Heterogeneous Treatment Effects and Policy Assignment without Unconfoundedness

Susan Athey – Stanford University

Athey (The Impact of Machine Learning on Economics, forth.)
Athey, Tibshirani and Wager (Generalized Random Forests, AOS, 2019)
Athey and Wager (Efficient Policy Learning, 2016)

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Holland, 1986, Imbens and Rubin, 2015, Rosenbaum and Rubin, 1983, Rubin, 1974), we posit the existence of quantities $Y_i^{(0)}$ and $Y_i^{(1)}$.

- These correspond to the response we **would have measured** given that the i-th subject received treatment ($W_i = 1$) or no treatment ($W_i = 0$).

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Goal is to estimate the **conditional average treatment effect**

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,\middle|\, X = x\right].$$

NB: In experiments, we only get to see $Y_i = Y_i^{(W_i)}$.

The potential outcomes framework

If we make no further assumptions, estimating $\tau(x)$ is not possible.

- Literature often assumes **unconfoundedness** (Rosenbaum and Rubin, 1983)

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp\!\!\!\perp W_i \mid X_i.$$

- When this assumption holds, methods based on matching or propensity score estimation are usually consistent.

ML Methods for Causal Inference: More general models

- Much recent literature bringing ML methods to causal inference focus on single binary treatment in environment with unconfoundedness
- Economic models often have more complex estimation approaches
- Athey, Tibshirani, and Wager (2016) tackle general GMM case:
  - Quantile regression
  - Instrumental Variables
  - Panel regression
  - Consumer choice
  - Euler equations
  - Survival analysis

Forests for GMM Parameter Heterogeneity

- Local GMM/ML uses kernel weighting to estimate personalized model for each individual, weighting nearby observations more.
  - Problem: curse of dimensionality
- We propose forest methods to determine what dimensions matter for "nearby" metric, reducing curse of dimensionality.
  - Estimate model for each point using "forest-based" weights: the fraction of trees in which an observation appears in the same leaf as the target
- We derive splitting rules optimized for objective
- Computational trick:
  - Use approximation to gradient to construct pseudo-outcomes
  - Then apply a splitting rule inspired by regression trees to these pseudo-outcomes

Solving estimating equations with random forests

We have $i = 1, \ldots, n$ i.i.d. samples, each of which has an **observable** quantity O_i, and a set of **auxiliary covariates** X_i.

Examples:

- Non-parametric regression: $O_i = \{Y_i\}$.
- Treatment effect estimation: $O_i = \{Y_i, W_i\}$.
- Instrumental variables regression: $O_i = \{Y_i, W_i, Z_i\}$.

Our **parameter of interest**, $\theta(x)$, is characterized by an estimating equation:

$$\mathbb{E}\left[\psi_{\theta(x),\,\nu(x)}(O_i) \,\middle|\, X_i = x\right] = 0 \quad \text{for all} \quad x \in \mathcal{X},$$

where $\nu(x)$ is an optional **nuisance parameter**.

The GMM Setup: Examples

Our parameter of interest, $\theta(x)$, is characterized by

$$\mathbb{E}\left[\psi_{\theta(x),\,\nu(x)}(O_i) \,\middle|\, X_i = x\right] = 0 \quad \text{for all} \quad x \in \mathcal{X},$$

where $\nu(x)$ is an optional **nuisance parameter**.

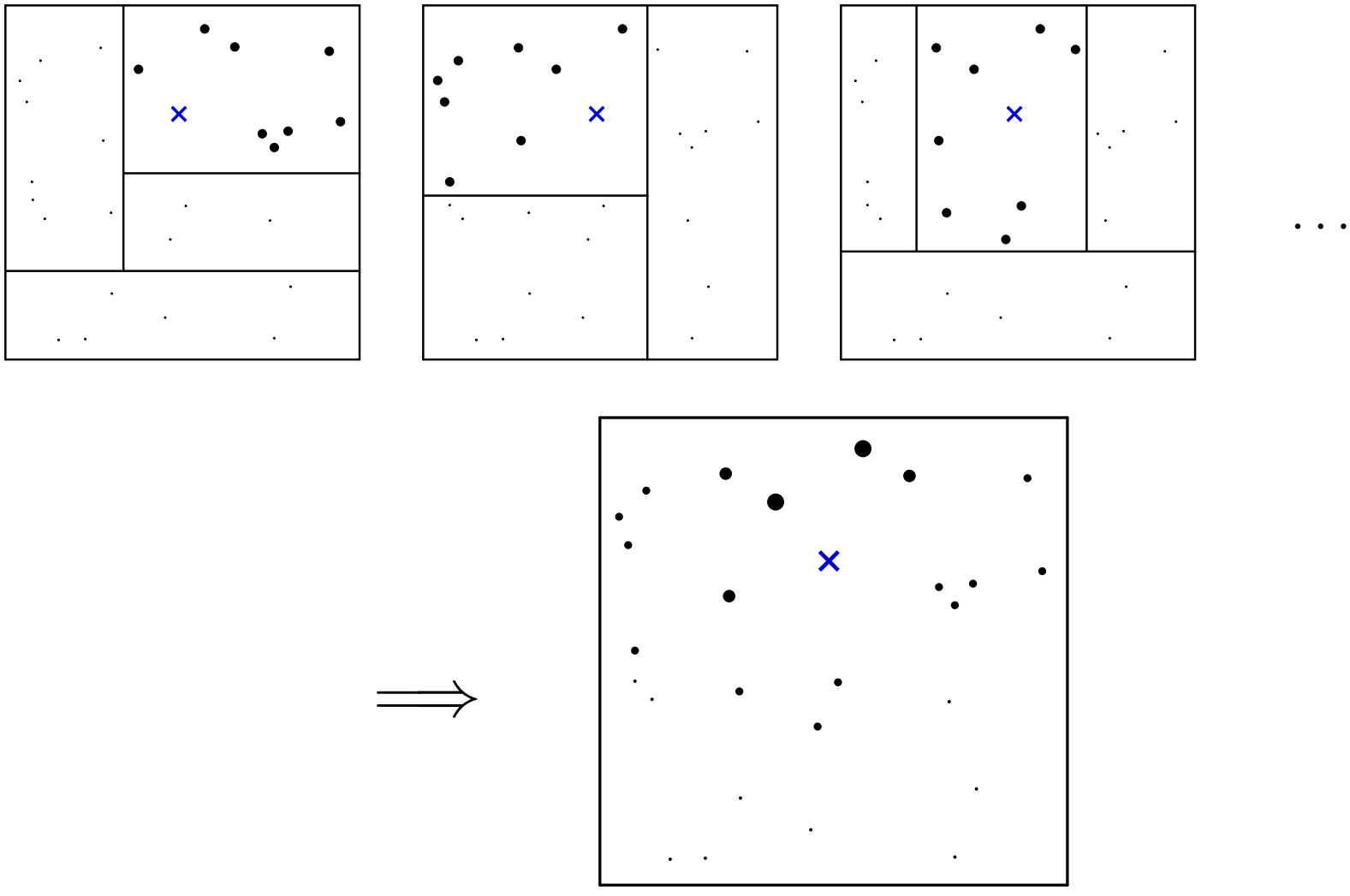
- **Quantile regression**, where $\theta(x) = F_x^{-1}(q)$ for $q \in (0,\, 1)$:

$$\psi_{\theta(x)}(Y_i) = q\,\mathbf{1}(\{Y_i > \theta(x)\}) - (1-q)\,\mathbf{1}(\{Y_i \leq \theta(x)\})$$

- **IV regression**, with treatment assignment W and instrument Z. We care about the treatment effect $\tau(x)$:

$$\psi_{\tau(x),\,\mu(x)} = \begin{pmatrix} Z_i(Y_i - W_i\,\tau(x) - \mu(x)) \\ Y_i - W_i\,\tau(x) - \mu(x) \end{pmatrix}.$$

The random forest kernel



Forests induce a kernel via **averaging tree-based neighborhoods**.
This idea was used by Meinshausen (2006) for quantile regression.

Solving estimating equations with random forests

We want to use an estimator of the form

$$\sum_{i=1}^{n} \alpha(x;\, X_i)\; \psi_{\hat{\theta}(x),\, \hat{\nu}(x)}\left(O_i\right) = 0,$$

where the weights $\alpha(x;\, X_i)$ are from a random forest.

Key Challenges:

- How do we grow trees that yield an **expressive** yet **stable** neighborhood function $\alpha(\cdot;\, X_i)$?
- We do not have access to **"prediction error"** for $\theta(x)$, so how should we **optimize splitting**?
- How should we account for **nuisance parameters**?
- Split evaluation rules need to be **computationally efficient**, as they will be run many times for each split in each tree.

Step #1: Conceptual motivation

Following CART (Breiman et al., 1984), we use **greedy splits**. Each split directly seeks to improve the fit as much as possible.

- For regression trees, in large samples, the **best split** is that which **increases the heterogeneity** of the predictions the most.
- The same fact also holds **locally** for estimating equations.

We split a parent node P into two children C_1 and C_2. In **large samples** and with **no computational constraints**, we would like to maximize

$$\Delta\left(C_1,\, C_2\right) = n_{C_1} n_{C_2} \left(\hat{\theta}_{C_1} - \hat{\theta}_{C_2}\right)^2,$$

where $\hat{\theta}_{C_1}$, $\hat{\theta}_{C_2}$ **solve the estimating equation in the children**.

Step #2: Practical realization

Computationally, solving the estimating equation in each possible child to get $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ can be **prohibitively expensive**.

To avoid this problem, we use a **gradient-based approximation**. The same idea underlies gradient boosting (Friedman, 2001).

$$\hat{\theta}_C \approx \tilde{\theta}_C := \hat{\theta}_P - \frac{1}{|\{i : X_i \in C\}|} \sum_{\{i:X_i \in C\}} \xi^\top A_P^{-1} \, \psi_{\hat{\theta}_P, \hat{\nu}_p}(O_i),$$

$$A_P = \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i:X_i \in P\}} \nabla \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i),$$

where $\hat{\theta}_P$ and $\hat{\nu}_P$ are obtained by solving the estimating equation once in the parent node, and ξ is a vector that picks out the θ-coordinate from the (θ, ν) vector.

Step #2: Practical realization

In practice, this idea leads to a **split-relabel** algorithm:

1. **Relabel step:** Start by computing pseudo-outcomes

$$\tilde{\theta}_i = -\xi^\top A_P^{-1} \, \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \in \mathbb{R}.$$

2. **Split step:** Apply a CART-style regression split to the $\widetilde{Y}_i$.

This procedure has several advantages, including the following:

- **Computationally**, the most demanding part of growing a tree is in scanning over all possible splits. Here, we reduce to a regression split that can be efficiently implemented.
- **Statistically**, we only have to solve the estimating equation once. This reduces the risk of hitting a numerically unstable leaf—which can be a risk with methods like IV.
- From an **engineering** perspective, we can write a single, optimized split-step algorithm, and then use it everywhere.

Step #3: Variance correction

Conceptually, we saw that—in large samples—we want splits that maximize the heterogeneity of the $\hat{\theta}(X_i)$. In small samples, we need to account for **sampling variance**.

We need to penalize for the following two sources of variance.

- Our **plug-in estimates** for the heterogeneity of $\hat{\theta}(X_i)$ will be **overly optimistic** about the large-sample parameter heterogeneity. We need to correct for this kind of over-fitting.
- We **anticipate "honest" estimation**, and want to avoid leaves where the **estimating equation is unstable**. For example, with IV regression, we want to avoid leaves with an unusually weak 1st-stage coefficient.

This is a generalization of the analysis of Athey and Imbens (2016) for treatment effect estimation.

Generalized Random Forests

Our label-and-regress splitting rules can be used to grow an ensemble of trees that yield a forest kernel. We call the resulting procedure a **generalized random forest**.

- Regression forests are a special case of generalized random forests with a squared-error loss.

Available as an `R`-package, `grf`.

Asymptotic normality of generalized random forests

Theorem. (Athey, Tibshirani and Wager, 2016) Given regularity of both the estimating equation and the data-generating distribution, generalized random forests are **consistent** and **asymptotically normal**:

$$\frac{\hat{\theta}_n(x) - \theta(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0, 1), \quad \sigma_n^2 \to 0.$$

Proof sketch.

- Influence functions: Hampel (1974); also parallels to use in Newey (1994).
- Influence function heuristic motivates approximating generalized random forests with a class of regression forests.
- Analyze the approximating regression forests using Wager and Athey (2018)
- Use coupling result to derive conclusions about generalized random forests.

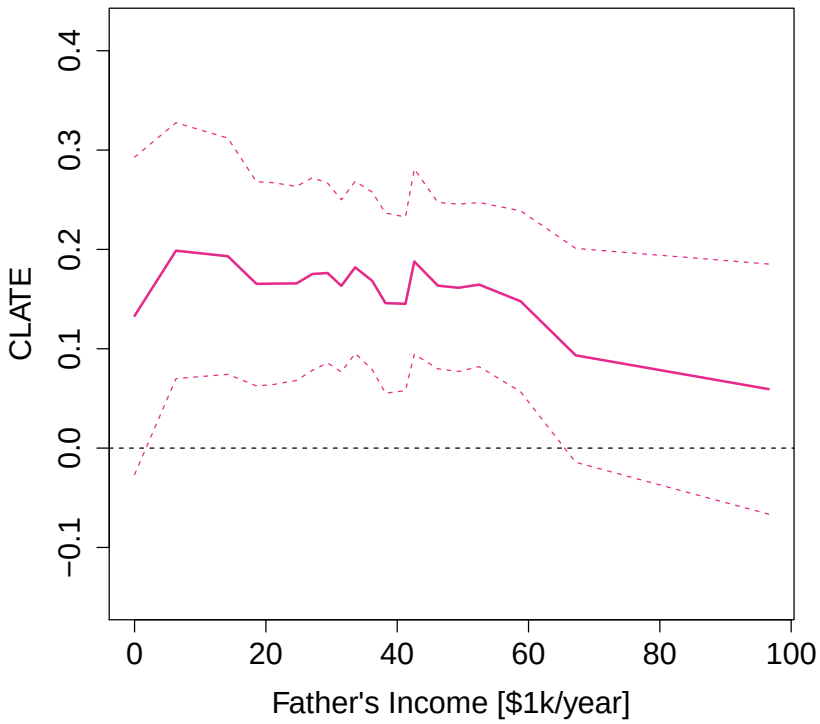
Empirical Application: Family Size

Angrist and Evans (1998) study the effect of family size on women's labor market outcomes. Understanding heterogeneity can guide policy.

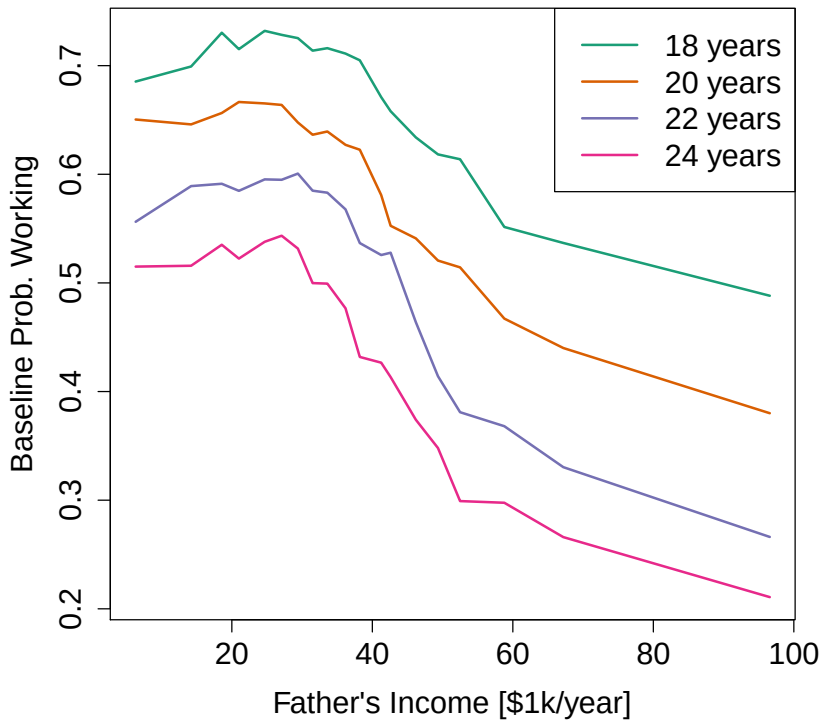
- Outcomes: participation, female income, hours worked, etc.
- Treatment: more than two kids
- Instrument: first two kids same sex
- First stage effect of same sex on more than two kids: .06
- Reduced form effect of same sex on probability of work, income: .008, $132
- LATE estimates of effect of kids on probability of work, income: .133, $2200

Treatment Effects: Magnitude of Decline

Effect on Participation

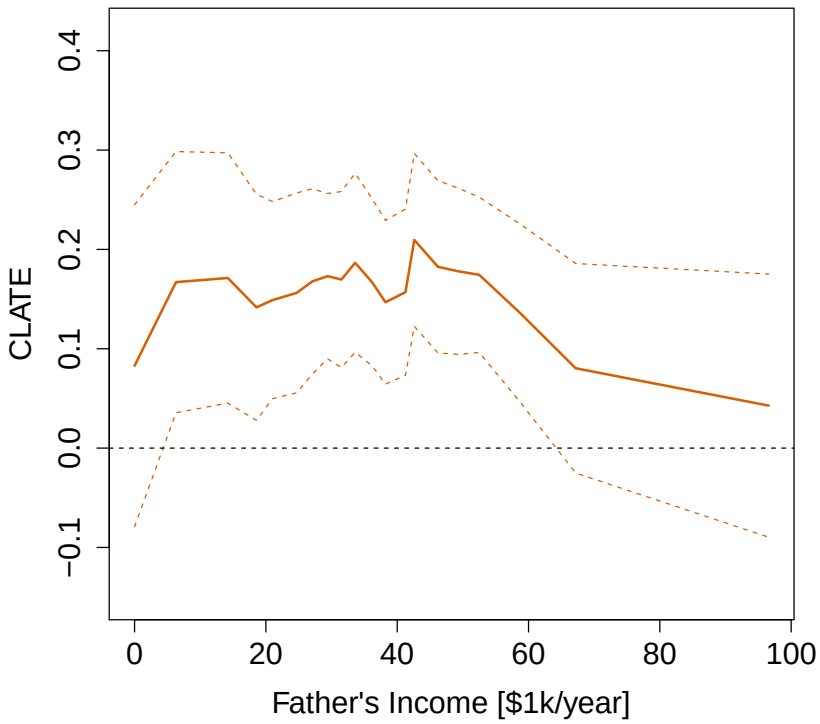


Baseline Probability of Working

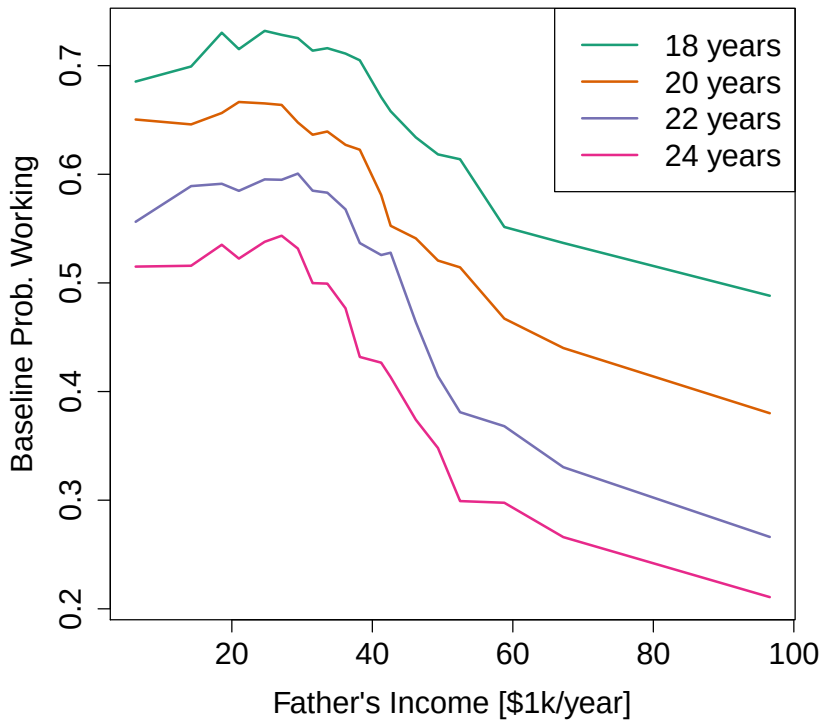


Treatment Effects: Magnitude of Decline

Effect on Participation

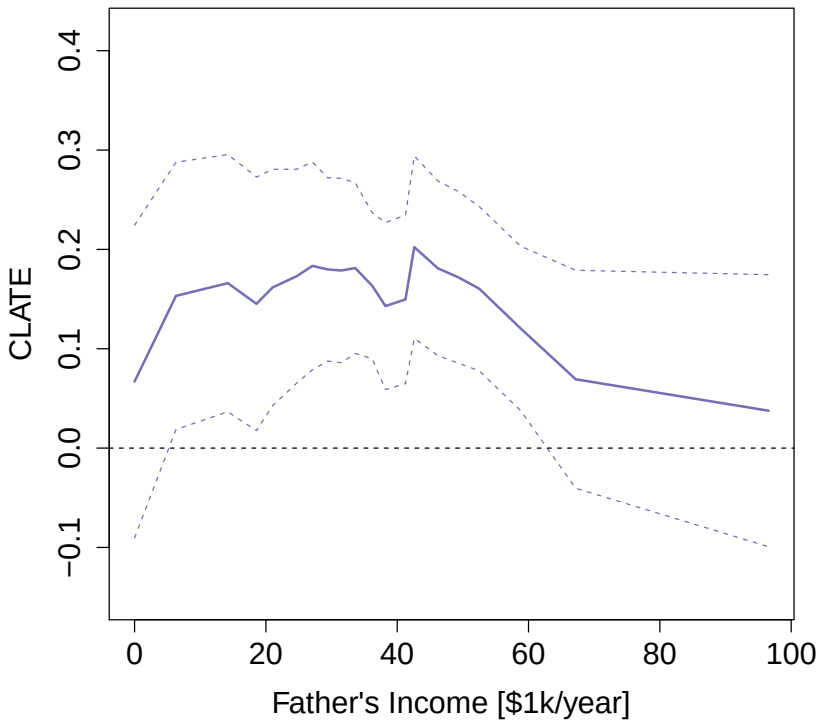


Baseline Probability of Working

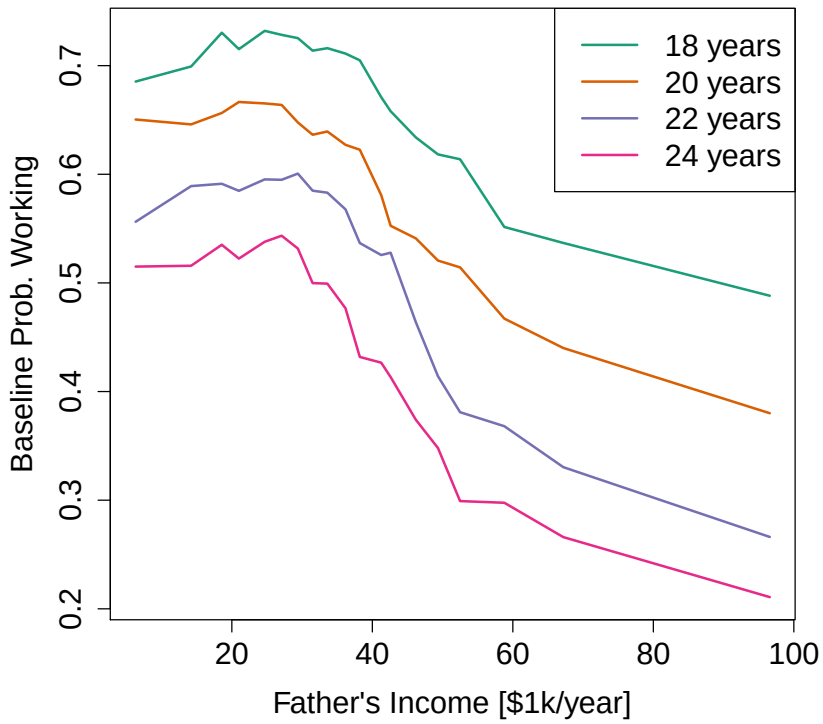


Treatment Effects: Magnitude of Decline

Effect on Participation

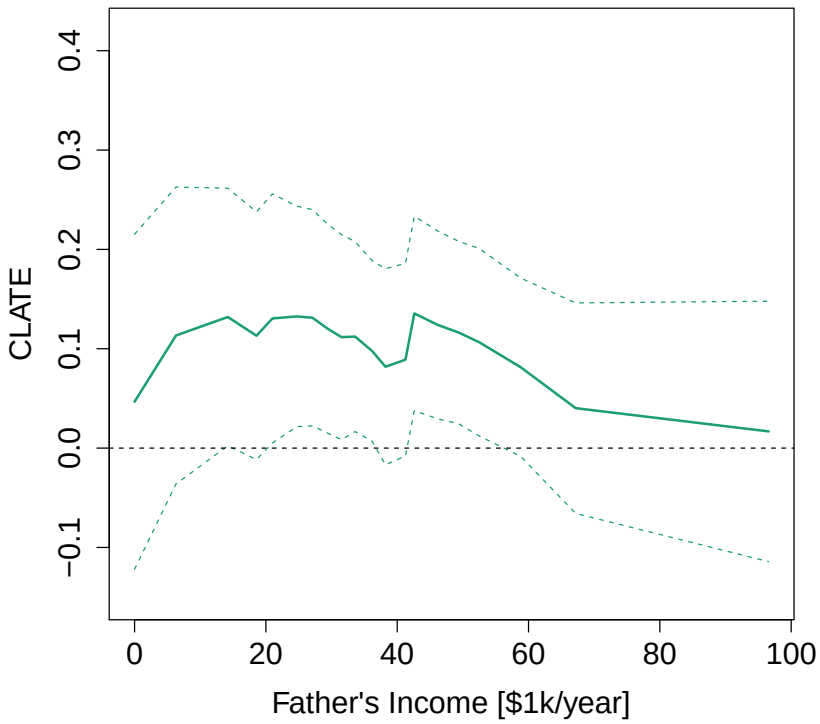


Baseline Probability of Working

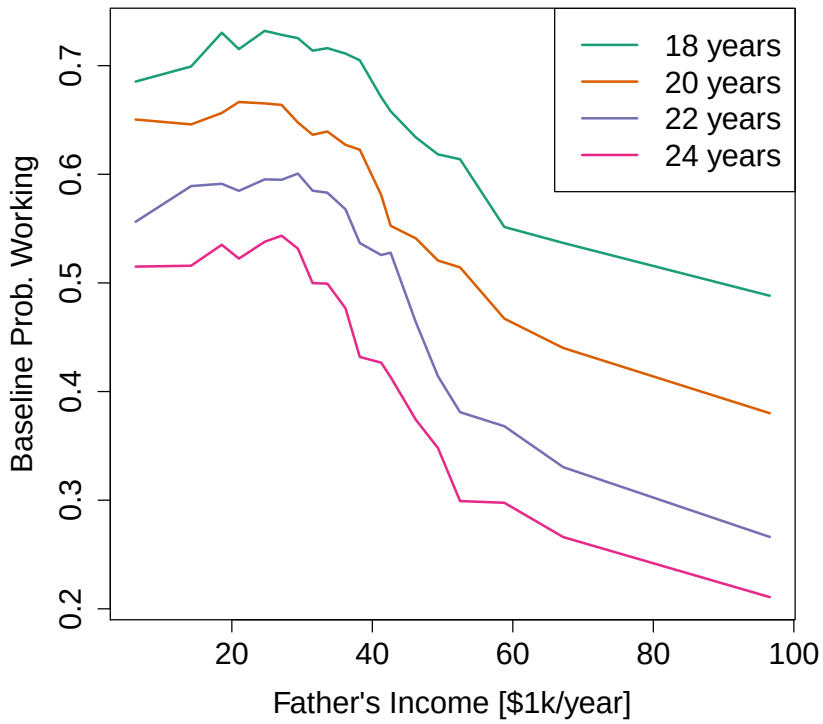


Treatment Effects: Magnitude of Decline

Effect on Participation



Baseline Probability of Working



Using Heterogeneous Treatment Effect Estimates for Optimal Targeted Policies

- Problem: find a policy $\pi : \mathcal{X} \to \mathcal{W}$ to maximize $\mathbb{E}[Y_i(\pi(X_i))]$.
- First approach: non-parametric treatment assigment
  - Treat if $\hat{\tau}(X_i) > 0$
  - Hirano and Porter (2009) show efficient (under conditions)
- How to evaluate success?
- If $\hat{\tau}(X_i)$ OOB estimate from random forest, then the implied $\hat{\pi}(X_i)$ is independent of Y_i.
  - Define Group $G^w = \{i : \hat{\pi}(X_i) = w\}$, proportion in G^w is q^w.
  - Define $\hat{\gamma}^w$ as sample average treatment effect in G^w.
  - Improvement of $\hat{\pi}(\cdot)$ over treating no one: $q^1 \cdot \hat{\gamma}^1$
  - ... over random policy: $\frac{1}{2}(q^1\hat{\gamma}^1 - q^0\hat{\gamma}^0)$.
  - Standard errors straightforward

Policy Learning

The utilitarian **value** of a policy $\pi : \mathcal{X} \to \{0, 1\}$ is

$$V(\pi) = \mathbb{E}\left[Y_i(\pi(X_i))\right] = \mathbb{E}\left[Y_i(0)\right] + \mathbb{E}\left[\tau(X)\pi(X)\right].$$

In the abstract, we maximize utility by treating according to a **thresholding rule** $\tau(X_i) > c$.

But estimating the conditional average **treatment effect function** $\tau(\cdot)$ and learning a good **policy** $\pi(\cdot)$ are different problems.

- The correct **loss function** for policy learning is not mean-squared error on $\tau(\cdot)$.
- The $\tau(x)$ function may change with variables we cannot use for **targeting** (e.g., variables only measured after the fact).
- We may wish to impose other constraints on policy functions

Policy Learning

We seek to maximize the **utility** of the learned policy subject to **constraints** encoded via a class Π of allowed policies.

As in Manski (2004), we focus on **minimax regret** (Savage, 1951) relative to the policy class Π. We define utility regret as $R(\pi)$,

$$R(\pi) = \sup \left\{ V(\pi') : \pi' \in \Pi \right\} - V(\pi),$$

and seek a policy $\hat{\pi} \in \Pi$ satisfying a high-probability **regret bound**.

We can also write policy regret in terms of $\tau(x)$,

$$R(\pi) = \sup \left\{ \mathbb{E}\left[\tau(X)\pi'(X)\right] : \pi' \in \Pi \right\} - \mathbb{E}\left[\tau(X)\pi(X)\right],$$

meaning that baseline effects don't affect policy regret.

Imposing **structure** on Π is essential in many applications. In observational studies, we use many features with a non-parametric specification to make **unconfoundedness plausible**,

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i.$$

Conversely, the policy $\pi(\cdot)$ must be **implementable in practice**. Features that should not be used in $\pi(\cdot)$ include:

- **Unreliably available features** (e.g., collected by specialist).
- **Gameable features** (e.g., self-reported preferences).
- **Legally protected classes** (e.g., religion, national origin).

Moreover, we may want Π to encode constraints on:

- **Total budget** or marginal **subgroup treatment rates** (e.g., Bhattacharya and Dupas, 2012).
- **Functional form** for easier implementation or audit.

We study policy learning in a way that is aware of such constraints.

From Evaluation to Learning

Recall that we want to pick a good **policy** $\pi : \mathcal{X} \to \{0, 1\}$ among a class Π of allowable interventions.

The **regret** from choosing π depends on the CATE function $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$:

$$R(\pi) = \sup\left\{\mathbb{E}\left[\tau(X)\pi'(X)\right] : \pi' \in \Pi\right\} - \mathbb{E}\left[\tau(X)\pi(X)\right].$$

Before discussing how to **learn** a policy, we review how how to estimate an **average effect**

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

We build on unifying results from Chernozhukov, Escanciano, Ichimura, Newey and Robins (CEINR, 2018).

From Evaluation to Learning

We have access to features X_i, an outcome Y_i, a treatment W_i, and an instrument Z_i. We assume that the exclusion restriction holds, such that potential outcomes only depend on W_i, and

$$m(x, w) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right], \quad \tau_m(x) = m(x, 1) - m(x, 0).$$

As in CEINR, suppose $\tau(x)$ can be represented via **weighting**:

$$\mathbb{E}\left[\tau_m(X) - g(X, Z)Y \,\middle|\, X = x\right] = 0 \text{ for all } x,\ m(\cdot).$$

CEINR then show that the **doubly robust** estimator is **efficient**,

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i, Z_i)\left(Y_i - \hat{m}(X_i, W_i)\right),$$

provided we use **cross-fitting** and have nuisance components that converge fast enough in L_2 (4th-root rates are sufficient).

From Evaluation to Learning

As in CEINR, suppose $\tau(x)$ can be represented via **weighting**:

$$\mathbb{E}\left[\tau_m(X) - g(X, Z)m(X, W) \,\middle|\, X = x\right] = 0 \text{ for all } x,\ m(\cdot).$$

CEINR then show that the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i, Z_i)\left(Y_i - \hat{m}(X_i, W_i)\right).$$

Example: Selection on observables, $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i$. In this case, the **propensity score** can be used for weighting:

$$\tau(x) = \mathbb{E}\left[g(X_i, W_i)Y_i \,\middle|\, X_i = x\right], \; g(X_i, W_i) = \frac{(W_i - e(X_i))}{e(X_i)(1 - e(X_i)}.$$

The corresponding doubly robust estimator is **augmented IPW** (Robins, Rotnitzky, and Zhao, 1994).

From Evaluation to Learning

As in CEINR, suppose $\tau(x)$ can be represented via **weighting**:

$$\mathbb{E}\left[\tau_m(X) - g(X, Z)m(X, W) \,\middle|\, X = x\right] = 0 \text{ for all } x,\ m(\cdot).$$

CEINR then show that the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i, Z_i)\left(Y_i - \hat{m}(X_i, W_i)\right).$$

Example: Endogenous treatment with instrument and **conditional homogeneity**, $\tau(x) = \text{Cov}\left[Y, Z \,\middle|\, X = x\right] / \text{Cov}\left[W, Z \,\middle|\, X = x\right]$. Now use the **compliance score** (Aronow and Carnegie, 2013),

$$g(X_i, Z_i) = \frac{1}{\Delta(X_i)}\frac{Z_i - z(X_i)}{z(X_i)(1 - z(X_i)}, \quad z(x) = \mathbb{P}\left[Z_i \,\middle|\, X_i = x\right],$$
$$\Delta(x) = \mathbb{P}\left[W \,\middle|\, Z = 1, X = x\right] - \mathbb{P}\left[W \,\middle|\, Z = 0, X = x\right],$$

to construct a doubly robust estimator.

From Evaluation to Learning

In many problems, the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i, Z_i)\,(Y_i - \hat{m}(X_i, W_i))\,.$$

Our main result is that we can also use the same scores of **learning**

$$\hat{\pi} = \text{argmax}\left\{ \frac{1}{n}\sum_{i=1}^{n} (2\pi(X_i) - 1)\,\widehat{\Gamma}_i : \pi \in \Pi \right\}.$$

Regret bounds depend on n, Π, and the semiparametric efficient variance for policy evaluation.

From Evaluation to Learning

In many problems, the **doubly robust** estimator is **efficient**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_i, \quad \widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i, Z_i)\,(Y_i - \hat{m}(X_i, W_i))\,.$$

Our main result is that we can also use the same scores of **learning**

$$\hat{\pi} = \text{argmax}\left\{ \frac{1}{n}\sum_{i=1}^{n} (2\pi(X_i) - 1)\,\widehat{\Gamma}_i : \pi \in \Pi \right\}.$$

Regret bounds depend on n, Π, and the semiparametric efficient variance for policy evaluation.

NB: The policy $\pi^* = \text{argmax}\left\{\mathbb{E}\left[(2\pi(X_i) - 1)\tau(X_i)\right] : \pi \in \Pi\right\}$ gets **zero regret**. Our estimator effectively replaces $\tau(X_i)$ with $\widehat{\Gamma}_i$.

Back to the California GAIN Study

Each county enrolled participants with a **different covariate mix**, and randomized to treatment with **different probabilities**. Once we remove county information, this is not a **randomized study**, but Hotz et al. present evidence that **unconfoundedness** holds.

We set the **cost** C of treatment to match the **ATE**; thus, we need to find heterogeneity in order to get non-zero utility.

We estimate nuisance components with **forests**, and then optimize over the class Π of low-depth **trees**:

$$\hat{\pi} = \text{argmax} \left\{ \frac{1}{n} \sum_{i=1}^{n} (2\pi(X_i) - 1) \, \widehat{\Gamma}_i : \pi \in \Pi \right\},$$

$$\widehat{\Gamma}_i = \hat{\tau}^{(-i)}(X_i) - C + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i) \left(1 - \hat{e}^{(-i)}(X_i)\right)} \times \left(Y_i - \hat{y}^{(-i)}(X_i) - (W_i - \hat{e}^{(-i)}(X_i)) \hat{\tau}^{(-i)}(X_i) \right).$$

Main Result

Goal is to show that we can use doubly robust scores of **learning**

$$\hat{\pi} = \text{argmax} \left\{ \frac{1}{n} \sum_{i=1}^{n} (2\pi(X_i) - 1) \widehat{\Gamma}_i : \pi \in \Pi \right\},$$
$$\widehat{\Gamma}_i = \tau_{\hat{m}}(X_i) + \hat{g}(X_i, Z_i) (Y_i - \hat{m}(X_i, W_i)).$$

Theorem. (Athey and Wager, 2018) Suppose $g(X, Z) \leq \eta^{-1}$, and that our nuisance estimates satisfy (we use **cross-fitting**)

$$\mathbb{E}\left[(\hat{m}(X, W) - m(X, W))^2\right] \mathbb{E}\left[(\hat{g}(X, Z) - g(X, Z))^2\right] = o_P\left(\frac{1}{n}\right).$$

Suppose moreover that Π has finite **VC-dimension**. Then,

$$R(\hat{\pi}) = \mathcal{O}_P\left(\sqrt{S \, \text{VC}(\Pi) / n}\right),$$

with $S = \mathbb{E}\left[(\tau_m(X_i) + g(X_i, Z_i)(Y_i - m(X_i, W_i)))^2\right]$.

Proof Ingredients

- ATE literature looks at efficient coupling of the **efficient score**: $|\widehat{A}(\pi) - \widetilde{A}(\pi)| = o_P(1/\sqrt{n})$, where

$$\widetilde{A}(\pi) = \sum_{i=1}^{n} (2\pi(X_i) - 1) \left(\tau_m(X_i) + g(X_i, Z_i)\left(Y_i - m(X_i, W_i)\right)\right).$$

- Here, need **uniform coupling**:
  $\sup\left\{\left|\widehat{A}(\pi) - \widetilde{A}(\pi)\right| : \pi \in \Pi\right\} = o_P(1/\sqrt{n})$.
  - Our uniform coupling result is specific to the **doubly robust** construction, and may not hold for other estimators that are efficient at a single π, e.g., empirical IPW (Hirano et al., 2003).
- Next: **concentration** of $\widetilde{A}(\pi)$ over the class $\pi \in \Pi$. Defining

$$S = \mathbb{E}\left[\left(\tau_m(X_i) + g(X_i, Z_i)\left(Y_i - m(X_i, W_i)\right)\right)^2\right],$$

  remix Dudley's classical **chaining argument** to verify that

$$\sup\left\{\left|\widetilde{A}(\pi) - A(\pi)\right| : \pi \in \Pi\right\} = \mathcal{O}_P\left(\sqrt{\frac{S\ VC(\Pi)}{n}}\right).$$

  Bound **Rademacher complexity** via chaining.

Lower bounds

Any statement about lower bounds depends on how **general** we want to be, and how **adaptive** we want to be to problem structure. We proved that $R(\hat{\pi}) = \mathcal{O}_P(\sqrt{S \operatorname{VC}(\Pi)/n})$, and argue that this is optimal. We first note, however:

- If treatment effects are **smaller** than $1/\sqrt{n}$, bound is loose.
- If treatment effects are **very large**, this bound is loose as finding the optimal rule is easy (Luedtke and Chambaz, 2017).
- VC-dimension may be a loose summary of the **complexity** of Π (Bartlett and Mendelson, 2006).

We show that our bound is **sharp** when our treatment effects scale as $1/\sqrt{n}$ and we summarize complexity via VC-dimension. Similar **local asymptotics** are also used by Hirano and Porter (2009).

NB: Our upper bound allows the data-generating distribution (and Π) to change with n, so changing $\tau(\cdot)$ with n is valid.

Lower bounds

In the **unconfoundedness** setting, define a sequence of problems

$$X_i \sim \text{Uniform}\left(\mathcal{X}_s\right), \quad W_i \,\Big|\, X_i \sim \text{Bernoulli}(e(X_i)),$$
$$Y_i \,\Big|\, X_i,\, W_i \sim \left(y(X_i) + (W_i - e(X_i))\,\frac{\tau(X_i)}{\sqrt{n}},\, \sigma^2(X_i)\right).$$

Theorem. (Athey and W., 2018) In this setting, there is a class Π with $\text{VC}(\Pi) = d$ whose **minimax regret** satisfies

$$\liminf_{n\to\infty}\left\{\sqrt{n}\,\inf_{\hat{\pi}_n}\left\{\sup_{|\tau(x)|\leq C}\left\{\mathbb{E}\left[R_n\left(\hat{\pi}_n\right)\right]\right\}\right\}\right\} \geq 0.33\sqrt{Sd},$$

where $S = \mathbb{E}\left[\sigma^2(X)/(e(X)(1-e(X)))\right]$.

Our method **achieves this bound** up to a universal constant. Other methods do not, e.g., for **IPW** with known propensity scores, Kitagawa & Tetenov (2018) prove a bound that depends on $\sup\{|Y_i|\} / \inf\{e(X_i), (1-e(X_i))\}$ instead of $\sqrt{S}$.

Conclusion

- Machine learning based methods very useful to analyze heterogeneous treatment effects and targeted policies
- Methods work in a variety of design settings (experiments, unconfounded, IV)
- Methods can give either simple or very complex policies, with statistical guarantees
- Online learning can help discover good policies
- Semi-parametric efficiency literature guides methods