Evaluating and Testing for Predictive Treatment Effect Heterogeneity

Nicholas C. Henderson

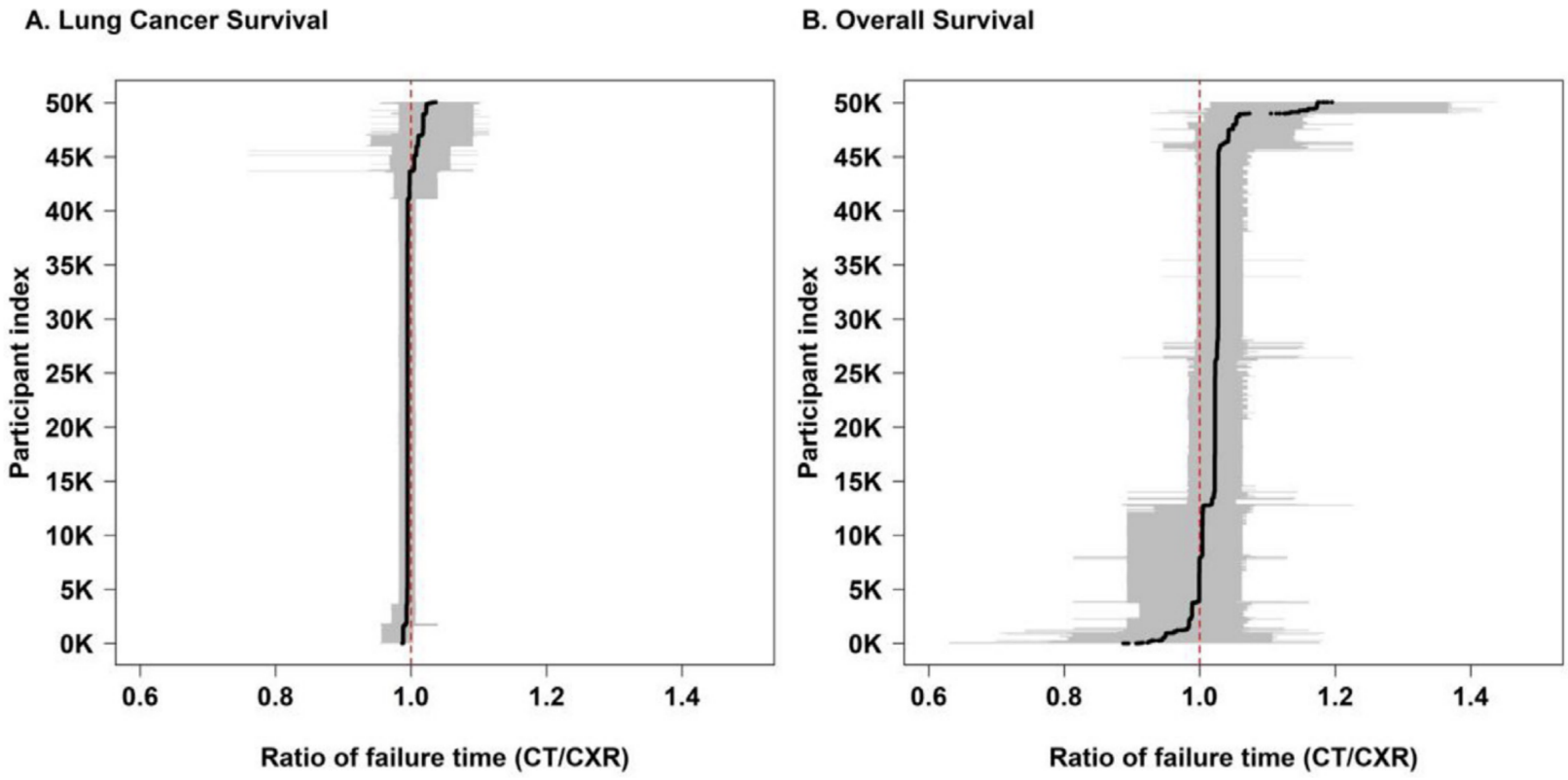
Department of Biostatistics
University of Michigan

April 4, 2025

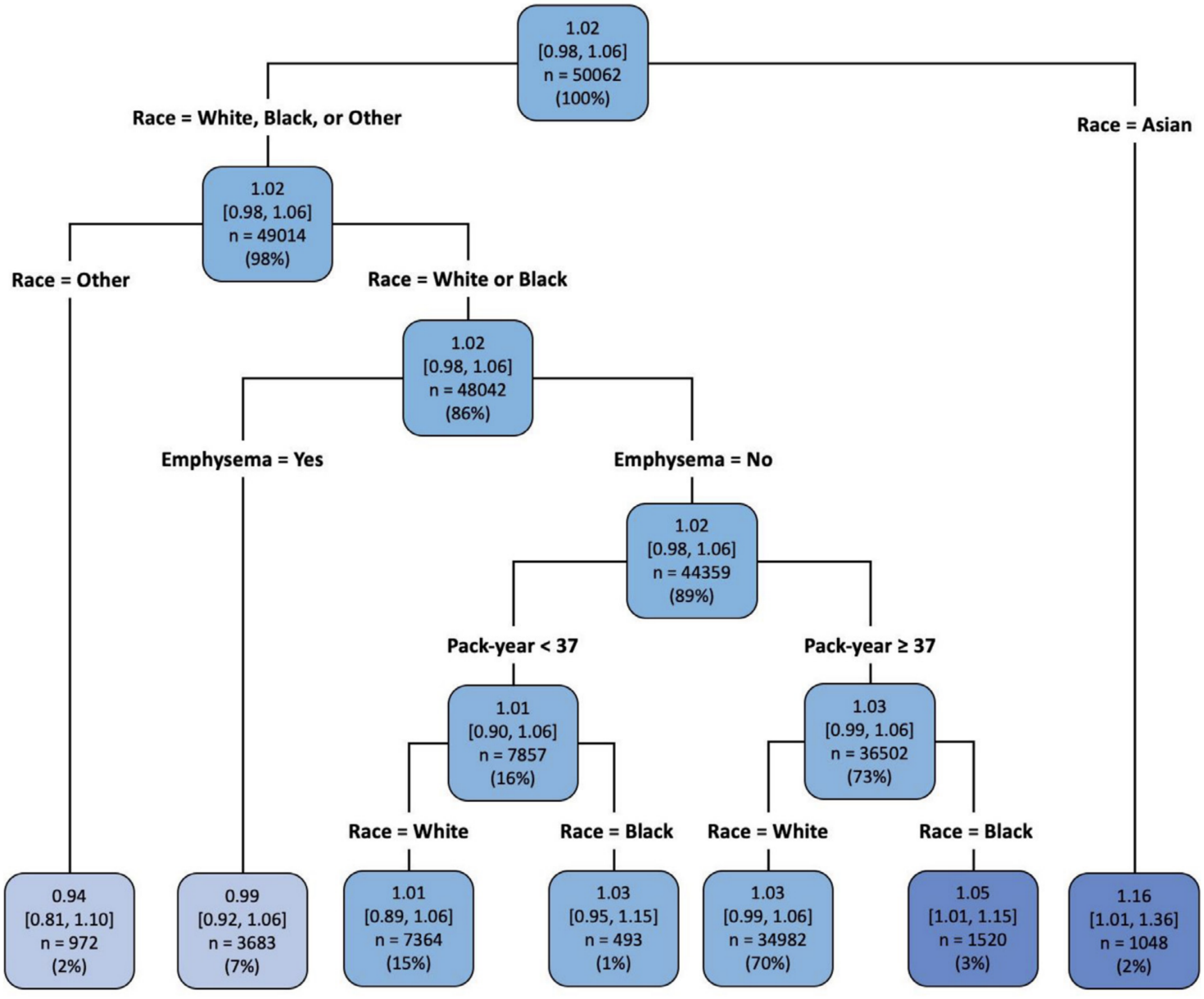
Background/Motivation

- Many statistical tools to investigate **heterogeneous treatment effects** (HTE)/precision medicine goals have emerged in recent years.
- Weak variation in treatment effects and limited sample sizes can hinder efforts to capture treatment effect variability.
- Can often be unclear how much models which incorporate HTE improve upon those that do not incorporate HTE in a predictive sense.
  - In the past 5-10 years, there has been a marked increase in the use of machine learning methods for estimating individualized treatment effects.

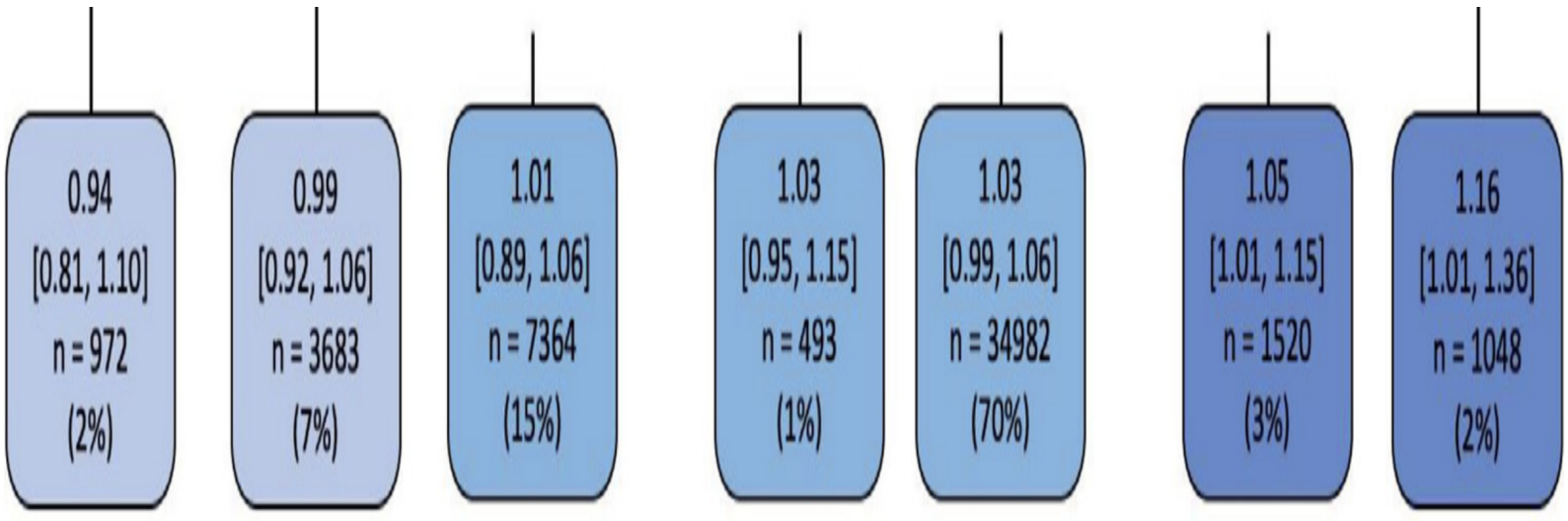
The NLS Trial (Hu et al. (2021))



The NLS Trial - CART for treatment effects



The NLS Trial



Aims

- **Goal:** Want to compare predictive performance of an HTE model with the closest "non-HTE" version of the model.
- This will be an overall assessment of whether or not your statistical modeling approach supports incorporating HTE.
  - Similar idea to an overall F-test in linear regression
- We want to be able to apply this predictive evaluation to any procedure.
  - This includes "black box" machine learning methods or other "AI" tools

Data Structure

- Available dataset: $\mathcal{D} = \{(Y_i, A_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$
  - Outcome - Y_i
  - Randomized treatment assignment A_i

$$A_i = \begin{cases} 1 & \text{assigned to "active treatment" group} \\ 0 & \text{assigned to control group} \end{cases}$$

  - Vector of covariates - $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$.

- Potential outcomes: $\big(Y_i(0), Y_i(1)\big)$

- $Y_i(a)$ - outcome if individual was assigned treatment a.

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

Data Structure

- You can think of the (unobserved) treatment effect as: $Y_i(1) - Y_i(0)$
- Our main aim is to examine the extent to which modeling variation in treatment effect predictive performance.
- To do this, the **conditional mean function** of Y_i will be useful:

$$\begin{aligned} f(A, \boldsymbol{x}) &= E\{Y_i \mid A_i = A, \boldsymbol{x}_i = \boldsymbol{x}\} \\ &= AE\{Y_i(1) \mid \boldsymbol{x}_i = \boldsymbol{x}\} + (1 - A)E\{Y_i(0) \mid \boldsymbol{x}_i = \boldsymbol{x}\}, \end{aligned}$$

- For example,

$$Y_i = f(A_i, \boldsymbol{x}_i) + \varepsilon_i,$$

ITE function

- The **Individualized treatment effect** (ITE) function (or, conditional average treatment effect (CATE)) is:

$$\begin{aligned} \theta(\mathrm{x}_i) &= E\{Y_i \mid A_i = 1, \boldsymbol{x}_i\} - E\{Y_i \mid A_i = 0, \boldsymbol{x}_i\} \\ &= f(1, \boldsymbol{x}_i) - f(0, \boldsymbol{x}_i) \end{aligned}$$

- HTE is **present** if $\theta(\boldsymbol{x}_i)$ varies across different values of x_i.
- HTE is **not present** if $\theta(\boldsymbol{x}_i)$ is constant for all x_i.
- A statistical model allows for HTE if $\hat{\theta}(\boldsymbol{x})$ is not constant for all $\boldsymbol{x}$.

ITE function

- We are **not interested** in "testing" whether or not $\theta(\boldsymbol{x})$ is exactly constant.
  - Some HTE is likely to be present in most settings.
- We are interested in whether allowing for HTE improves the predictive performance of a statistical model.
  - Specifically, we want to determine if there is substantial evidence that predictive performance is improved.

Overall Strategy

1. Find an **"unrestricted estimator"** which does not place any restrictions on the form of the individualized treatment effect function.

2. Find a **"restricted estimator"** of the conditional mean function that does not allow for treatment effect variability.
   - This will be the estimator that is the "closest" zero-HTE estimator to the unrestricted one.

3. Choose a loss function that compares predictive performance of the restricted estimator and unrestricted estimators.

4. Construct a confidence interval for the expected value of this loss on "future observations".

Restricted Estimation: Treatment-shifted outcomes

- Under an assumption of **no HTE** and a constant treatment effect **equal to** τ, a reasonable outcome model for continuous outcomes would be

$$Y_i = f_\tau(\mathrm{x}_i) + A_i\tau + e_i.$$

- $f_\tau(\mathrm{x}_i)$ can be thought of as a **"baseline risk function"** when there is no HTE:

$$f_\tau(\mathrm{x}) = E\{Y_i \mid A_i = 0, \mathrm{x}_i = \mathrm{x}\}$$

- If we define $Z_i^\tau = Y_i - A_i\tau$, then

$$Z_i^\tau = f_\tau(\mathrm{x}_i) + e_i.$$

The τ-adjusted baseline risk estimator

$$Y_i = f_\tau(\boldsymbol{x}_i) + A_i\tau + e_i.$$

- We can estimate $f_\tau(\mathrm{x}_i)$ in the above model, by fitting a model with **shifted outcomes** $Z_i^\tau = Y_i - \tau A_i$ and covariates $\boldsymbol{x}_i$
  - Treatment assignment **should not** be included in the covariates when estimating $f_\tau(\mathrm{x})$.
- $\hat{f}_\tau(\boldsymbol{x})$ - "τ-adjusted" baseline risk estimator.

The τ-adjusted baseline risk estimator

$$Y_i = f_\tau(\boldsymbol{x}_i) + A_i\tau + e_i.$$

- Find the best value τ^* of τ by minimizing

$$\tau^* = \arg\min_\tau \sum_{i=1}^{n} \left(Y_i - \tau A_i - \hat{f}_\tau(\boldsymbol{x}_i) \right)^2.$$

- τ^* can be thought of as a treatment effect estimate **under the restriction** that there is no HTE.

The zero-HTE restricted estimator

- Using τ^*, the **restricted estimator** of the conditional mean function for Y_i is:

$$\hat{g}(A_i, \mathrm{x}_i) = \hat{f}_{\tau^*}(\mathrm{x}_i) + \tau^* A_i.$$

- Using the restricted estimator guarantees that there **no HTE in the fitted model**.
- Our procedure is model agnostic and can be applied when using **any complicated nonparametric/machine learning tool** to model $f(A, \mathrm{x})$ or $f_\tau(\mathrm{x})$.
- The restricted estimator $\hat{g}(A, \mathrm{x})$ can be a very flexible estimator of the mean function $f(A, \mathrm{x})$. The main restriction is that the ITE function estimator $\hat{g}(1, \mathrm{x}) - \hat{g}(0, \mathrm{x})$ is constant.

The zero-HTE restricted estimator

- The restricted estimator can be thought of as a **more general version** of a linear model that only has a main effect for treatment.
- If you are considering a linear model for the unrestricted conditional mean function

$$f(A, \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta} + A\boldsymbol{x}\boldsymbol{\gamma},$$

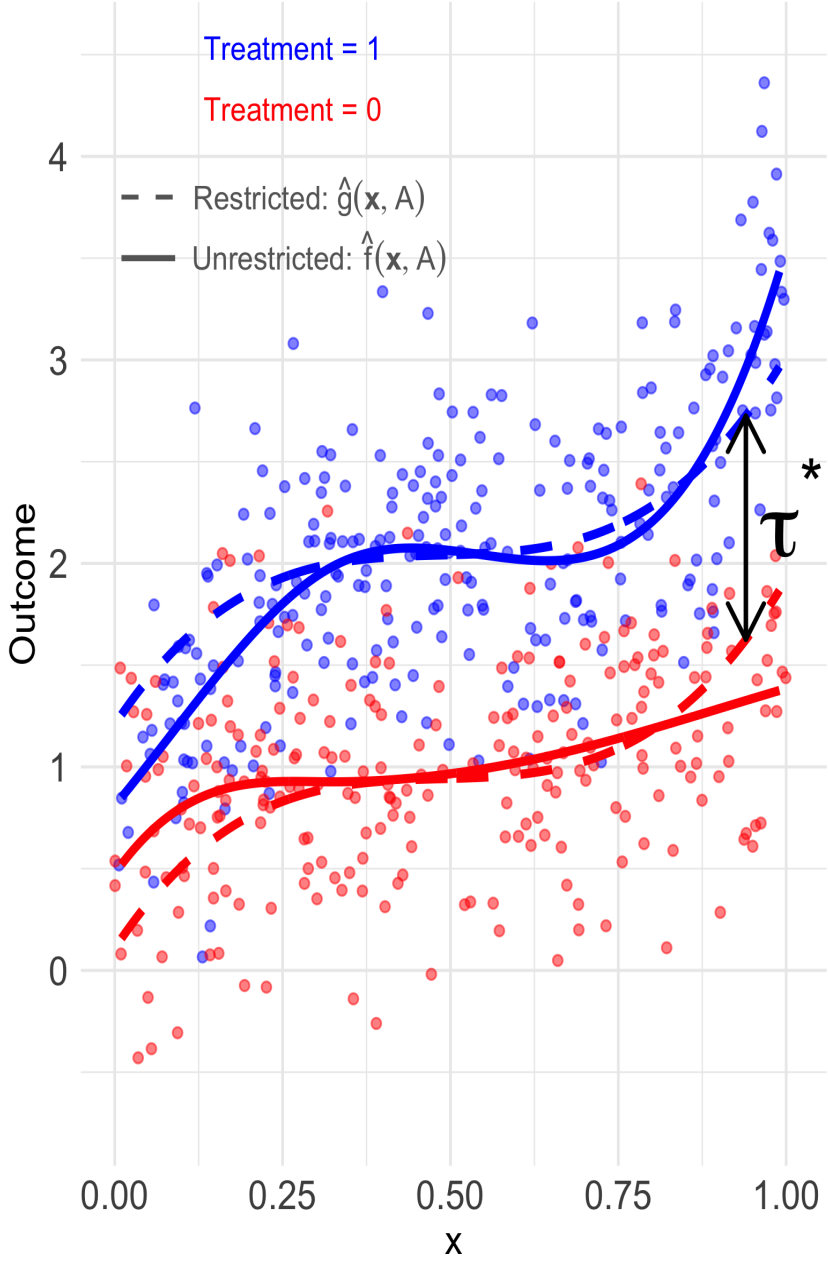
- Then, the restricted estimator has the form

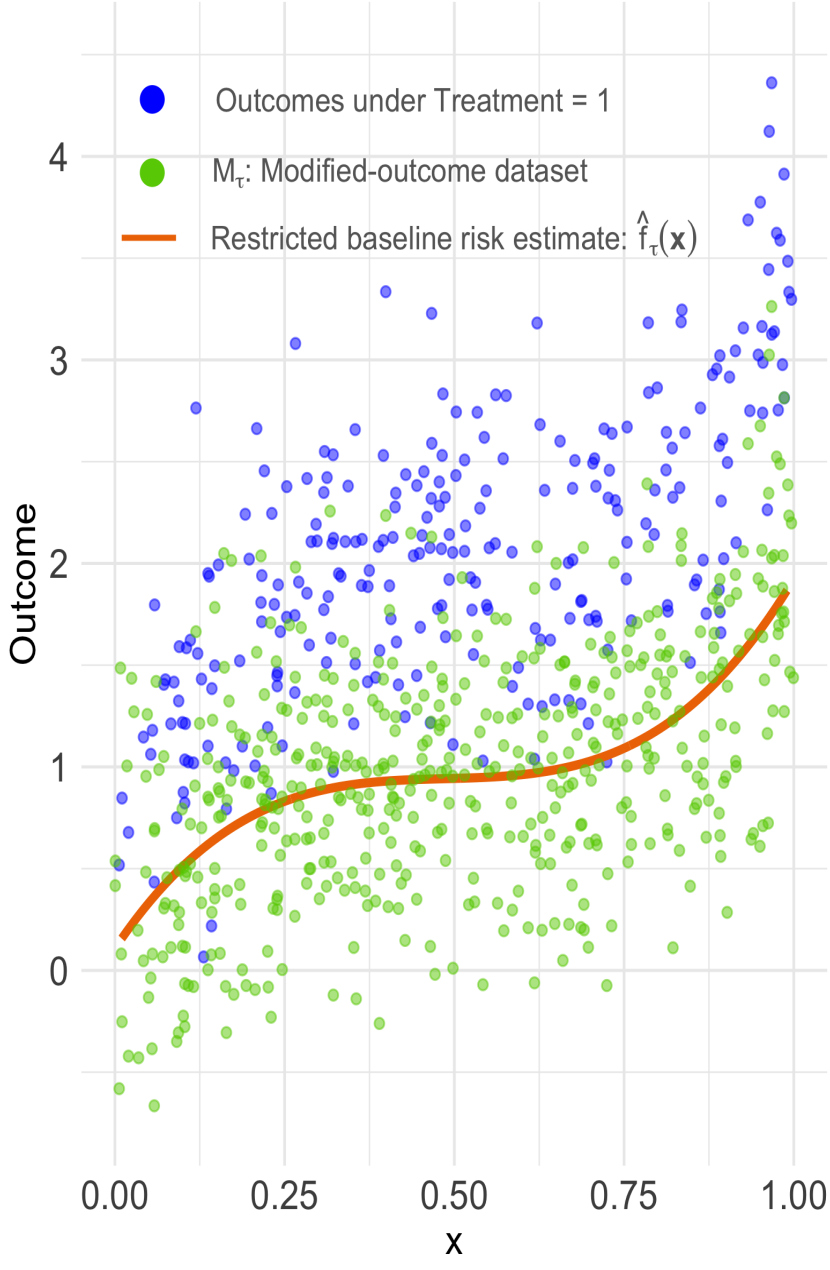
$$\hat{g}(A, \boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}_0 + A\hat{\gamma}_0,$$

where $\hat{\boldsymbol{\beta}}_0$ and $\hat{\gamma}_0$ are the least-squares estimates for the model:

$$Y_i = \mathrm{x}_i^T \boldsymbol{\beta} + A_i\gamma + \epsilon_i.$$

Illustration with Simulated Data





Predictive Comparison of Restricted and Unrestricted

- Our main goal is to evaluate the **predictive performance** of the following two mean function estimators:
  - The **unrestricted estimator** $\hat{f}(A, \mathrm{x})$ which allows for HTE.
  - The **restricted estimator** $\hat{g}(A, \mathrm{x})$ which has no HTE.

- Consider a "new observation" that one could observe

$$(Y_{n+1}, A_{n+1}, \mathrm{x}_{n+1})$$

- Measure the **difference in performance** of $\hat{f}(A, \mathrm{x})$ and $\hat{g}(A, \mathrm{x})$ with a loss function ℓ

$$\ell_{new} = \ell\Big(\hat{f}(\mathrm{x}_{n+1}, A_{n+1}), \hat{g}(\mathrm{x}_{n+1}, A_{n+1}), Y_{n+1}\Big).$$

Loss Function for Comparison

- The most direct loss function is the difference in squared-error loss:

$$\begin{aligned} \ell_{new} &= \left(Y_{n+1} - \hat{f}(\mathsf{x}_{n+1}, A_{n+1}; \mathcal{D})\right)^2 \\ &\quad - \left(Y_{n+1} - \hat{g}(\mathsf{x}_{n+1}, A_{n+1}; \mathcal{D})\right)^2. \end{aligned}$$

- One could also consider difference in absolute error or another measure of interest

$$\begin{aligned} \ell_{new} &= \left|Y_{n+1} - \hat{f}(\mathsf{x}_{n+1}, A_{n+1}; \mathcal{D})\right| \\ &\quad - \left|Y_{n+1} - \hat{g}(\mathsf{x}_{n+1}, A_{n+1}; \mathcal{D})\right| \end{aligned}$$

Comparing performance of restricted and unrestricted estimators

- We want our measure of comparative predictive performance to utilize **all the data**.
- While splitting into a training dataset/validation set is reasonable, this would measure predictive performance for models built on a smaller dataset (the size of the training set).
- **Alternative:** Construct confidence interval for the quantity that **cross-validation** (on the full dataset) best estimates.
- According to Bates et al. (2024), this is

$$\theta_{XY} = E\{\ell_{new} \mid \text{full dataset}\}.$$

- θ_{XY} is the expected future loss conditional on the dataset used to construct estimators

Confidence Intervals for Difference in Performance

- We want to perform inference on the following expected value of difference in performance

$$\theta_{XY} = E\{\ell_{new} \mid \mathcal{D}\}.$$

- Construct a $100 \times (1-\alpha)\%$ **confidence interval** $C_\alpha(\mathcal{D})$ for θ_{XY} using **nested cross-validation** techniques

$$C_\alpha(\mathcal{D}) = \hat{E}^{(ncv)} - \widehat{\mathrm{b}}^{(ncv)} \pm z_{1-\alpha/2}\sqrt{\widehat{mse}^{(ncv)}}$$

- The **h-value** is defined as the **smallest value of** α at which $C_\alpha(\mathcal{D})$ does not contain zero.

- The h-value can be computed with

$$h = 2\Phi\left(|\hat{E}^{(ncv)} - \widehat{\mathrm{b}}^{(ncv)}| \Big/ \sqrt{\widehat{mse}^{(ncv)}}\right)$$

Confidence Intervals for Difference in Performance

- The **h-value** is defined as the **smallest value of** α at which $C_\alpha(\mathcal{D})$ does not contain zero.
- If we write $\mathcal{C}_\alpha(\mathcal{D}) = \big(L_\alpha(\mathcal{D}), U_\alpha(\mathcal{D})\big)$:

$$h \quad = \quad \inf\left\{\alpha \in (0,1) : \operatorname{sign}(L_\alpha(\mathcal{D})) = \operatorname{sign}(U_\alpha(\mathcal{D}))\right\}$$

- The h-value can be computed with

$$h = 2\Phi\left(|\hat{E}^{(ncv)} - \widehat{\mathrm{b}}^{(ncv)}| \Big/ \sqrt{\widehat{mse}^{(ncv)}}\right)$$

One-sided h-values

- One-sided confidence interval $\tilde{C}_\alpha(\mathcal{D}) = (-\infty, U_{2\alpha}(\mathcal{D}))$.
- The one-sided h-value associated with this one-sided confidence interval is

$$\begin{aligned}\tilde{h}(\mathcal{D}) &= \inf\left\{\alpha \in (0,1) : U_{2\alpha}(\mathcal{D}) < 0\right\} \\ &= \Phi\left((\hat{E}^{(NCV)} - \widehat{\mathrm{b}}^{(NCV)}) \Big/ \sqrt{\widehat{MSE}^{(NCV)}} \right).\end{aligned}$$

Modified Outcome Methods

- An alternative is to use "modified outcomes" W_i

$$W_i = \begin{cases} Y_i / P\{A_i = 1 \mid \boldsymbol{x}_i\} & \text{if } A_i = 1 \\ -Y_i / P\{A_i = 0 \mid \boldsymbol{x}_i\} & \text{if } A_i = 0. \end{cases}$$

- The modified outcome W_i has expectation equal to $\theta(\boldsymbol{x}_i)$

$$E\{W_i \mid \boldsymbol{x}_i\} \quad = \quad f(1, \boldsymbol{x}_i) - f(0, \boldsymbol{x}_i).$$

- Estimate ITE directly by using W_i as outcomes.
- Do not need to estimate conditional mean function for both treatment arms.

Loss Function for Modified Outcomes

- The restricted estimator in the context of modified outcome methods is simply:

$$\tau^*_{MO} = \arg\min_{\tau} \sum_{i=1}^{n} (W_i - \tau)^2 = \bar{W}.$$

- To compare unrestricted $\hat{\theta}(x; \mathcal{D})$ and restricted estimators τ^*_{MO}, use the loss for a "future modified outcome" W_{n+1}:

$$\tilde{\ell}\Big(\hat{\theta}(\boldsymbol{x}_{n+1}; \mathcal{D}), \tau^*_{MO}, W_{n+1}\Big) = \Big(W_{n+1} - \hat{\theta}(\boldsymbol{x}_{n+1}; \mathcal{D})\Big)^2 - \Big(W_{n+1} - \tau^*_{MO}\Big)^2.$$

Loss Function for Modified Outcomes

- The **restricted estimator** in the context of modified outcome methods is simply:

$$\tau^*_{MO} = \arg\min_{\tau} \sum_{i=1}^{n} (W_i - \tau)^2 = \bar{W}.$$

- To compare unrestricted $\hat{\theta}(x; \mathcal{D})$ and restricted estimators τ^*_{MO}, use the loss for a "future modified outcome" W_{n+1}:

$$\tilde{\ell}\Big(\hat{\theta}(\boldsymbol{x}_{n+1}; \mathcal{D}), \tau^*_{MO}, W_{n+1}\Big) = \Big(W_{n+1} - \hat{\theta}(\boldsymbol{x}_{n+1}; \mathcal{D})\Big)^2 - \Big(W_{n+1} - \tau^*_{MO}\Big)^2.$$

Loss Function for Modified Outcomes

- You can show that

$$\begin{aligned}
\tilde{\theta}_{XY} &= E\left\{\tilde{\ell}\left(\hat{\theta}(\boldsymbol{x}_{n+1};\mathcal{D}), \tau^*_{MO}, W_{n+1}\right)\middle|\mathcal{D}\right\} \\
&= E\left\{\left(\theta(\boldsymbol{x}_{n+1}) - \hat{\theta}(\mathrm{x}_{n+1};\mathcal{D})\right)^2\middle|\mathcal{D}\right\} \\
&- E\left\{\left(\theta(\boldsymbol{x}_{n+1}) - \tau^*_{MO}\right)^2\middle|\mathcal{D}\right\}.
\end{aligned}$$

Simulations

- The main thing we want to verify is that the confidence intervals for performance difference have **good coverage** across a range of scenarios.
- h-values should not be consistently small in cases with no predictive advantages of HTE.
- The second thing we would want is to have reasonable power to detect cases where incorporating HTE does improve predictive performance.

Simulation 1: A Linear Model

- Generated outcomes Y_i from the following model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 A_i + \beta_4 A_i x_{i1} + \beta_5 A_i x_{i2} + \varepsilon_i$$

- *Setting A*: $\beta_4 = 0$ and $\beta_5 = 0$, (no HTE)
- *Setting B*: $\beta_4 = 0.5$ and $\beta_5 = -2$, (HTE is present)

Simulation 1: Setting A

n	Simulation Setting	Method	θ_{XY}	Coverage Proportion	Median h-value
100	A	linear	0.022	0.986	0.637
		Lasso	0.012	0.980	0.838
		boosting	0.017	0.992	0.657
500	A	linear	0.004	0.986	0.609
		Lasso	0.001	0.972	0.871
		boosting	0.006	0.980	0.432

Simulation 1: Setting B

n	Simulation Setting	Method	θ_{XY}	Coverage Proportion	Median h-value
100	B	linear	-1.10	0.950	0.001
		Lasso	-1.25	0.952	0.002
		boosting	-1.08	0.948	0.001
500	B	linear	-1.07	0.954	< 0.001
		Lasso	-1.13	0.964	< 0.001
		boosting	-1.05	0.950	< 0.001

Simulation 2: Nonlinear Conditional Mean Function

- Generated outcomes Y_i from the following model

$$Y_i = \mu(\boldsymbol{x}_i) + A_i\theta(\boldsymbol{x}_i) + \varepsilon_i,$$

- Three possible choices of baseline risk function $\mu(\boldsymbol{x}_i)$.
- Four possible choices of ITE function $\theta(\boldsymbol{x}_i)$.
- ITE functions $\theta_3(\boldsymbol{x})$ and $\theta_4(\boldsymbol{x})$ exhibit HTE.
- For example,

$$\begin{aligned}
\mu_3(\mathrm{x}_i) &= \frac{1}{2}(x_1^2 + x_2 + x_3^2 + x_4 + x_5^2 + x_6 + x_7^2 + x_8 + x_9^2 - 11), \\
\theta_3(\mathrm{x}_i) &= 2 + 0.1/\{1 + \exp(-x_{i2})\},
\end{aligned}$$

Simulation 2 Results (Linear Regression)

$\mu(\cdot)$	$\theta(\cdot)$	θ_{XY}	Coverage Proportion	CI Width	Median* h-value
			n = 500		
μ_1	θ_1	2.262	0.992	13.919	0.773
μ_1	θ_4	4.709	0.974	36.289	0.714
μ_2	θ_1	0.019	0.986	0.105	0.781
μ_2	θ_4	-0.182	0.984	7.818	0.477
μ_3	θ_1	0.019	0.990	0.105	0.785
μ_3	θ_4	-0.104	0.984	8.415	0.488
			n = 2000		
μ_1	θ_1	0.236	0.992	1.418	0.771
μ_1	θ_4	-0.694	0.984	4.342	0.280
μ_2	θ_1	0.002	0.992	0.009	0.799
μ_2	θ_4	**-1.139**	0.970	1.412	**0.001**
μ_3	θ_1	0.002	0.990	0.009	0.794
μ_3	θ_4	**-1.138**	0.954	1.387	**0.001**

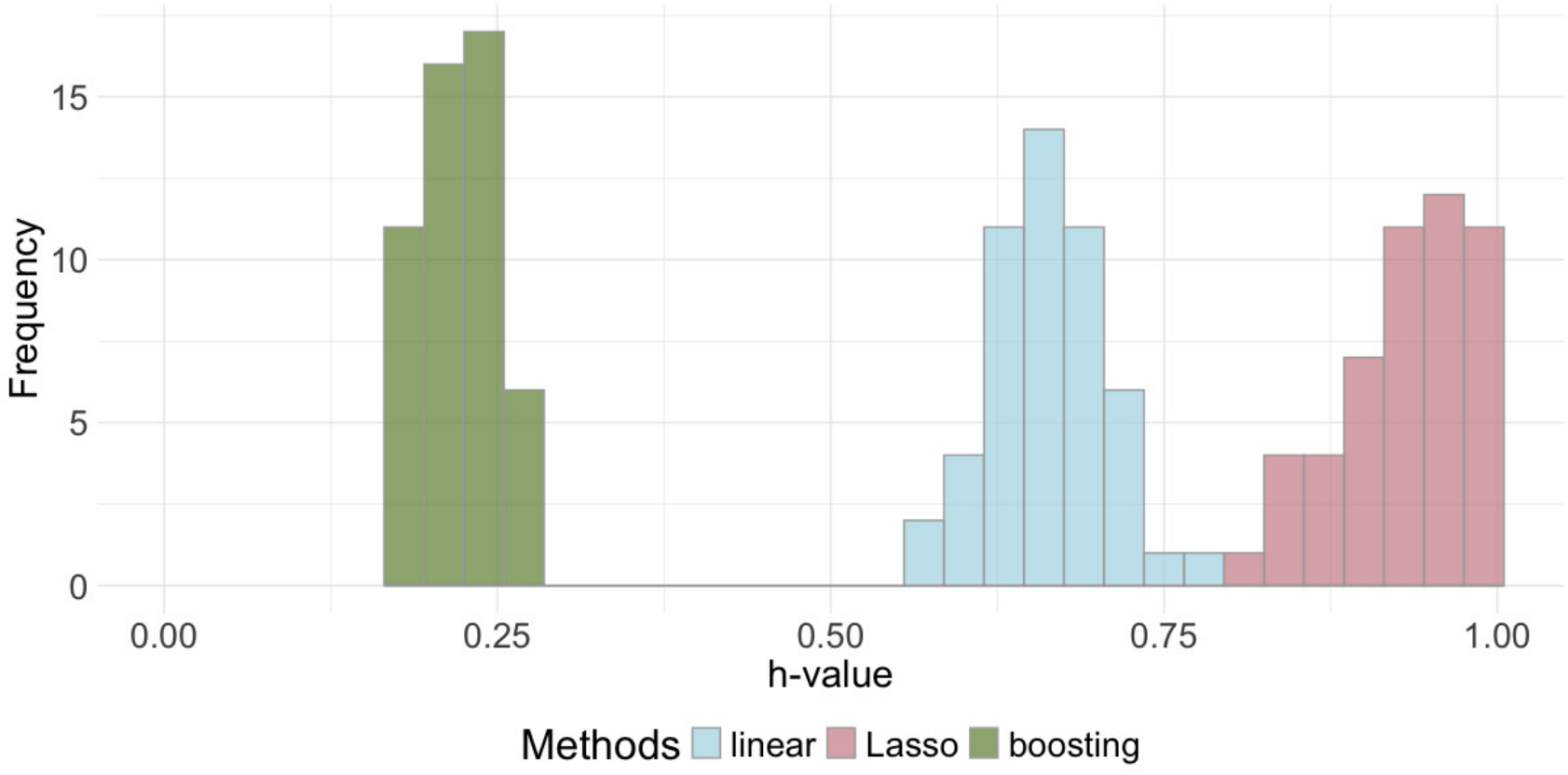
Simulation 2 Results (Boosting)

$\mu(\cdot)$	$\theta(\cdot)$	θ_{XY}	Coverage Proportion	CI Width	Median* h-value
		n = 500			
μ_1	θ_1	0.359	0.984	7.283	0.646
μ_1	θ_4	1.283	0.972	24.508	0.615
μ_2	θ_1	0.001	0.984	0.033	0.510
μ_2	θ_4	-0.433	0.982	6.064	0.425
μ_3	θ_1	0.001	0.978	0.040	0.525
μ_3	θ_4	-0.184	0.978	6.337	0.480
		n = 2000			
μ_1	θ_1	0.037	0.992	0.713	0.642
μ_1	θ_4	-0.958	0.980	3.431	0.154
μ_2	θ_1	0.000	0.980	0.003	0.358
μ_2	θ_4	**-0.966**	0.968	1.137	$<$ **0.001**
μ_3	θ_1	0.000	0.990	0.003	0.342
μ_3	θ_4	**-0.764**	0.956	1.180	**0.005**

Application: The IHDP Study

- The Infant Health and Development Program (IHDP).
- This program was a randomized trial focused on evaluating different interventions for low-birthweight premature infants.
- Treatment group received high-quality child care and home visits from a trained provider at eight clinical sites.
- **Outcome:** measure of cognitive performance at 36 months.
- $n = 908$
- $p = 10$

Application: The IHDP Study (h-values)



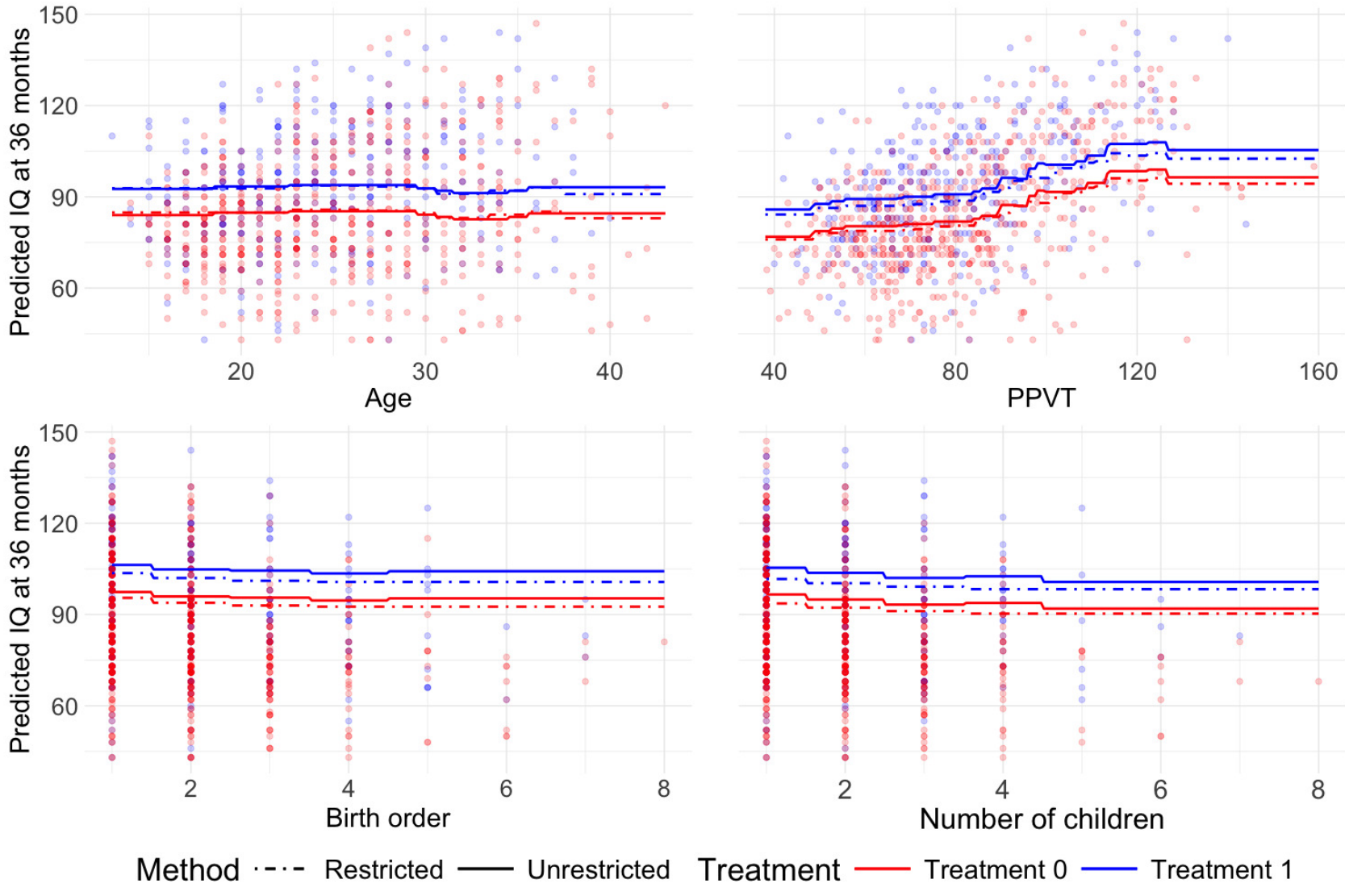
The IHDP Study (Partial Dependence Plots)

- Plotting partial dependence functions is a useful way to the role of certain covariates.
- For the conditional mean function $f(A, \mathrm{x})$, we define the partial dependence function for covariate k at (a, u) for the unrestricted estimator, as

$$\rho_k(a, u) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}\big(a, (u, \mathrm{x}_{i,-k}); \mathcal{D}\big),$$

- The notation $(\mathrm{x}_{i,-k}, u)$ denotes the vector where $x_{ik} = u$ and the remaining covariates are set equal to their observed values.

The IHDP Study (Partial Dependence Plots)



Concluding Points

- Coverage based on this nested cross-validation procedure is quite good for most procedures, but could potentially be improved with other approaches.
- As of now, this approach applies to continuous outcomes (extensions to binary or survival outcomes would be useful).
- Description and evaluation of our approach focused on randomized treatment assignment. Extensions to observational settings should be straightforward, but have not been done yet.

Thanks