Uncovering Causal Relationships in Sports Analytics: Methods and Applications

Shinpei Nakamura-Sakai

Yale University, Department of Statistics and Data Science

* Parts of this presentation are based on slides by Jasjeet Sekhon (edited for this workshop)

S&DS

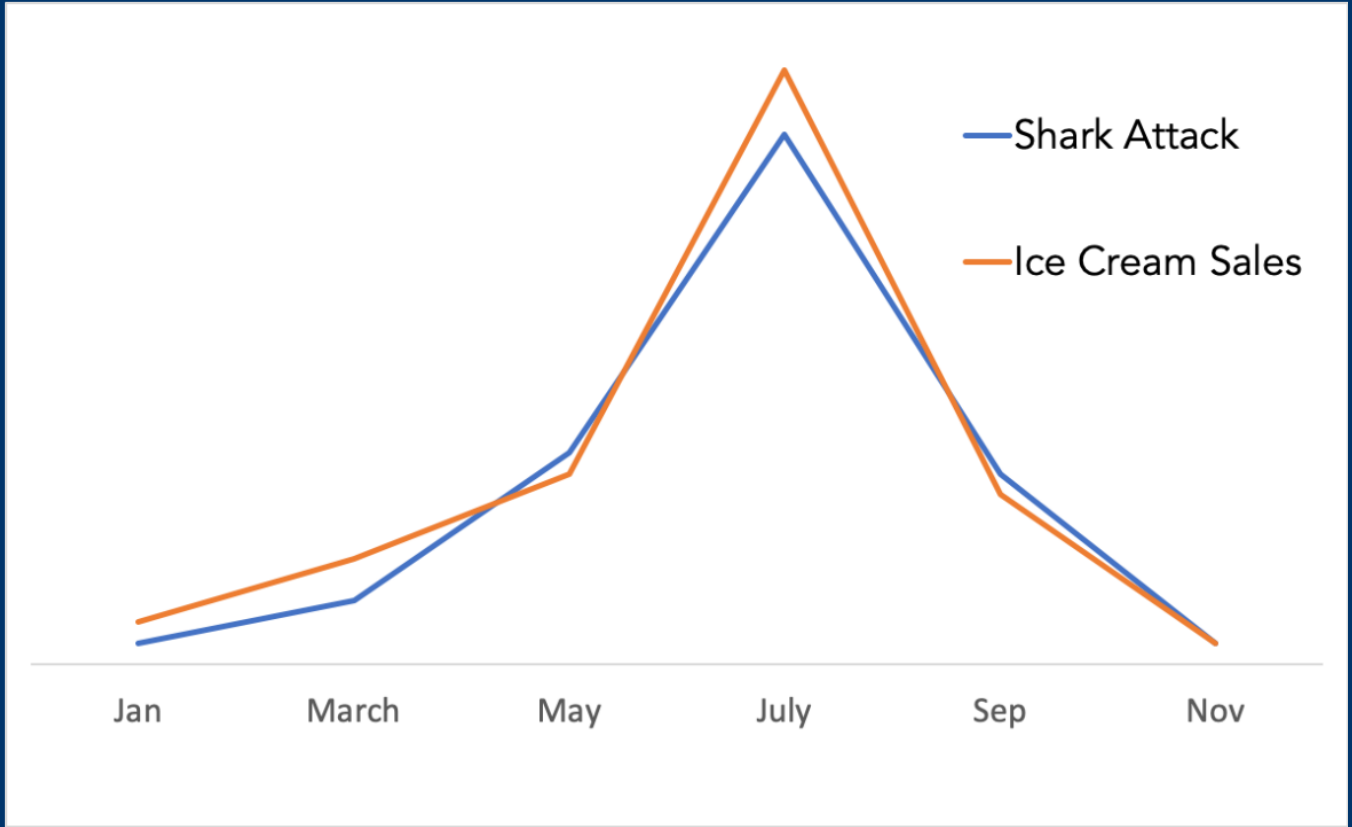
Yale SCHOOL OF PUBLIC HEALTH

Overview

- What is causal inference?
- Why causal inference in sport?
- How to combine causal inference and ML?
- What is an age-curve?
- Coding:
  - Which causal ML model is the most appropriate for our study?
  - How to validate?

Causal Inference

- Correlation vs causation
  - Does eating ice cream cause to be attacked by shark?

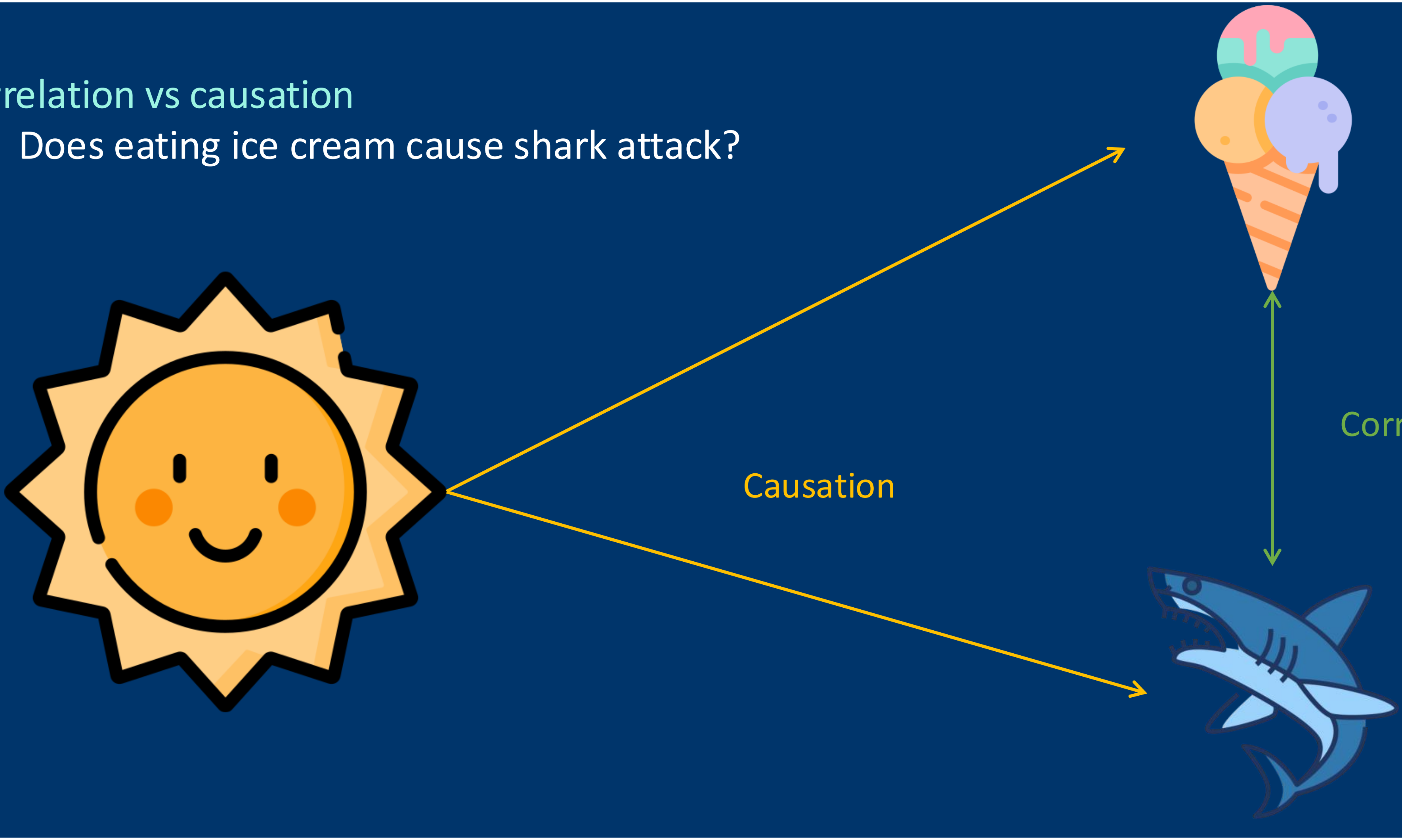


S&DS Yale University
Graduate School of Arts & Sciences

Reference: www.linkedin.com/pulse/understanding-difference-between-correlation-shark-candela/

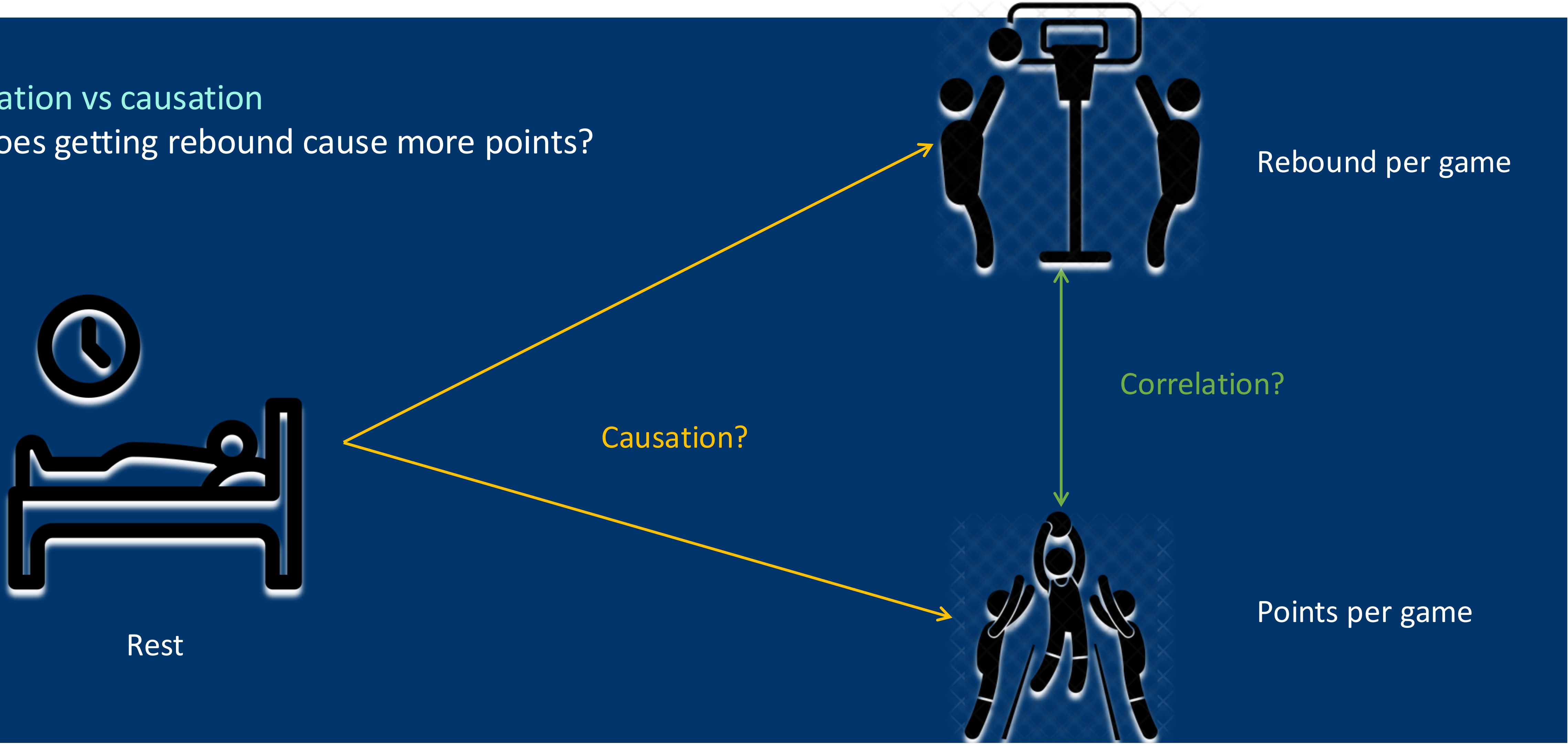
Why causal inference?

- Correlation vs causation
  - Does eating ice cream cause shark attack?



Why causal inference?

- Correlation vs causation
  - Does getting rebound cause more points?



Rebound per game

Correlation?

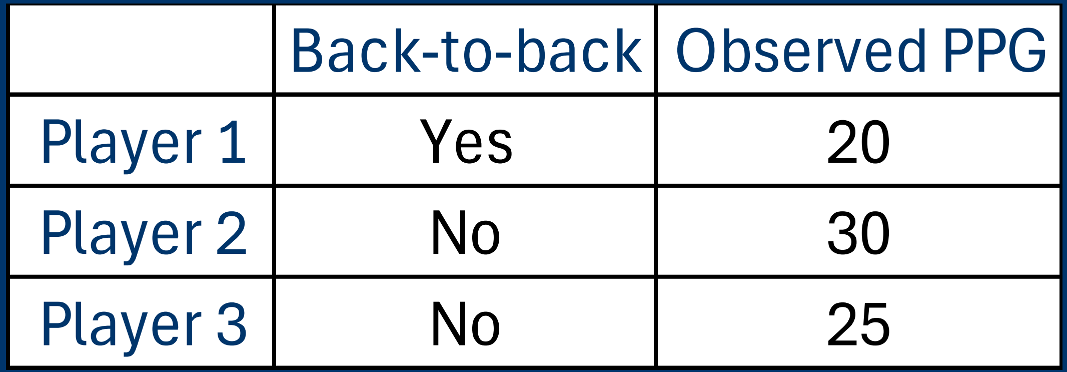
Causation?

Rest

Points per game

Causal Inference

- Example: 3 players, Resting at least one day is the treatment, outcome is Points per game (PPG)



	Back-to-back	Observed PPG
Player 1	Yes	20
Player 2	No	30
Player 3	No	25

- Question: Is resting improve PPG?

Causal Inference

- Example: 3 players, Resting at least one day is the treatment, outcome is Points per game (PPG)

	Back-to-back	Observed PPG
Player 1	Yes	20
Player 2	No	30
Player 3	No	25

- Question: Does resting improve PPG?
    - Maybe but maybe not. Let’s use the potential outcome framework

	Back-to-back	Observed PPG	Potential Outcome Y(0)	Potential Outcome Y(1)
Player 1	Yes	20	?	20
Player 2	No	30	30	?
Player 3	No	25	25	?

Causal Inference

- Example: 3 players, Resting at least one day is the treatment, outcome is Points per game (PPG)

	Back-to-back	Observed PPG
Player 1	Yes	20
Player 2	No	30
Player 3	No	25

- Question: Does resting improve PPG?
  - Maybe but maybe not. Let’s use the potential outcome framework
  - For example, in this case, treatment effect is zero, but there is a Fundamental Problem of Causal inference

	Back-to-back	Observed PPG	Potential Outcome Y(0)	Potential Outcome Y(1)
Player 1	Yes	20	20	20
Player 2	No	30	30	30
Player 3	No	25	25	25

Fundamental Problem of Causal Inference

- Fundamental Problem of Causal Inference: We observe only one version of potential outcome!!

- If Y(1) (player performance when rested) is observed, then Y(0) (player performance without rest) cannot be observed or it’s a missing value. Causal Inference is a missing data problem!

- Formally, given T_i, treatment for individual i, the observed outcome can be expressed as
    - $Y_i^{obs} = T_i Y_i(1) + (1 - T_i) Y_i(0)$

- And the individual treatment effect of interest for individual i is
    - $\tau_i = Y_i(1) - Y_i(0)$

- However, due to Fundamental Problem of Causal Inference we only observe either $Y(1)$ or $Y(0)$.

- What can we do to estimate the treatment effect?

Average Treatment Effect Estimation

- Instead, we can estimate the Average Treatment Effect (ATE) for finite sample
  - $ATE_{FS} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$

- But again, this is not observable as we only observe one of the outcomes ☹

- What would be a good estimator for this estimand?

Average Treatment Effect Estimation

- Instead, we can estimate the Average Treatment Effect (ATE) for finite sample
  - $ATE_{FS} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$
- But again this is not observable as we only observe one of the outcomes ☹
- What would be a good estimator for this estimand?
  - $\widehat{ATE}_{FS} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{obs} T_i - \frac{1}{N}\sum_{i=1}^{N} Y_i^{obs}(1 - T_i)$
- We can prove that $\mathbb{E}_T[ATE_{FS}] = ATE_{FS}$ under some assumptions
  - Homogeneity: difference of POs are constant across units
  - Random Treatment assignment
  - SUTVA: No interference and consistency
- Is ATE enough for sports analytics?

Heterogeneous Treatment Effect (HTE) Estimation

- We might be interested on the treatment effect for specific group of players, not just the average.

- Let’s consider T_i to be the whether the player i did not play a game the day before. Namely, $T_i = 0$ implies, playing back-to-back games

- Does homogeneity hold for every players?
  - Probably, no. Rest might affect more older players than younger players

- Other examples
  - Shooting practice program: This might affect differently through positions, point guard and center does not get the same treatment effect, especially if center does not shoot a lot, i.e. Shaq
  - Going for 4th down: Win percentage increase 0.5% on ATE. However, HTE measure the win percentage depending of X-yard line and score differential (Ex. 4th and 2 on opponents 45-yard line down by 3 (Q4))

Conditional Average Treatment Effect (CATE)

- Conditional Average Treatment Effect (CATE): Average Treatment effect for a specific subgroup or given a set of covariates
  - $CATE(x_i) \coloneqq \tau(x_i) \coloneqq \mathbb{E}[D_i|X = x_i] = \mathbb{E}[Y(1) - Y(0)|X = x_i]$
- x_i can be high dimensional, such as center with age greater than 30
- Let $\hat{\tau}$ be estimator for D then the MSE is defined as $\mathbb{E}[(D_i - \hat{\tau})^2|X_i = x_i]$ by bias-variance decomposition we have
  - $\mathbb{E}[(D_i - \hat{\tau})^2|X_i = x_i]=\underbrace{\mathbb{E}\left[\left(D_i - \tau(x_i)\right)^2\middle|X_i = x_i\right]}_{\text{Approximation error}} + \underbrace{\mathbb{E}[(\tau(x_i) - \hat{\tau})^2|X_i = x_i]}_{\text{Estimation error}}$
- How do we shrink each of these errors?
  - Causal Inference 🤝 Machine Learning

How do we estimate CATE?

- Meta-learners: A meta-learner decomposes the problem of estimating the CATE into several sub-regression problems. The estimator which solve those sub-regression problems are called base-learners
  - We can use ANY (regression, ML, statistical) model for base-learner
    - Neural nets, random forest, XGBoost, regressions etc.
    - Compared to traditional causal inference which is heavily based on regressions, those ML models has higher accuracy hence reducing the Estimation error
- But how do we combine those base learners?

How do we estimate CATE?

- $CATE(x_i) \coloneqq \tau(x_i) \coloneqq \mathbb{E}[D_i|X = x_i] = \mathbb{E}[Y(1) - Y(0)|X = x_i]$
  - By linearity of expectation, $CATE(x_i)$= $\mathbb{E}[Y(1)|X = x_i] - \mathbb{E}[Y(0)|X = x_i] \coloneqq \mu_1(x_i) - \mu_0(x_i)$

- It's all about how to estimate μ_1 and μ_0

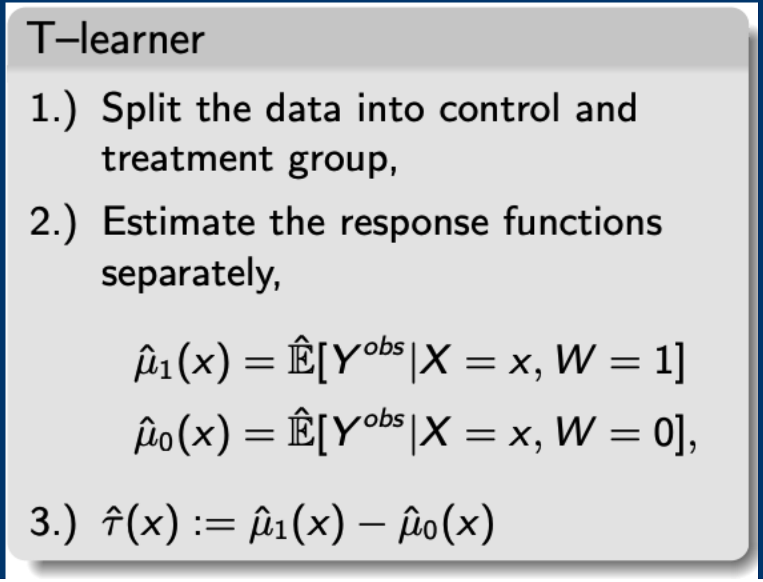
- What is the most intuitive way?

How do we estimate CATE?

- $CATE(x_i) \coloneqq \tau(x_i) \coloneqq \mathbb{E}[D_i|X = x_i] = \mathbb{E}[Y(1) - Y(0)|X = x_i]$
  - By linearity of expectation, $CATE(x_i)$= $\mathbb{E}[Y(1)|X = x_i] - \mathbb{E}[Y(0)|X = x_i] \coloneqq \mu_1(x_i) - \mu_0(x_i)$

- It's all about how to estimate μ_1 and μ_0

- What is the most intuitive way?



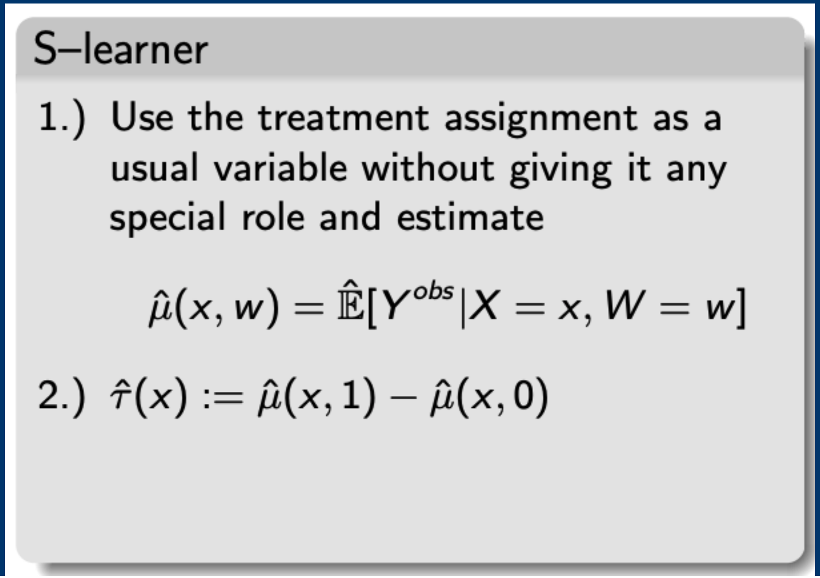
T–learner

1.) Split the data into control and treatment group,

2.) Estimate the response functions separately,

$$\hat{\mu}_1(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 1]$$
$$\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 0],$$

3.) $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$



S–learner

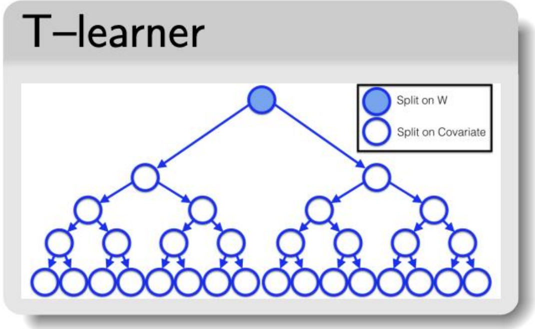
1.) Use the treatment assignment as a usual variable without giving it any special role and estimate

$$\hat{\mu}(x, w) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = w]$$

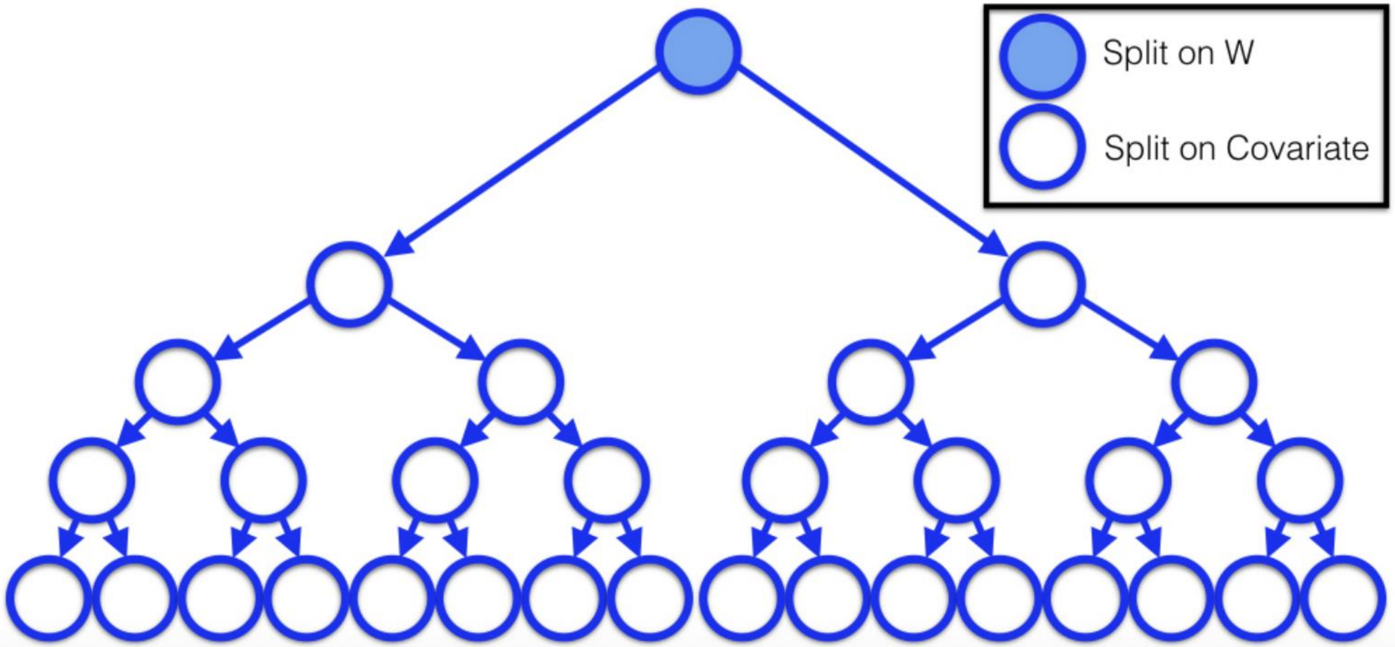
2.) $\hat{\tau}(x) := \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$

How do we estimate CATE?

- T-learner

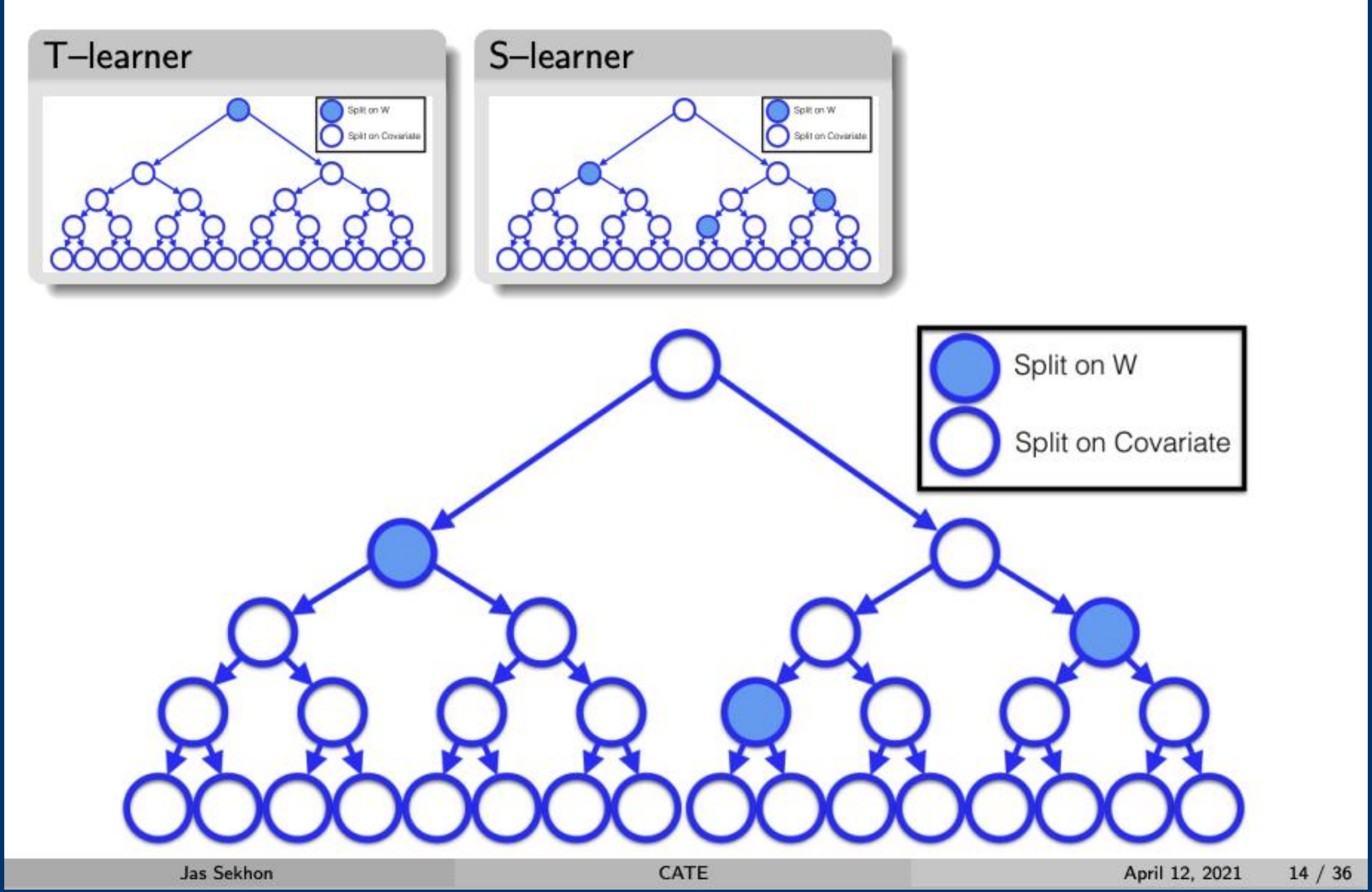


T–learner



How do we estimate CATE?

- S-learner



T–learner

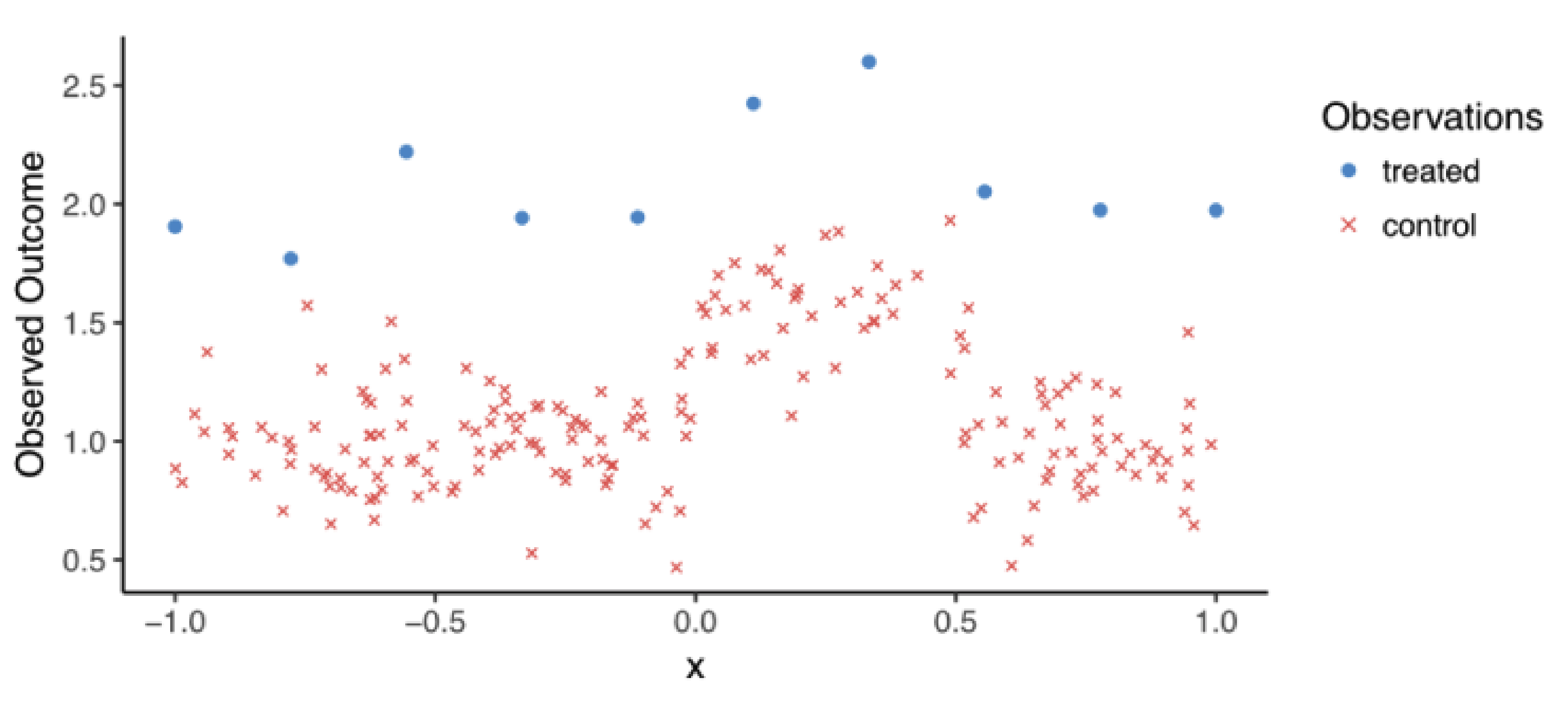
S–learner

Jas Sekhon CATE April 12, 2021 14 / 36

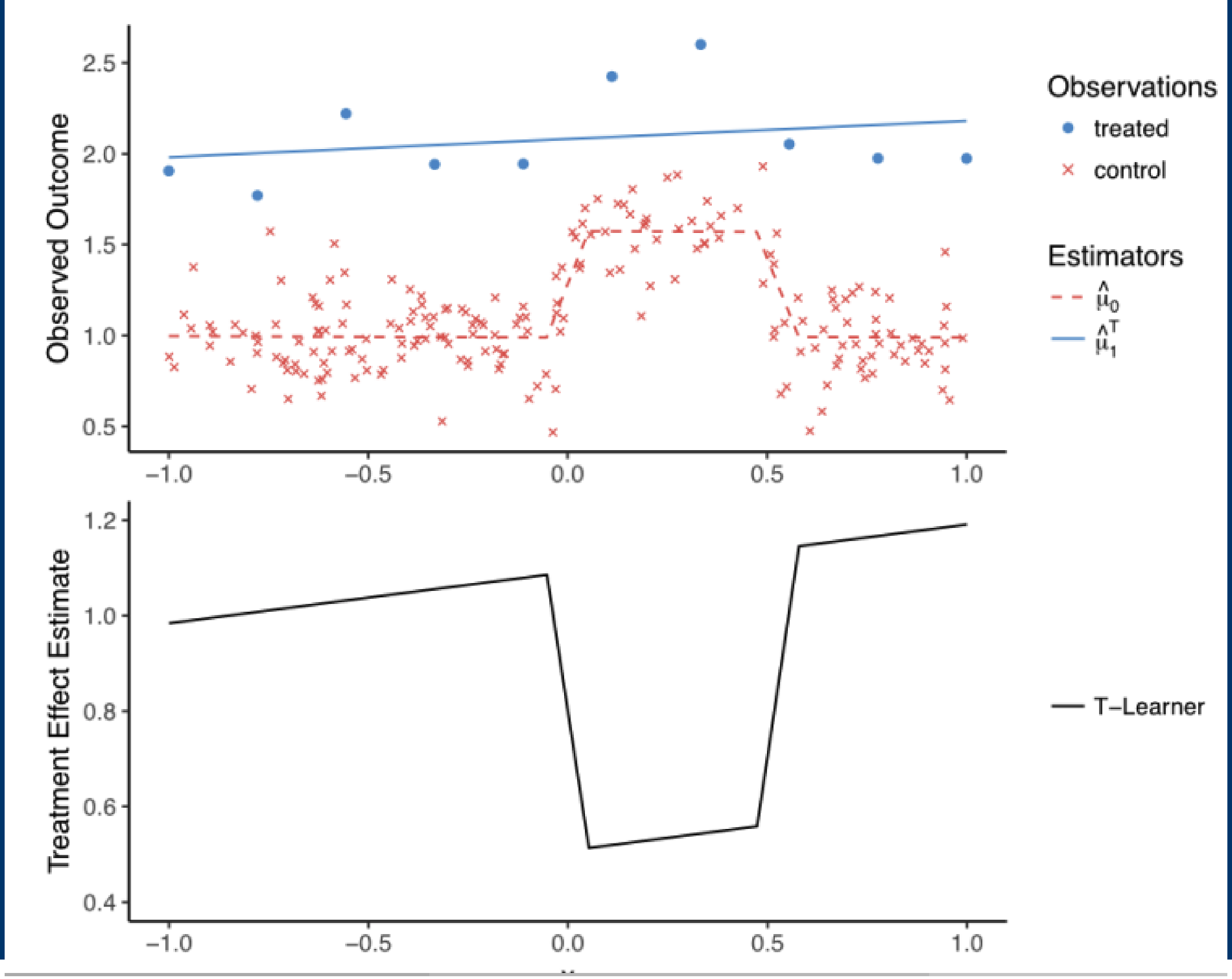
- Are these learners enough?

X-learner Motivation

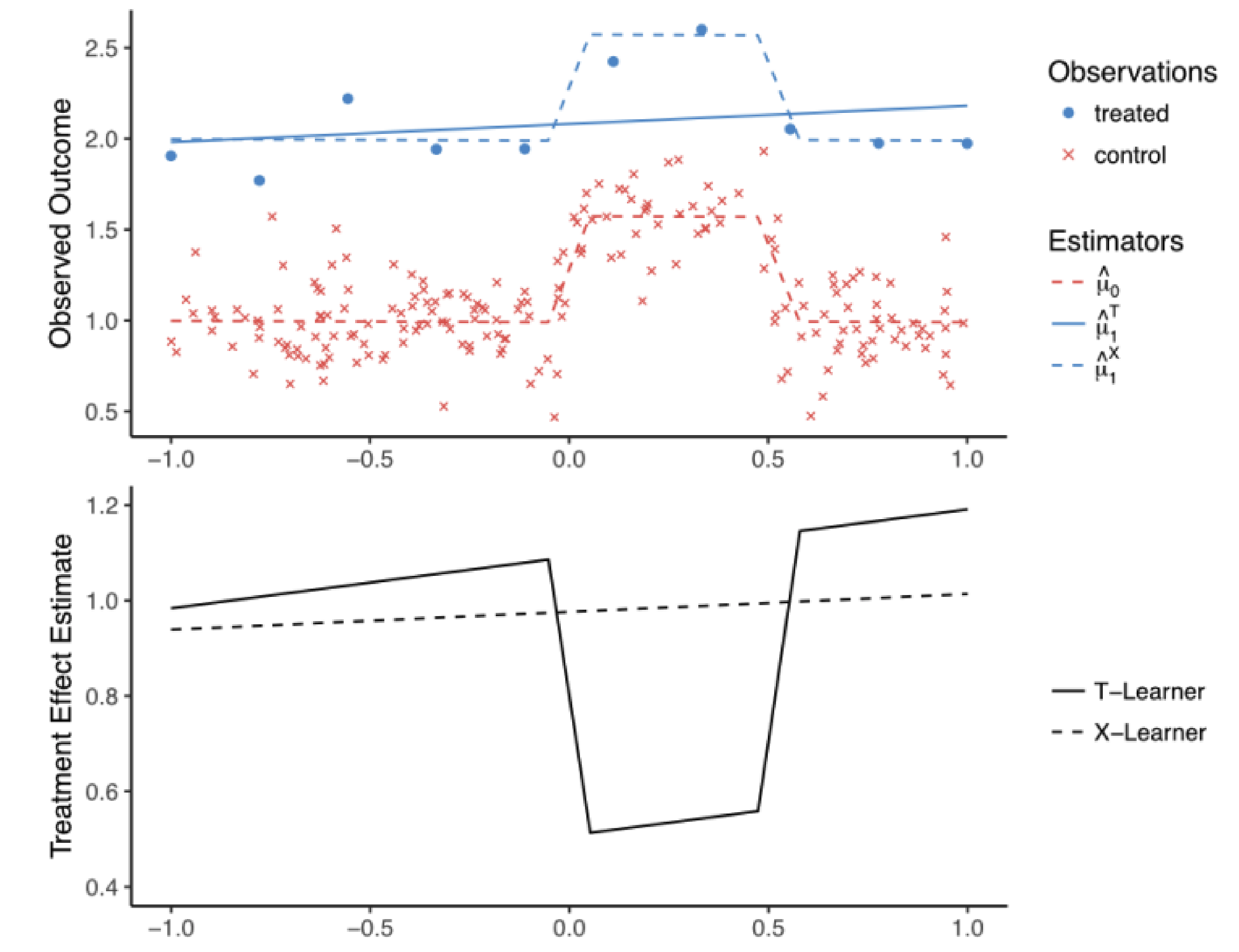
- What happen if if fit T-learner on this data?



X-learner Motivation



X-learner Motivation

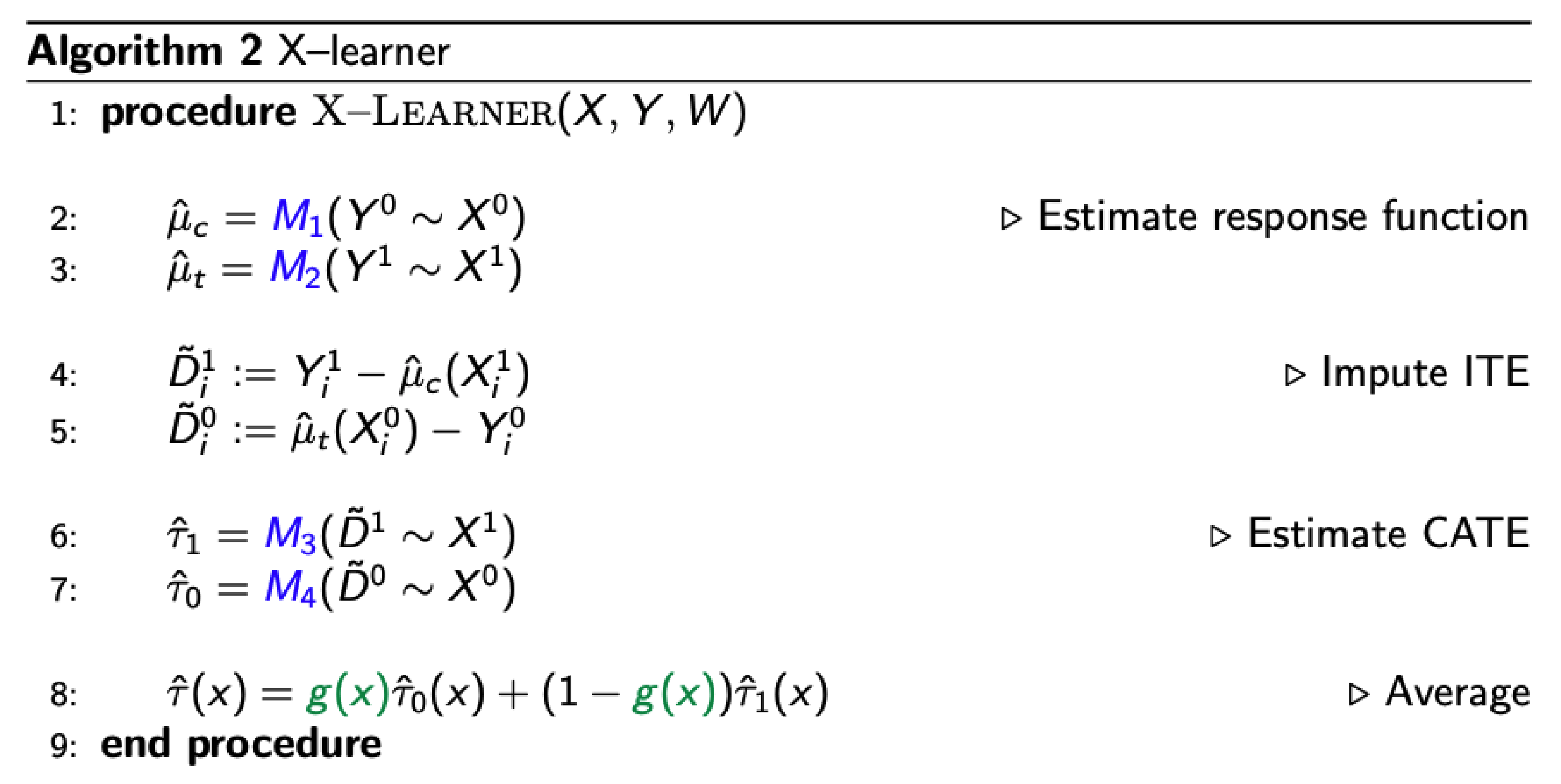


X-learner

- X-learner: X-learner crosses the information of treatment and control groups. Using total of four (or five) models

- Which model will be better? $\hat{\mu}_c$ or $\hat{\mu}_t$
  - Depends on data size
  - Complexity of M_1 and M_2
  - Regularly, which is easier to estimate? M_1/M_2 or M_3/M_4

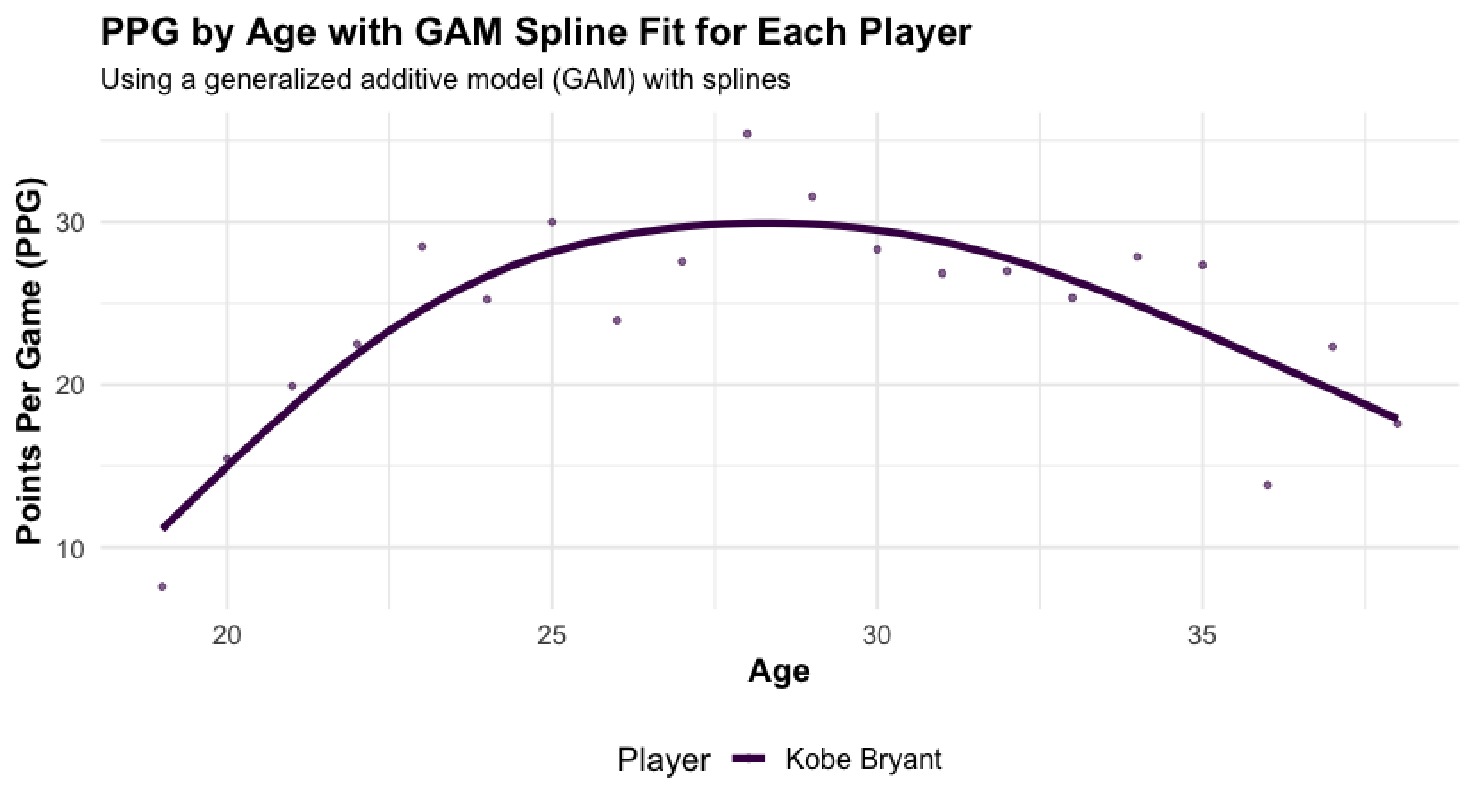
- How do we estimate uncertainty?



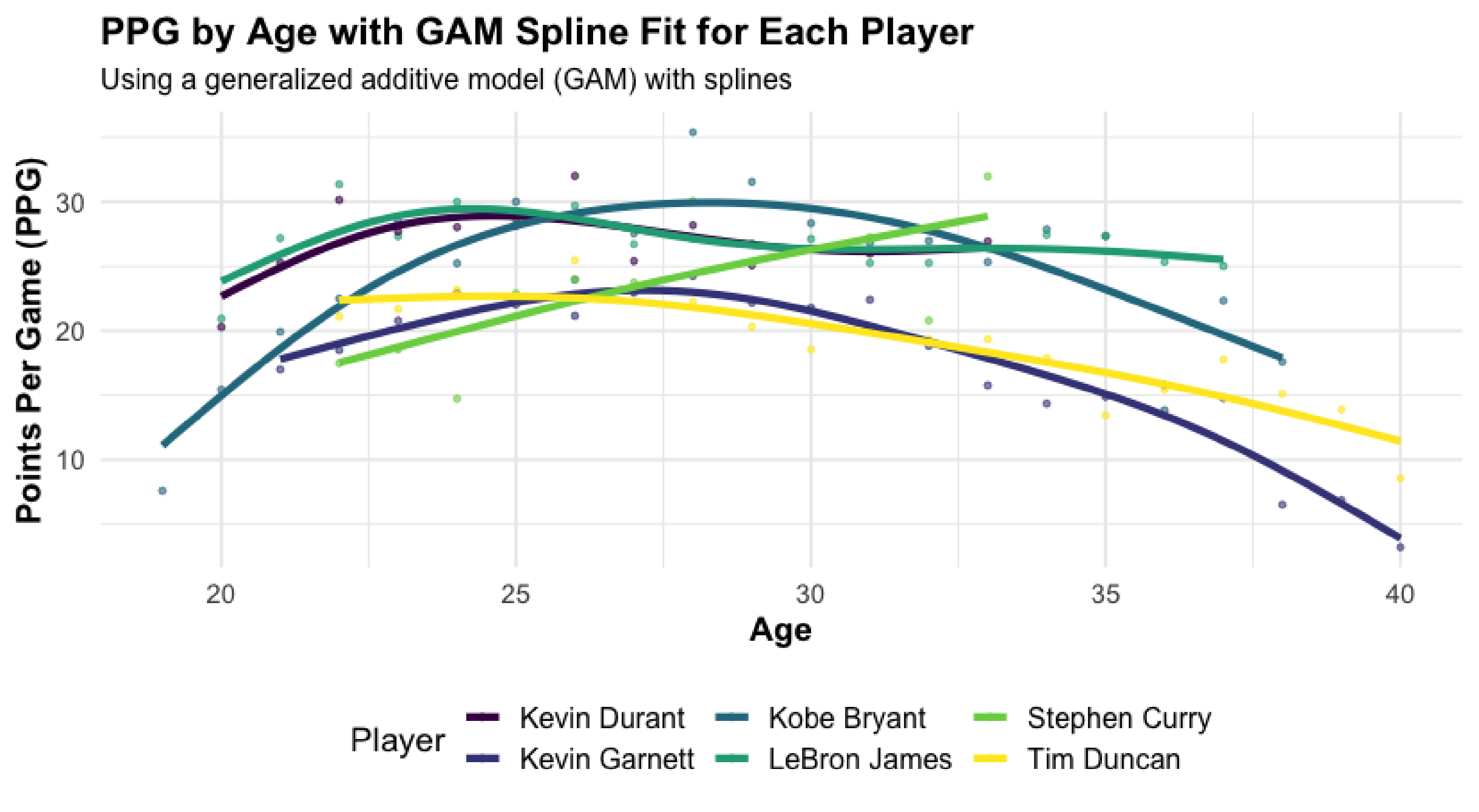
Algorithm 2 X–learner

```
procedure X–LEARNER($X, Y, W$)

    $\hat{\mu}_c = M_1(Y^0 \sim X^0)$                              ▷ Estimate response function
    $\hat{\mu}_t = M_2(Y^1 \sim X^1)$

    $\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_c(X_i^1)$                  ▷ Impute ITE
    $\tilde{D}_i^0 := \hat{\mu}_t(X_i^0) - Y_i^0$

    $\hat{\tau}_1 = M_3(\tilde{D}^1 \sim X^1)$                     ▷ Estimate CATE
    $\hat{\tau}_0 = M_4(\tilde{D}^0 \sim X^0)$

    $\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$   ▷ Average
end procedure
```

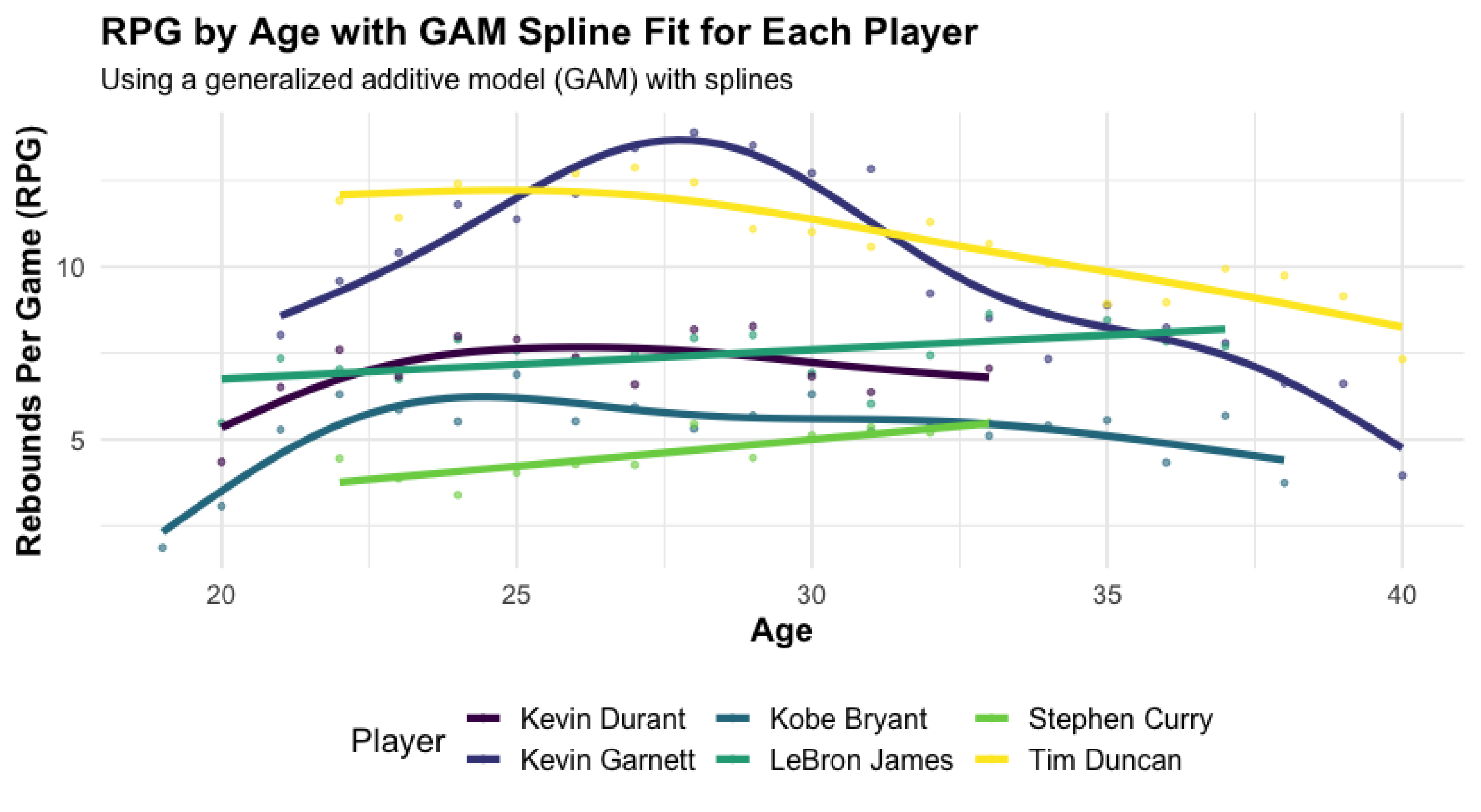
What is an age-curve?



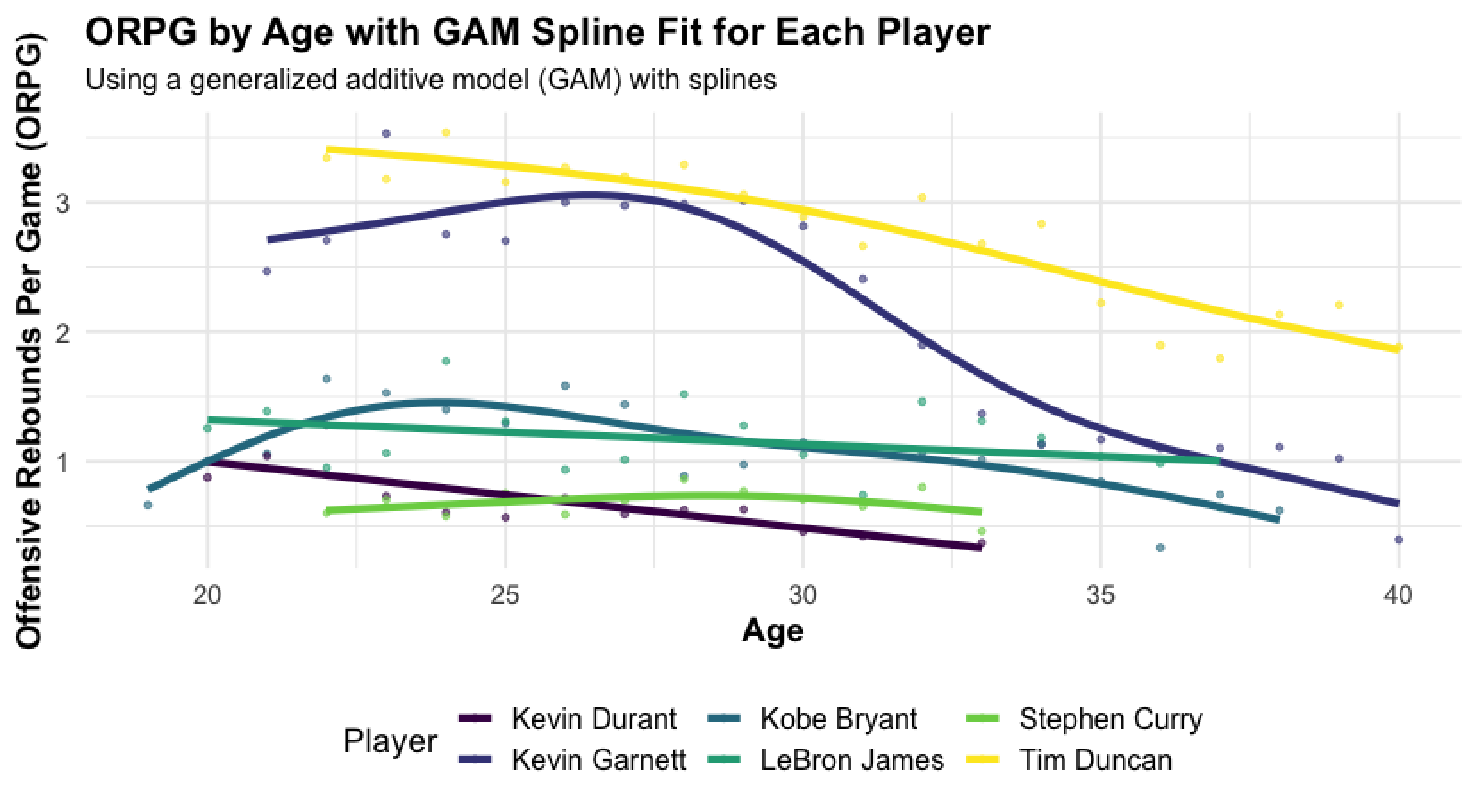
Age-Curves are heterogeneous



Age-Curves are heterogeneous

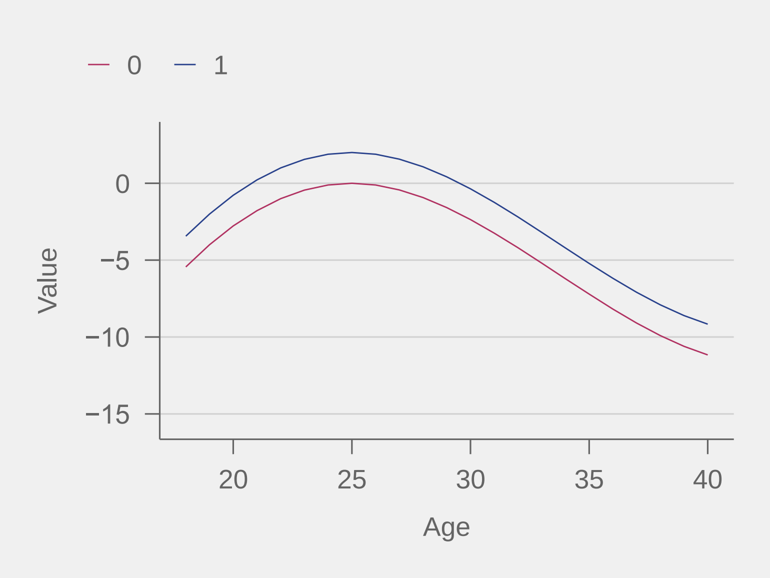


Age-Curves are heterogeneous

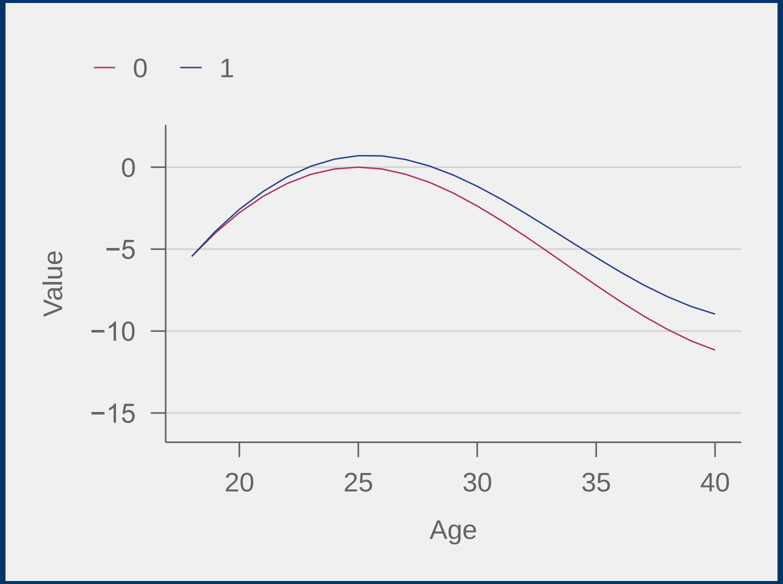


Simulations Scenarios

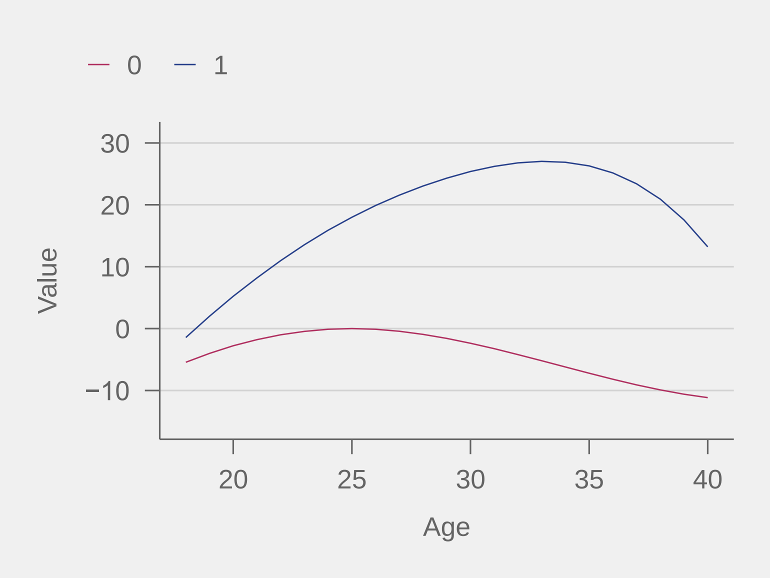
Scenario 1: Constant treatment effect



Scenario 2: Linear treatment effect



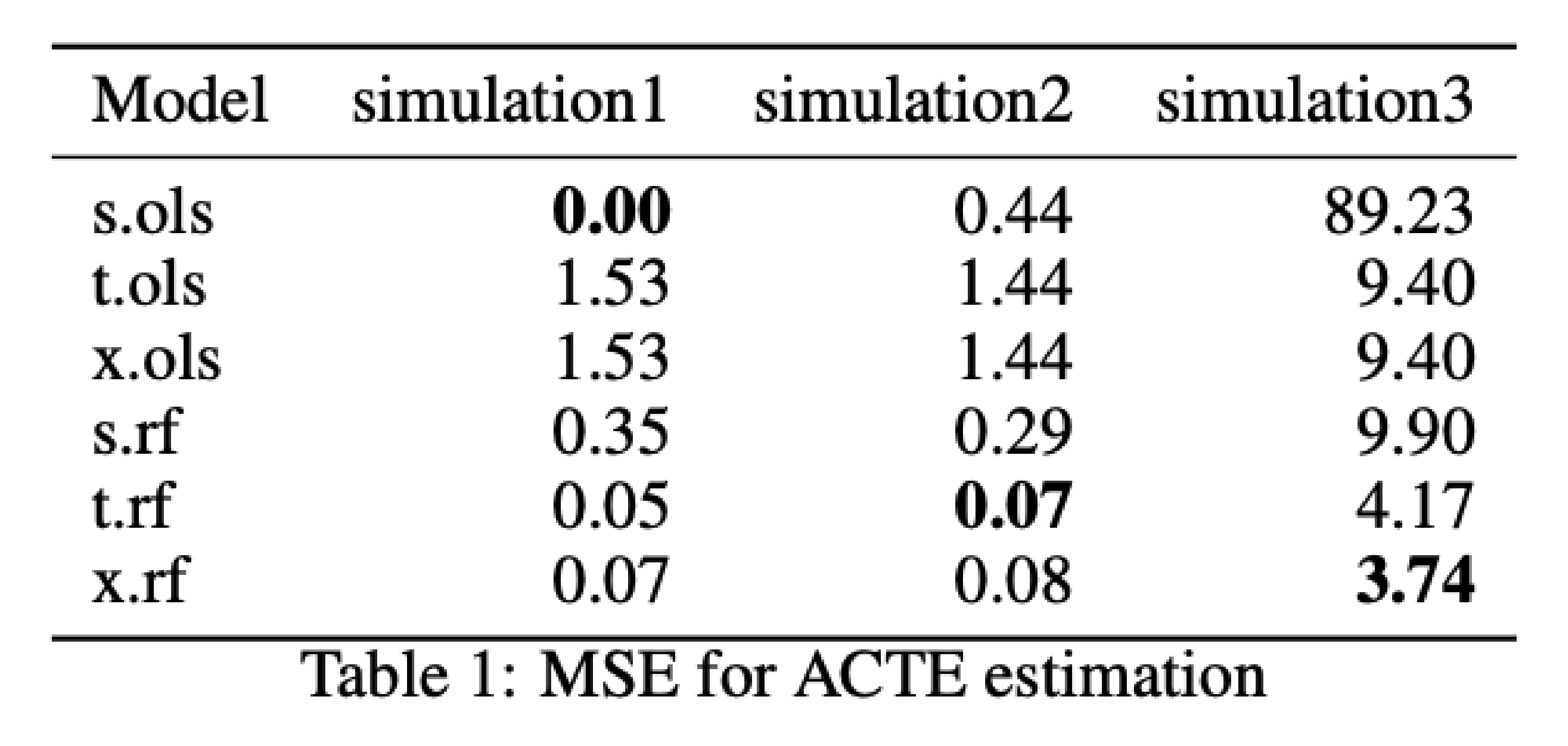
Scenario 3: Non-linear treatment effect



Rmarkdown

Simulations Results

- Flexibility is good but which learner and estimator should we choose?



Model	simulation1	simulation2	simulation3
s.ols	**0.00**	0.44	89.23
t.ols	1.53	1.44	9.40
x.ols	1.53	1.44	9.40
s.rf	0.35	0.29	9.90
t.rf	0.05	**0.07**	4.17
x.rf	0.07	0.08	**3.74**

Table 1: MSE for ACTE estimation

Framework

- Conditional Expectation Function (CEF)

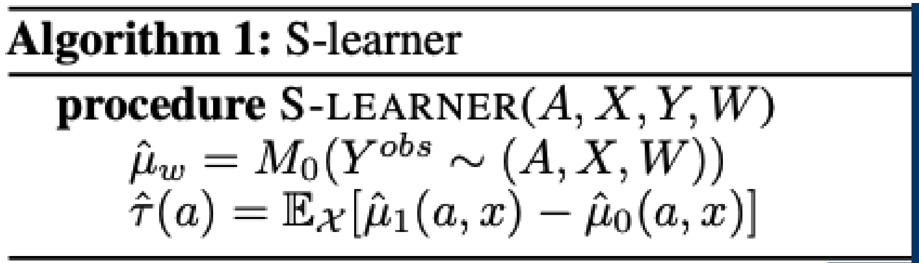
  - Consider N units, age $A \in \mathbb{Z}$, treatment $W \in \{0,1\}$, $Y(w)$ potential outcome of unit i for treatment w Then the CEF is defined as:
    $\mu_0(a,x) \coloneqq \mathbb{E}[Y_i(0)|A_i = a, X_i = x]$ and $\mu_1(a,x) \coloneqq \mathbb{E}[Y_i(1)|A_i = a, X_i = x]$

  - Under this framework, our main causal estimand of interest is the ACTE, that is:
    $\tau(a) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0)|A = a] = \mathbb{E}_{\mathcal{X}}[\mu_1(a,x) - \mu_0(a,x)]$

  - Takeaway: μ_0 is the expectation for the control group and μ_1 is for treatment group and $\tau(\mathrm{a})$ is the difference at age a

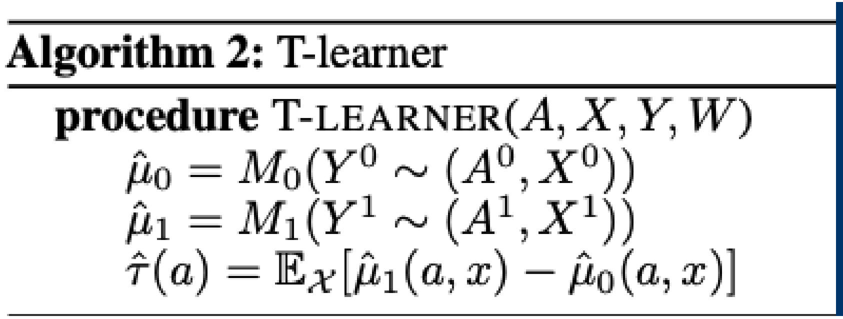
Estimation of ACTE

- To estimate the ACTE/$\tau(a)$ we use three meta-learning framework by Künzel et. al (2019) where it allows us to use ANY predictive model (or any supervised ML model)
- S-learner: S-learner uses a `single’ model
- T-learner: T-learner uses `two’ models, one for treatment and one for control group
- X-learner: X-learner `crosses’ the information of treatment and control groups. Using total of `four’ models



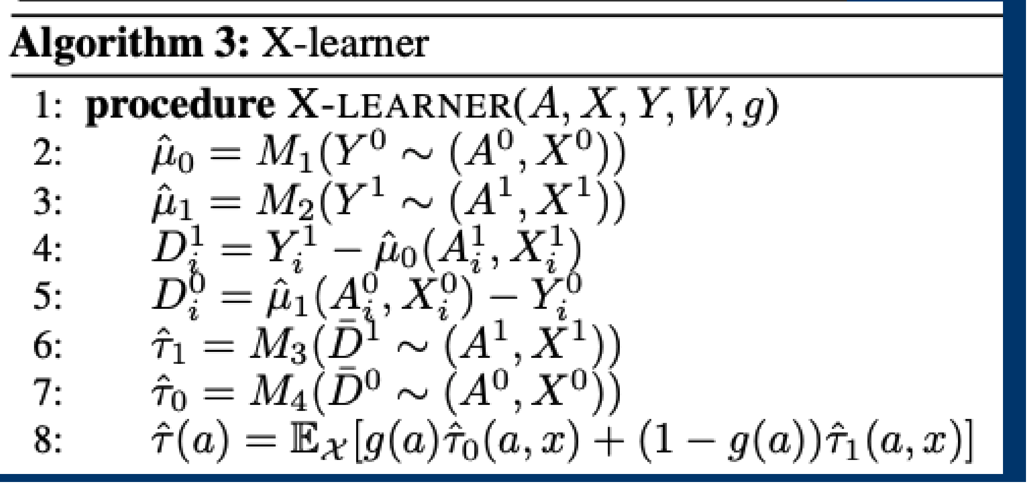
Algorithm 1: S-learner

```
procedure S-LEARNER($A, X, Y, W$)
    $\hat{\mu}_w = M_0(Y^{obs} \sim (A, X, W))$
    $\hat{\tau}(a) = \mathbb{E}_{\mathcal{X}}[\hat{\mu}_1(a, x) - \hat{\mu}_0(a, x)]$
```



Algorithm 2: T-learner

```
procedure T-LEARNER($A, X, Y, W$)
    $\hat{\mu}_0 = M_0(Y^0 \sim (A^0, X^0))$
    $\hat{\mu}_1 = M_1(Y^1 \sim (A^1, X^1))$
    $\hat{\tau}(a) = \mathbb{E}_{\mathcal{X}}[\hat{\mu}_1(a, x) - \hat{\mu}_0(a, x)]$
```

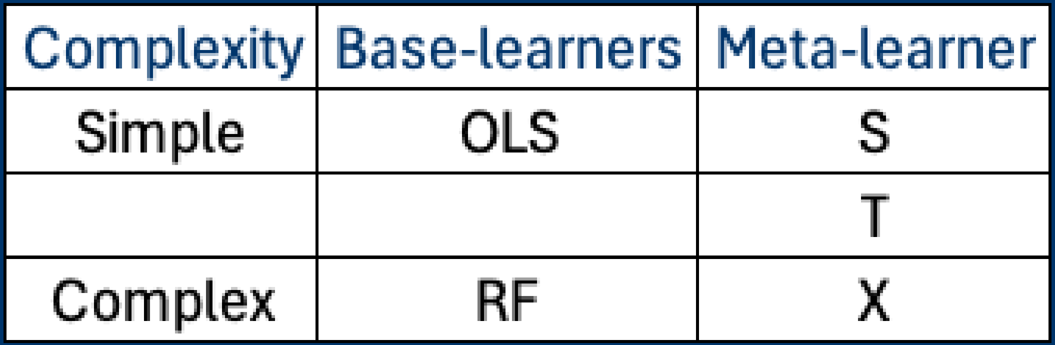


Algorithm 3: X-learner

```
procedure X-LEARNER($A, X, Y, W, g$)
    $\hat{\mu}_0 = M_1(Y^0 \sim (A^0, X^0))$
    $\hat{\mu}_1 = M_2(Y^1 \sim (A^1, X^1))$
    $D_i^1 = Y_i^1 - \hat{\mu}_0(A_i^1, X_i^1)$
    $D_i^0 = \hat{\mu}_1(A_i^0, X_i^0) - Y_i^0$
    $\hat{\tau}_1 = M_3(\bar{D}^1 \sim (A^1, X^1))$
    $\hat{\tau}_0 = M_4(\bar{D}^0 \sim (A^0, X^0))$
    $\hat{\tau}(a) = \mathbb{E}_{\mathcal{X}}[g(a)\hat{\tau}_0(a, x) + (1 - g(a))\hat{\tau}_1(a, x)]$
```

How to pick the right estimator and learner?

- Flexibility is good but which learner and estimator should we choose?
  - As always, the answer depends on the concrete application, effect size, and complexity of the hypothesis
  - S-learner is effective when the dataset is relatively small, or treatment effect is uniform across the observations
  - T-learner excels with larger datasets that can support separate models for treated and control groups
  - X-learner stands out in scenarios where treatment effects are expected to be highly heterogeneous or when there is an imbalance between treated and control group



Complexity	Base-learners	Meta-learner
Simple	OLS	S
		T
Complex	RF	X

Contribution

1. Game-level data diverging from traditional season-level data approach
   - There are numerous game-level confounders to consider including the back-to-back games, team you play for, the team you compete against, home court advantage, the teammates and opponents you encounter, the geographic location of the game, the season year, and various other relevant factors

2. Provide framework to estimate the Age-Conditioned Treatment Effect (ACTE) which
   - Enables the identification of causal effects under certain assumptions
   - Capture non-linear trends easier than the previous regression-based methods

3. Applied the methodology to study the effects of rest on multiple performance metrics across different ages
   - We find that the rest generally affects positively, but not constantly where the heterogeneity is driven by multiple factors

Application to NBA data to assess effect of load-management

- Success: Kawhi Leonard 2018-2019 season. Played only 60/82 regular season games, he was in top form for the playoffs, leading the Raptors to their championship

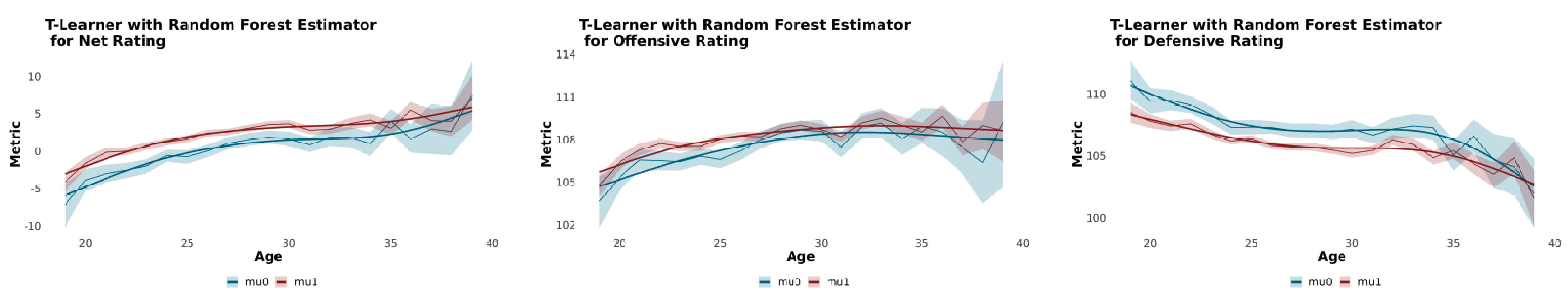
- Critiques: Impact on regular season's significance, diminishing fan engagement. Leading to Player Participation Policy for the 2023-2024 season, requiring star players to participate more frequently.

- Contribution: Quantitative analysis of load management's precise impact remains scarce, particularly in comparison to the effects of playing in back-to-back games. In this study, we apply the ACTE framework to measure the impact of rest on players, segmented by age.

- Note: While randomized controlled trial would be ideal it is infeasible for practical challenge in the NBA. Hence, we do observational.

Conditional Expectation Function (CEF) for T-learner with Random Forest

- Net, offensive, and defensive ratings
  - Not significant for young and old players due to lack of sample size
  - Non rested players tend to be more defensive liability compared to effect on offence



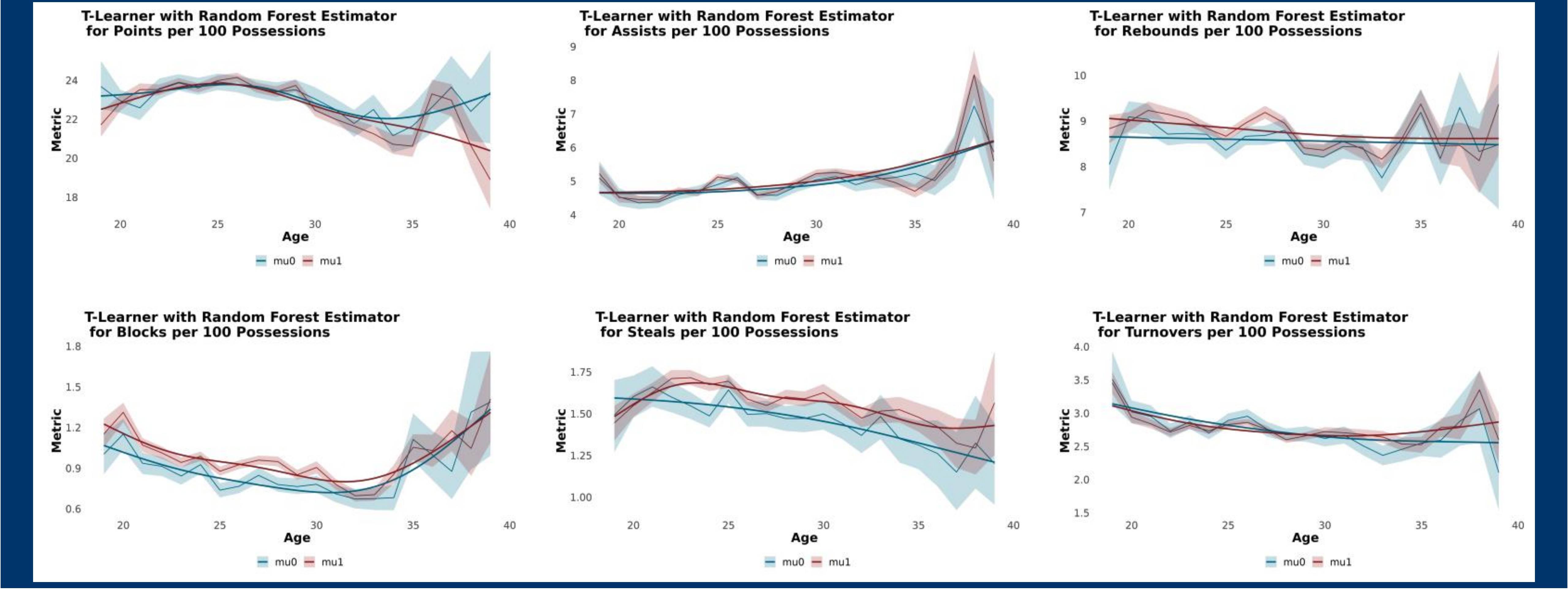
T-Learner with Random Forest Estimator for Net Rating

T-Learner with Random Forest Estimator for Offensive Rating

T-Learner with Random Forest Estimator for Defensive Rating

Conditional Expectation Function (CEF) for T-learner with Random Forest

- Normalized box score statistics



T-Learner with Random Forest Estimator for Points per 100 Possessions

T-Learner with Random Forest Estimator for Assists per 100 Possessions

T-Learner with Random Forest Estimator for Rebounds per 100 Possessions

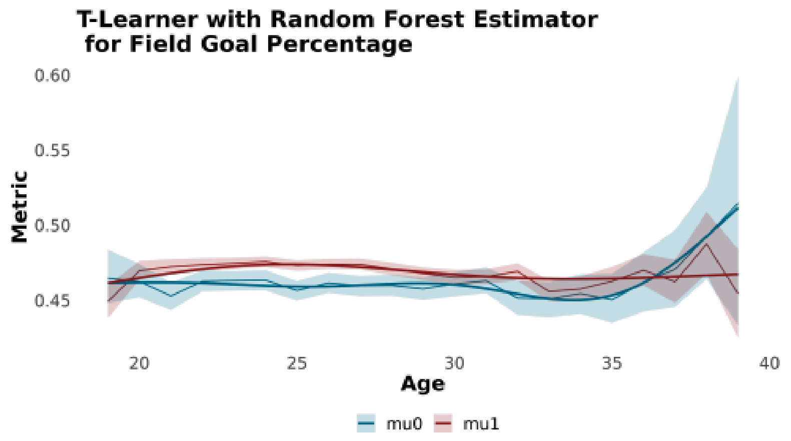
T-Learner with Random Forest Estimator for Blocks per 100 Possessions

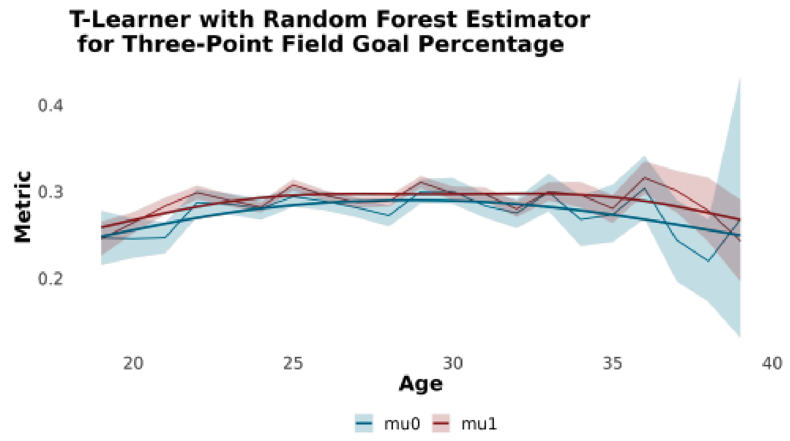
T-Learner with Random Forest Estimator for Steals per 100 Possessions

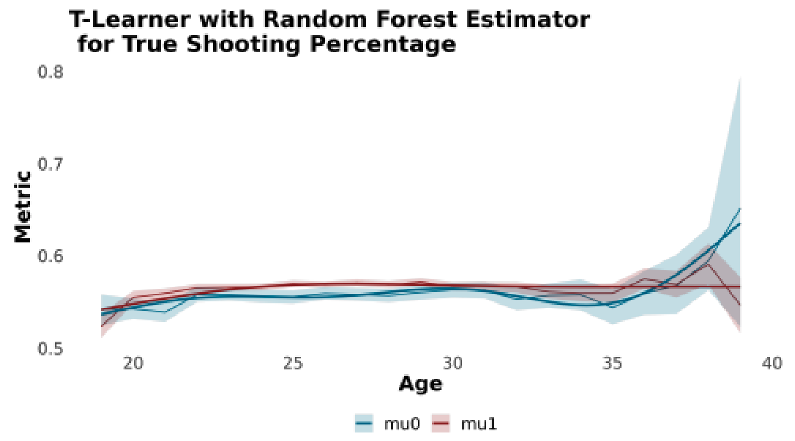
T-Learner with Random Forest Estimator for Turnovers per 100 Possessions

Conditional Expectation Function (CEF) for T-learner with Random Forest

- Shooting percentage
  - The effect is more pronounced for field goal percentage than for three-point field goal percentage. This is likely because the play becomes more physical as players get closer to the rim
  - Young and old player effects not significant due to sample size

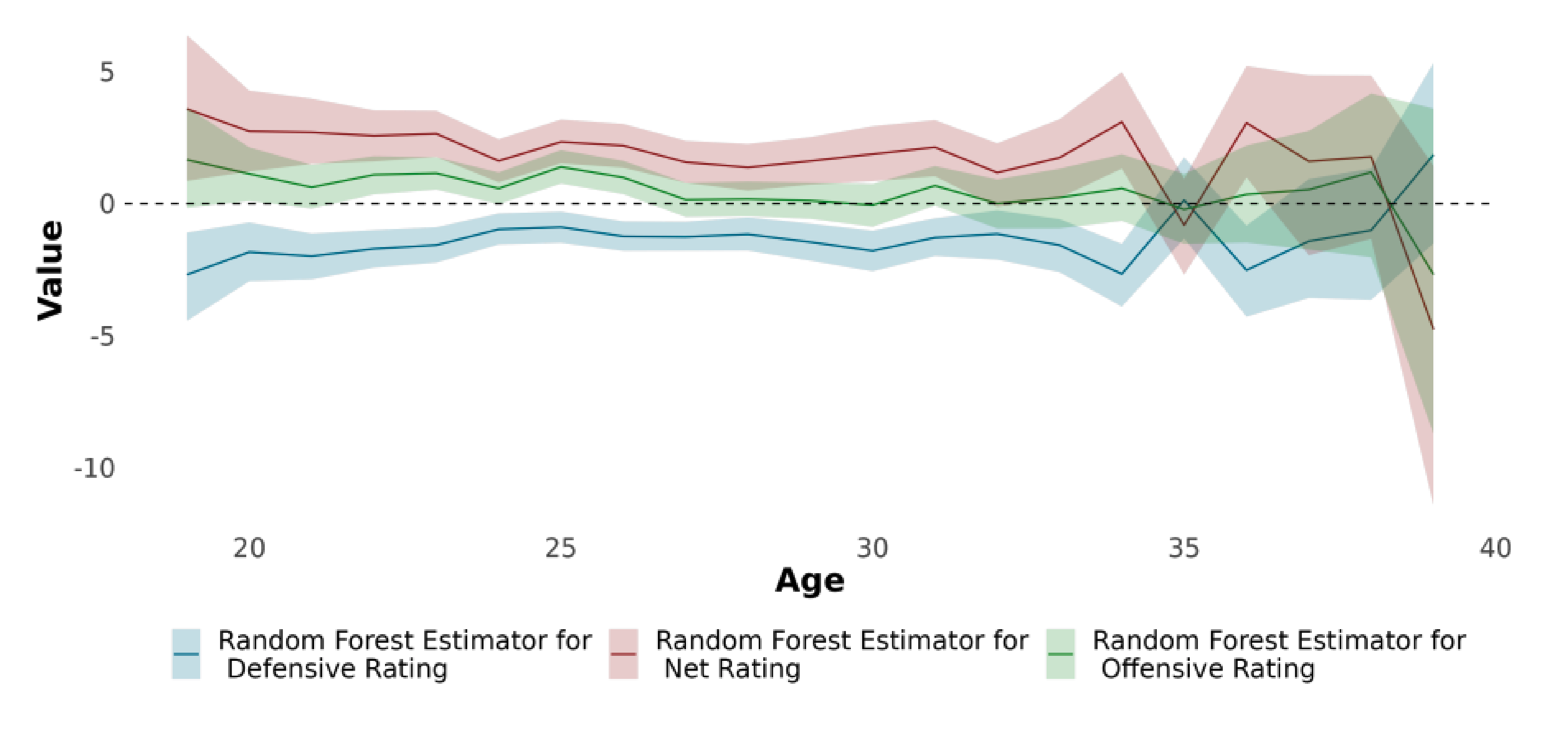

T-Learner with Random Forest Estimator for Field Goal Percentage


T-Learner with Random Forest Estimator for Three-Point Field Goal Percentage


T-Learner with Random Forest Estimator for True Shooting Percentage

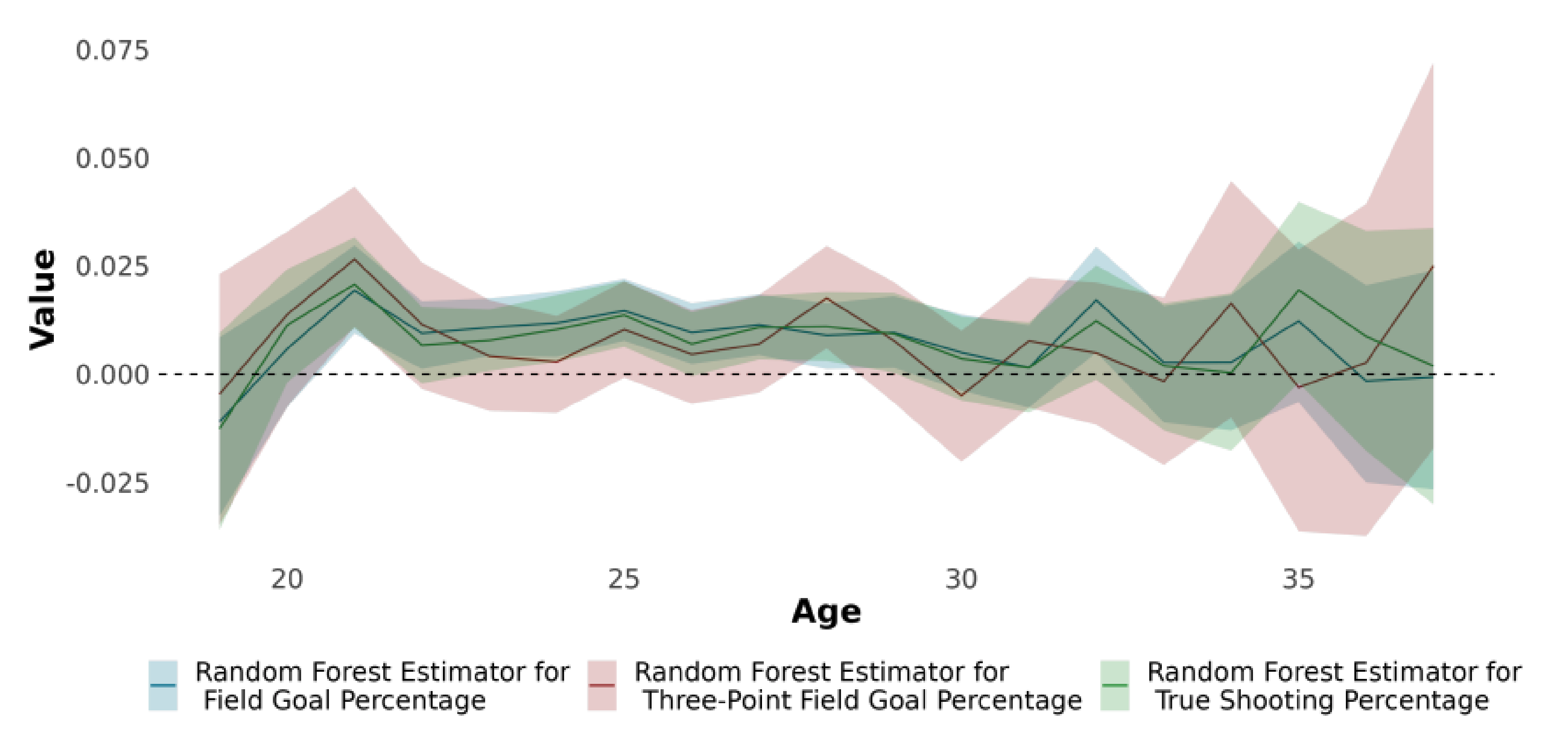
ACTE for X-learner with Random Forest

- Net, offensive, and defensive ratings



ACTE for X-learner with Random Forest

- Shooting percentage



Discussion and future work

- Rest matters: More on defensive end than in offensive end

- Advances the age-curve literature by
  - incorporating rich game-level data
  - Devising framework for causal effect estimation
  - Allowing to capture non-linear trends using flexible machine learning models

- Meta-learning framework for multiple treatment problem: Allows to differentiate between one day vs two days rest

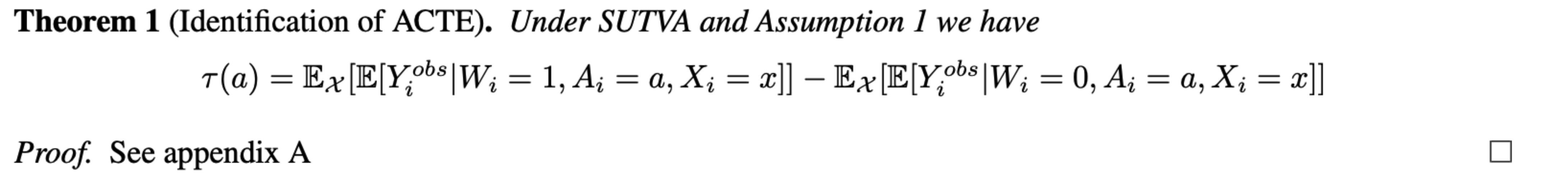
- Rest and fatigue: Potentially relate with the fatigue index

Framework

- Conditional Expectation Function (CEF)
  - Consider N units, age $A \in \mathbb{Z}$, treatment $W \in \{0,1\}$, $Y(w)$ potential outcome of unit i for treatment w Then the CEF is defined as:
    $\mu_0(a,x) \coloneqq \mathbb{E}[Y_i(0)|A_i = a, X_i = x]$ and $\mu_1(a,x) \coloneqq \mathbb{E}[Y_i(1)|A_i = a, X_i = x]$

  - Under this framework, our main causal estimand of interest is the ACTE, that is:
    $\tau(a) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0)|A = a] = \mathbb{E}_{\mathcal{X}}[\mu_1(a,x) - \mu_0(a,x)]$

- Identification of ACTE
  - Under some regularity conditions (SUTVA and unconfoundedness)



Theorem 1 (Identification of ACTE). *Under SUTVA and Assumption 1 we have*

$$\tau(a) = \mathbb{E}_{\mathcal{X}}[\mathbb{E}[Y_i^{obs}|W_i = 1, A_i = a, X_i = x]] - \mathbb{E}_{\mathcal{X}}[\mathbb{E}[Y_i^{obs}|W_i = 0, A_i = a, X_i = x]]$$

Proof. See appendix A □

Age-distribution

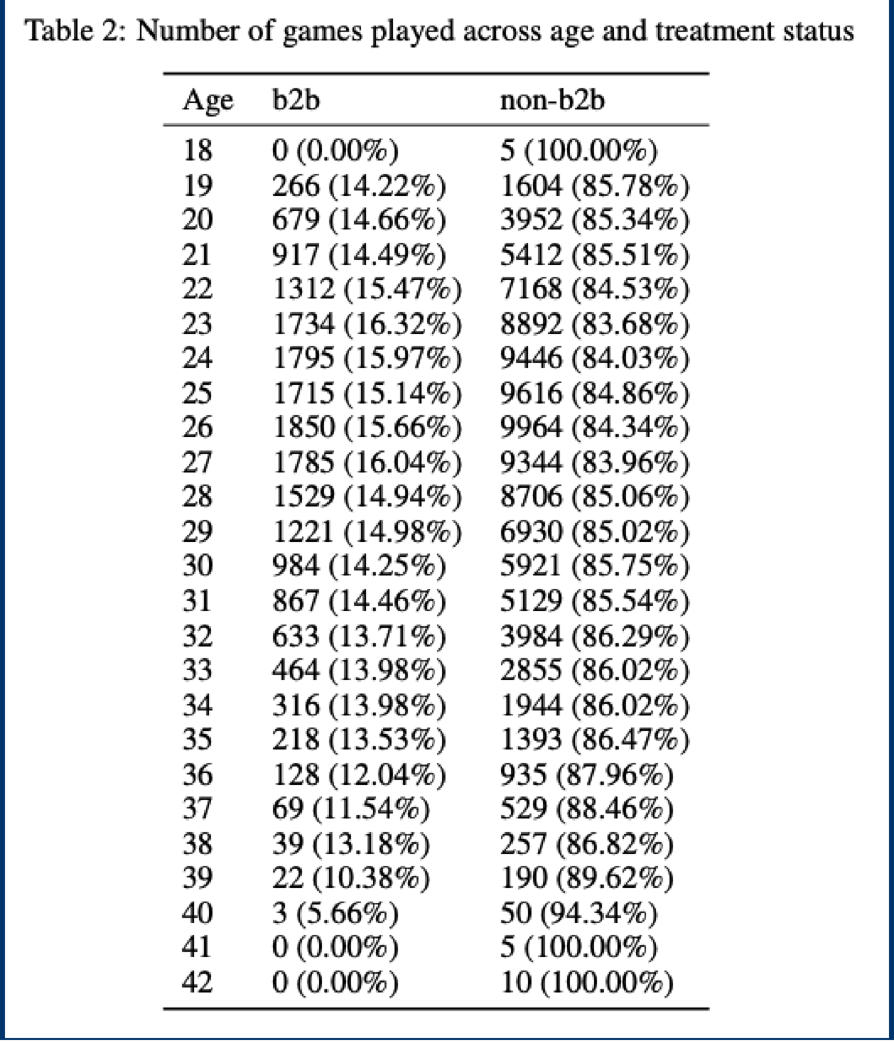


Table 2: Number of games played across age and treatment status

Age	b2b	non-b2b
18	0 (0.00%)	5 (100.00%)
19	266 (14.22%)	1604 (85.78%)
20	679 (14.66%)	3952 (85.34%)
21	917 (14.49%)	5412 (85.51%)
22	1312 (15.47%)	7168 (84.53%)
23	1734 (16.32%)	8892 (83.68%)
24	1795 (15.97%)	9446 (84.03%)
25	1715 (15.14%)	9616 (84.86%)
26	1850 (15.66%)	9964 (84.34%)
27	1785 (16.04%)	9344 (83.96%)
28	1529 (14.94%)	8706 (85.06%)
29	1221 (14.98%)	6930 (85.02%)
30	984 (14.25%)	5921 (85.75%)
31	867 (14.46%)	5129 (85.54%)
32	633 (13.71%)	3984 (86.29%)
33	464 (13.98%)	2855 (86.02%)
34	316 (13.98%)	1944 (86.02%)
35	218 (13.53%)	1393 (86.47%)
36	128 (12.04%)	935 (87.96%)
37	69 (11.54%)	529 (88.46%)
38	39 (13.18%)	257 (86.82%)
39	22 (10.38%)	190 (89.62%)
40	3 (5.66%)	50 (94.34%)
41	0 (0.00%)	5 (100.00%)
42	0 (0.00%)	10 (100.00%)