Estimating conditional average treatment effects for player performance over time:

assessing load-management in sports

Shinpei Nakamura-Sakai [1] Brian Macdonald [1]

[1]Yale University, Statistics and Data Science

Git

Shiny

Introduction

Athletes' performances improve, peak, and eventually decline. This curve is called the "age curve" and we expect this curve to have heterogeneity with respect to the characteristics of the players. In this work, we focus on estimating the effect of rest between games on performance for each age. This is helpful for making decisions about resting a player and so-called "load management". We make three main contributions: First, we construct a Conditional Expectation Function (CEF) to compare the age curve for different covariates and treatments. Second, using a causal inference approach, we propose a methodology to construct age conditioned treatment effect (ACTE) for a given treatment. The ACTE can test causal hypotheses for each age on the treatment and outcomes of interest. Third, we apply this method to assess the effect of days between games on multiple performance metrics conditional on age.

Notation, Framework, and Estimands

We assume a superpopulation or distribution $\mathcal{P}$ where we draw N independent tuples $(Y_i, A_i, W_i, X_i) \sim \mathcal{P}$ where Y is the observed outcome of interest, $A \in \mathbb{Z}$ is the age, $W \in \{0,1\}$ is the binary treatment of interest, and $X \in \mathcal{X}$ is the covariate matrix. Under Rubin's potential outcome framework [2] the Conditional Expectation Function (CFE) for a given age a is defined as

$$\mu_0(a,x) := \mathbb{E}[Y(0)|A=a, X=x] \text{ and } \mu_1(a,x) := \mathbb{E}[Y(1)|A=a, X=x]$$

In this work, our objective is not only the prediction but we have a strong belief that age is a driving factor to model $Y(w)$ so we decompose and assume that

$$Y_i(w) = g(a,w) + f_i(x,a,w) + \epsilon_i \tag{1}$$

following [3] where $g(t)$ is the average performance at age t for all players with treatment w, $f(x,a,w)$ represents a possible performance adjustment at age a for player with covariates x, and treatment w, and ϵ is the mean zero model error. Note that under this decomposition, g is the representation of all players and hence f represents the variation for all players at age a. Namely, $\mathbb{E}[f(x,a,w)|A=a] = 0$

Under this framework, note that

$$\begin{aligned}\tau(a) &:= \mathbb{E}[Y(1) - Y(0)|A=a] \\ &= \mathbb{E}[g(a,1) - g(a,0) + f(x,a,1) - f(x,a,0)|A=a] \\ &= \mathbb{E}[g(a,1) - g(a,0)|A=a]\end{aligned}$$

which indicates that:

$$\tau(a) := \mathbb{E}[g(a,1) - g(a,0)|A=a] \tag{2}$$

In order to properly identify this estimand we need the following assumptions:

- **Probabilistic Assignment**: The assignment probability is strictly between 0 and 1
$$0 < \mathbb{P}(W_i = 1|X_i, Y_i(0), Y_i(1)) < 1$$
- **Consistency**: There is only one version of the potential outcome for each w
$$Y_i^{obs} = Y_i(w_i)$$
- **No interference**: The treatment status of other players would not affect your outcome.
$$Y_i(\mathbf{W}) = Y_i(W_i)$$
- **Unconfoundedness**: There are no unmeasured factors affecting the probability of treatment allocation and the outcome.
$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i|X_i$$

Theorem 1(Identification of ACTE) Under above assumptions we have

$$\tau(a) = \mathbb{E}_{\mathcal{X}}[\mathbb{E}[Y|W=1] - \mathbb{E}[Y|W=0]|A=a, X=x]$$

S,T,X-learner with Splines

' **S-learner** is one of the causal ML meta-algorithms introduced by [1]. It is referred to as the S-learner because it uses a "single" estimator. In this learner, the treatment is included as one of the covariate and we learn the $\hat{g}$ using the next formulation

$$g(a,0) := \mathbb{E}[Y(0)|A=a, W=0, X=x] \text{ and } g(a,1) := \mathbb{E}[Y(1)|A=a, W=1, X=x]$$

then

$$\hat{\tau}(a) = \hat{g}(a,1) - \hat{g}(a,0)$$

T-learner uses "two" learners to estimate the τ. As first step, we learn $\hat{g}_0(a)$ and $\hat{g}_1(a)$ separately for treatment and control using the next set up

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

Then, the ACTE for each age and covariates is estimated by

$$\hat{\tau}(a) = \hat{g}_1(a) - \hat{g}_0(a)$$

X-learner consists of two-steps. As first step, we estimate two learners for treatment and control as in T-learner to obtain $\hat{g}_1(a)$ and $\hat{g}_0(a)$. For the second step, we calculate the "bias" using the model we trained in step one and feed the data set that we did not use to train.

$$\tilde{D}^1 = Y^1 - \hat{g}_0(a) \text{ and } \tilde{D}^0 = \hat{g}_1(a) - Y^0$$

Then, we use age spline regressions to estimate $\tau^0(a) = \mathbb{E}[\tilde{D}^1|A=a, X=x]$ and $\tau^1(a,x) = \mathbb{E}[\tilde{D}^0|A=a, X=x]$ and call its estimates $\hat{\tau}^0(a)$ and $\hat{\tau}^1(a)$.

Finally, the ACTE for each age and covariates is estimated by

$$\hat{\tau}(a) = e(a)\hat{\tau}_0(a) + (1 - e(a))\hat{\tau}_1(a)$$

where $e \in [0,1]$ is a weight function. Typically, we use the propensity score.
The advantage of S-learner is that the curve constructed by this approach is smooth and easier to interpret than X-learner. However, X-learner can handle more complex ACTE functions and treatment imbalance.
To estimate μ, we could use any base learner on the entire dataset. In particular, we will use the age spline regression method as described in [3] as our base learner and we will compare the result with random forest.

Simulation study

We realized three different simulation studies to assess which method we should use for ACTE estimation. The first scenario is the simplest one given by

1. **Simulation1 (Simple, constant treatment)**:
$$\tau(a) = 2$$
$$f(x,a,w) = \gamma_i + b_i(a - a_{max})^2 \mathbb{1}(a > a_{max})$$
2. **Simulation2 (Simple, linear treatment)**:
$$\tau(a) = 0.1(a - a_{min})$$
$$f(x,a,w) = \gamma_i + b_i(a - a_{max})^2 \mathbb{1}(a > a_{max})$$
3. **Simulation3 (Complex, non-linear treatment)**:
$$\tau(a) = 2(a-16) + 0.0005 * \mathbb{1}(a > 20) * (a - a_{max})^3 - 0.0005 * \mathbb{1}(a > a_{max}) * (a - a_{max})^4$$
$$f(x,a,w) = \gamma_i + b_i(a - a_{max})^2 \mathbb{1}(a > a_{max}) + (2 + 3w)x_i$$

where $g(a,w) = \omega + \beta_1(a - a_{max})^2 + \beta_2(a - a_{max})^2 \mathbb{1}(a > a_{max}) + \beta_3(a - a_{max})^3 \mathbb{1}(y > t_{max}) + \tau(a)w$, γ_i and b_i are mean zero normal distributions and other constant is from [3]. Recall that $Y_i(w) = g(a,w) + f_i(x,a,w) + \epsilon_i$

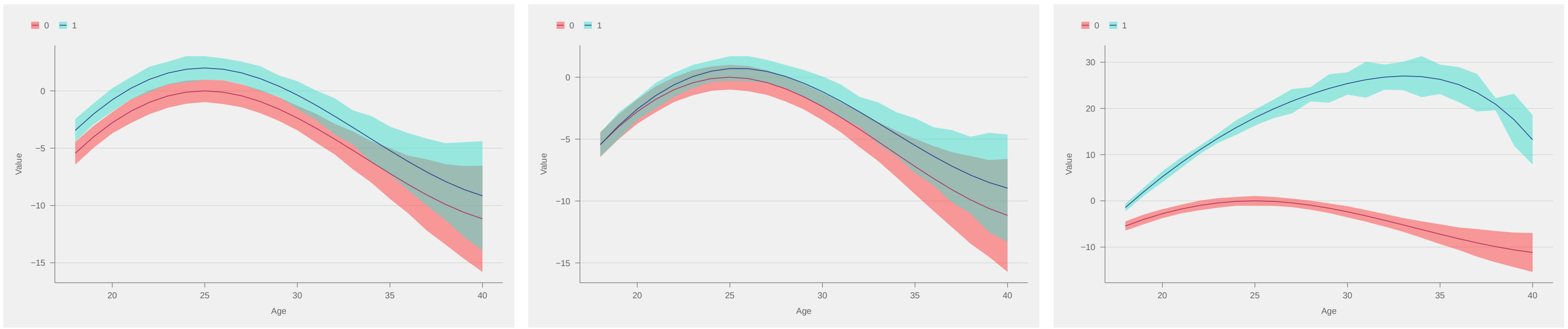

Figure 1. Simulation Scenarios

In the table 1 we can observe that the best method depends on each cases. If the treatment effect is complex then the X-learner with random forest would perform the best but if it is simple and treatment and control group have similar trend, S-learner with splines would be the best method. Researcher should make an appropriate assumptions before hand and decide carefully which model to use. In the real-world applications, by the fundamental problem of causal inference, we never can observe the counterfactual so we cannot calculate the MSE.

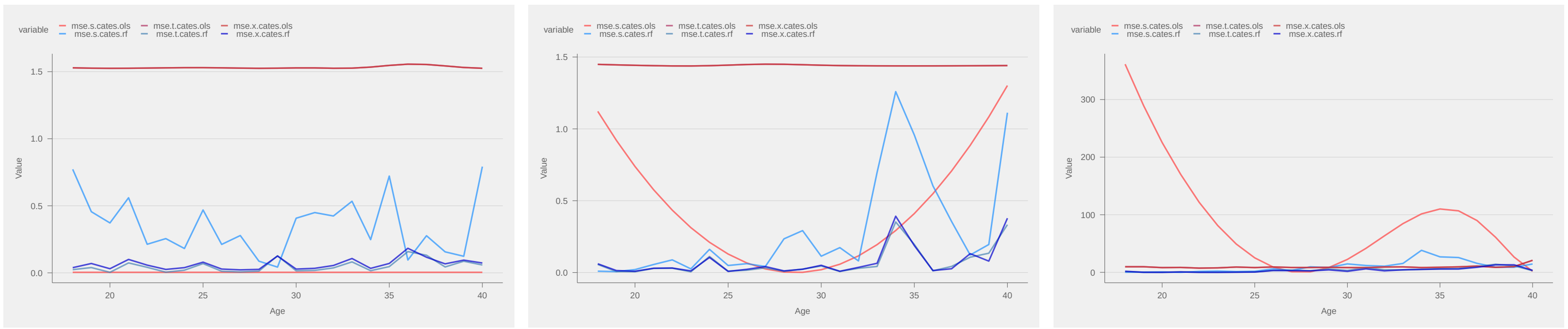

Figure 2. Simulation Results

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

Application to NBA data: Assessing load-management in sports

Using 10 seasons (2011-2022) of NBA data, we apply the above methods to estimate the effect of days of rest and age on points per 100 possessions played for NBA players. The treatment $w = 0$ indicates back-to-back (b2b) games and $w = 1$ indicates 1+day(s) rest, $Y(w)$ is the potential outcome when treatment is w, and X would be the covariate matrix. Then $\mu_w(a,x)$ indicates the conditional expectation of the potential outcome $Y(w)$ given age a. This function allows us to construct the age curve. By using $\tau(a,x)$, we get the ACTE, which is the effect of resting over playing b2b games given age.

As this is an observational study, we might not be able to remove entirely the confounders but we will control the model using player, team, team against, home-away indicator, and season fixed effect.

Results: S-learner/Splines

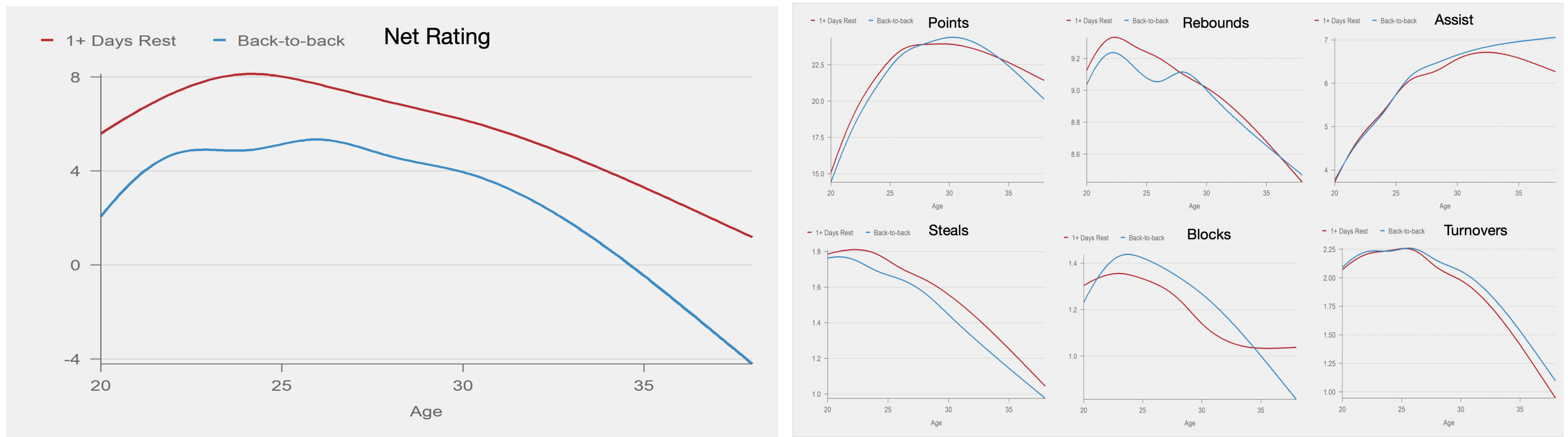

Figure 3. S-learner/Splines

The results of the S-learner can be observed in Figure 3. For net rating, we observe that b2b worsens the performance and moreover, the difference between the two curves increases as the player gets older.

We show the other metrics on the left. Points/100 poss. for ages considered as players' "peak" they successfully score without resting but for older ages, the scoring ability decreases faster. For rebounds, we observe that the scale is very small and we can conclude that there is a small difference between the two curves. Assists are interesting as players' do better for consecutive games. We believe that this happens as passing requires less physicality than scoring, an old player who played the game the day before might prefer passing rather than scoring due to fatigue. Resting is consistently better for steals and young players have higher steals per possession.

We believe the curve has a fitting problem for blocks as more than 60% of the players have 0 blocks. We could improve by fitting a Poisson regression instead of using the Gaussian link. Younger players make more turnovers and as they age, they tend to commit more turnovers in b2b games.

Results: X-learner/Splines

In Figure 4, we can observe the result of the X-learner using spline regressions.

On the left, we can observe the ACTE for each per 100 possessions variables. We can spot that points per 100 possessions are the most affected statistics using X-learner/Spline method. We are aware that the scaling for each of these variables might not be the same but for other outcomes except for points, the effect is marginal.

On the right, we observe that the net and offensive ratings have increasing trends. And the defensive rating is flat. We can observe that resting would affect offense more than defensive rating. Overall, the net rating improves and we notice that it is coming from the improvement in the offensive rating.

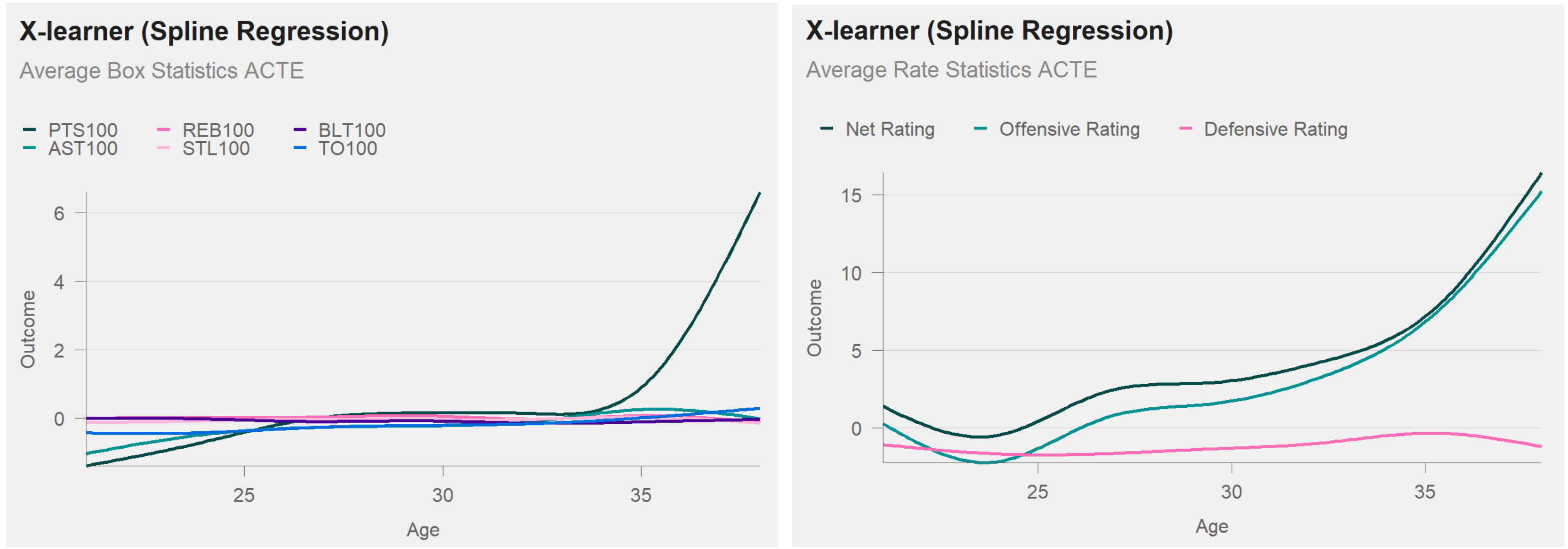

Figure 4. X-learners /Splines

Discussion

In this work, we connected the age curve with the causal ML literature in this work. We applied this to NBA data to assess players' performance changes in b2b games. This framework can be applied in numerous settings to determine the hidden causality in the data. Although we used the model that best performed on [3] where they explored different techniques to account for this bias, we would like to solve from an instrumental variable perspective by trying to estimate the Survival Average Causal effect (SACE) using principal stratification.

References

[1] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, March 2019. doi:10.1073/pnas.1804597116. URL https://www.pnas.org/doi/10.1073/pnas.1804597116. Publisher: Proceedings of the National Academy of Sciences.

[2] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974. ISSN 1939-2176. doi:10.1037/h0037350. Place: US Publisher: American Psychological Association.

[3] Michael Schuckers, Michael Lopez, and Brian Macdonald. What does not get observed can be used to make age curves stronger: estimating player age curves using regression and imputation, October 2021. URL http://arxiv.org/abs/2110.14017. arXiv:2110.14017 [stat].

NESSIS September 23rd, 2023,

s.nakamura.sakai@yale.edu