Machine Learning Theory

The R-Learner

David A. Hirshberg
February 6, 2025

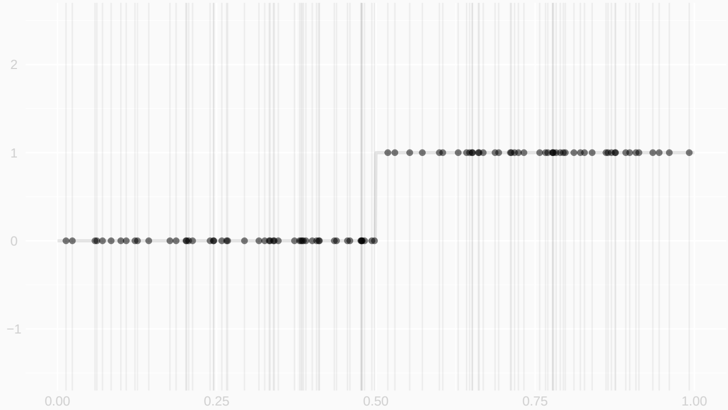
Emory University

Least Squares Review

What we've been doing

We started with a curve $\mu(x)$.

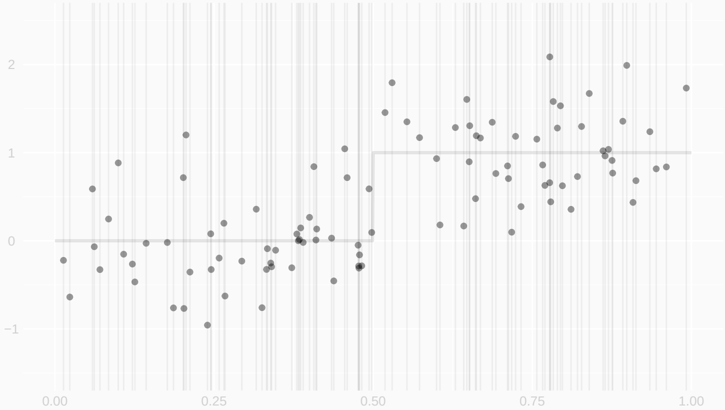
What we've been doing



We *sampled* it at some points X_i.

$$Y_i = \mu(X_i)$$

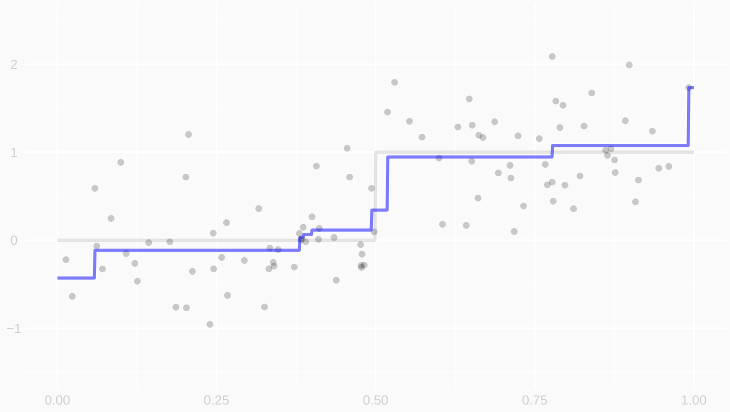
What we’ve been doing



We added *noise* to get our observations.

$$Y_i = \mu(X_i) + \varepsilon_i$$

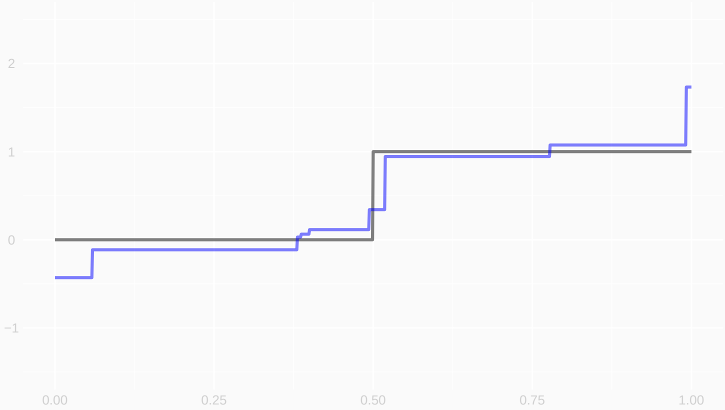
What we’ve been doing



We fit a curve, e.g. an increasing one, via least squares.

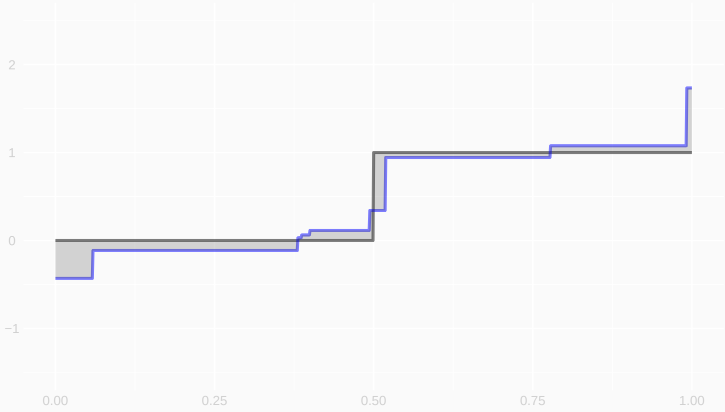
$$\hat{\mu} = \underset{\text{increasing } m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2 .$$

What we've been doing



We compared the curve we fit to the curve we started with.

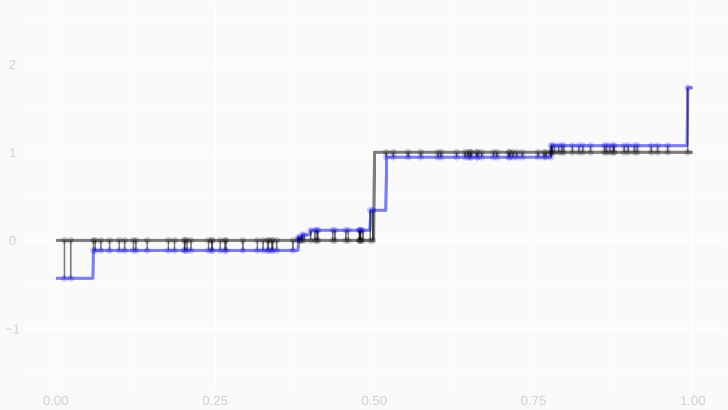
What we've been doing



We looked at mean squared distance over the whole interval.

$$\text{PMSE} = \int_0^1 \{\mu(x) - \hat{\mu}(x)\}^2 \, dx$$

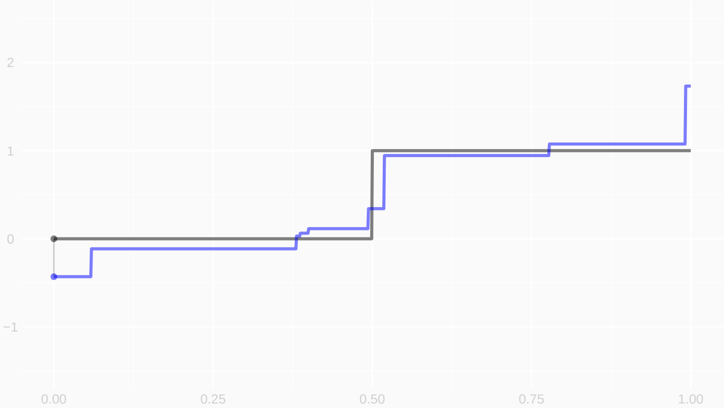
What we've been doing



And at mean squared distance over the sample.

$$\text{SMSE} = \frac{1}{n}\sum_{i=1}^{n}\{\mu(X_i) - \hat{\mu}(X_i)\}^2$$

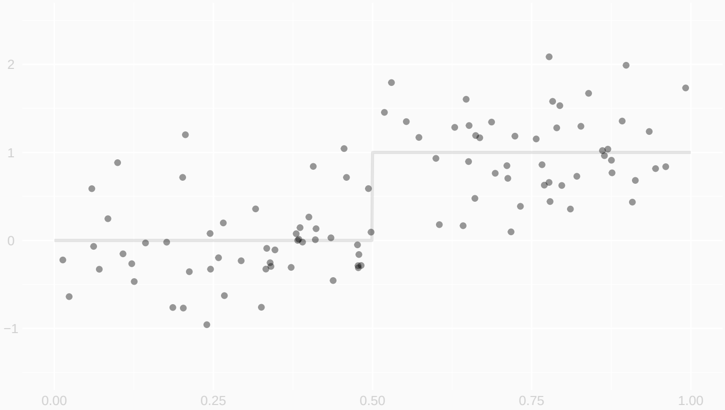
What we’ve been doing



And at squared distance at the left endpoint $x = 0$.

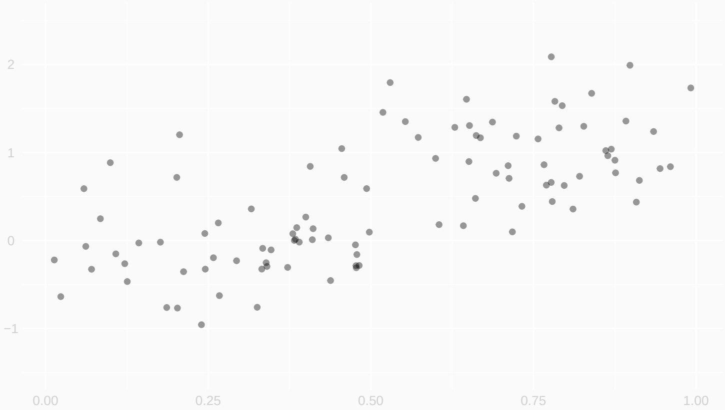
$$\text{MSE}_0 = \{\mu(0) - \hat{\mu}(0)\}^2$$

What we've been doing



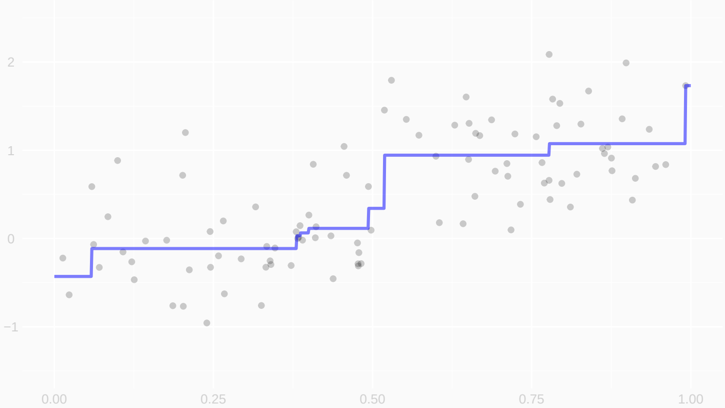
We could do all of this because we were using fake data.
We knew the curve μ that we'd sampled.

Working with real data is different



What we start with is the data.
We don't see any underlying curve μ.

Working with real data is different



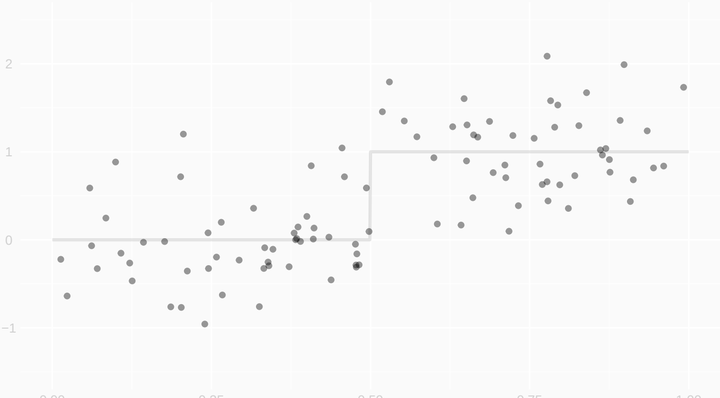
We can, of course, still fit a curve.
But what are we supposed to compare it to?
What curve are we trying to estimate?

What least squares estimates

The error we minimize, in large samples, approximates its expectation.

$$\frac{1}{n}\sum_{i=1}^{n}\{\,Y_i - m(X_i)\}^2 \quad \to \quad \mathrm{E}\{\,Y_i - m(X_i)\}^2$$

So what we might hope for is to estimate the curve μ minimizing that.
That's the *conditional mean* $\mu(x) = \mathrm{E}[\,Y_i \mid X_i = x]$.
It's the curve giving the mean value of Y_i at every value of X_i.



How do we know that?

Let's see what happens when we break Y_i into $\mu(X_i)$ and what's left over.
What's left over plays the role of our noise ε_i. What do we know about it?

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad ?$$

How do we know that?

Let's see what happens when we break Y_i into $\mu(X_i)$ and what's left over.
What's left over plays the role of our noise ε_i. What do we know about it?

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \mathrm{E}[\varepsilon_i \mid X_i] = 0.$$

Now let's use this to break down what we're minimizing.

How do we know that?

Let's see what happens when we break Y_i into $\mu(X_i)$ and what's left over. What's left over plays the role of our noise ε_i. What do we know about it?

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \mathrm{E}[\varepsilon_i \mid X_i] = 0.$$

Now let's use this to break down what we're minimizing.

$$\mathrm{E}\{Y_i - m(X_i)\}^2 = \mathrm{E}\{\varepsilon_i + \mu(X_i) - m(X_i)\}^2$$

How do we know that?

Let's see what happens when we break Y_i into $\mu(X_i)$ and what's left over. What's left over plays the role of our noise ε_i. What do we know about it?

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \mathrm{E}[\varepsilon_i \mid X_i] = 0.$$

Now let's use this to break down what we're minimizing.

$$\begin{aligned}
\mathrm{E}\{Y_i - m(X_i)\}^2 &= \mathrm{E}\{\varepsilon_i + \mu(X_i) - m(X_i)\}^2 \\
&= \mathrm{E}\,\varepsilon_i^2 + 2\,\mathrm{E}\,\varepsilon_i\{\mu(X_i) - m(X_i)\} + \mathrm{E}\{\mu(X_i) - m(X_i)\}^2
\end{aligned}$$

Why does this tell us the minimizer is μ?

How do we know that?

Let's see what happens when we break Y_i into $\mu(X_i)$ and what's left over.
What's left over plays the role of our noise ε_i. What do we know about it?

$$Y_i = \mu(X_i) + \varepsilon_i \quad \text{where} \quad \mathrm{E}[\varepsilon_i \mid X_i] = 0.$$

Now let's use this to break down what we're minimizing.

$$\begin{aligned}
\mathrm{E}\{Y_i - m(X_i)\}^2 &= \mathrm{E}\{\varepsilon_i + \mu(X_i) - m(X_i)\}^2 \\
&= \mathrm{E}\,\varepsilon_i^2 + 2\,\mathrm{E}\,\varepsilon_i\{\mu(X_i) - m(X_i)\} + \mathrm{E}\{\mu(X_i) - m(X_i)\}^2
\end{aligned}$$

Why does this tell us the minimizer is μ?

As we vary m, it's ...

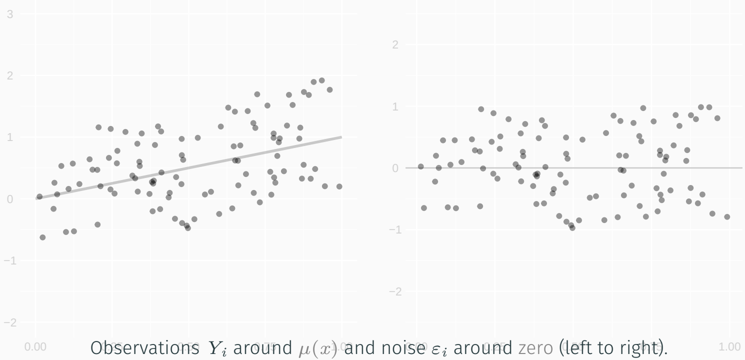
- a constant
- plus zero
- plus a positive term that's zero only if $m = \mu$

Signal and Noise in Least Squares Regression

If everything goes right, we'll approximate the conditional mean.

$$\mu(x) = \mathrm{E}[\,Y_i \mid X_i = x]$$

That's the *signal* we're trying to recover.
The *noise* it hides in has mean zero at each X_i.
This noise can be symmetric.



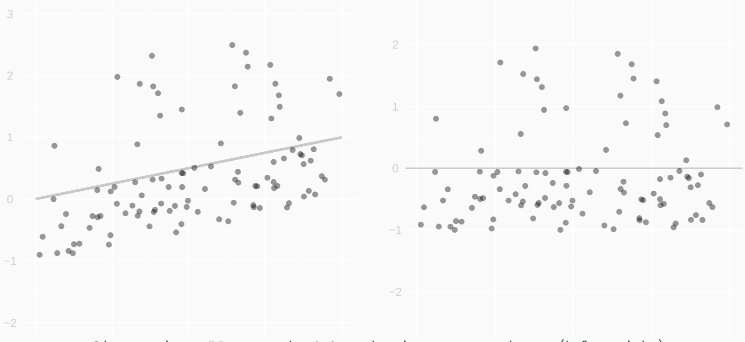
Observations Y_i around $\mu(x)$ and noise ε_i around zero (left to right).

$$Y_i = X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \text{Uniform}(-1, 1)$$

If everything goes right, we'll approximate the conditional mean.

$$\mu(x) = \mathrm{E}[\, Y_i \mid X_i = x]$$

That's the *signal* we're trying to recover.
The *noise* it hides in has mean zero at each X_i.
It can be asymmetric.



Observations Y_i around $\mu(x)$ and noise ε_i around zero (left to right).

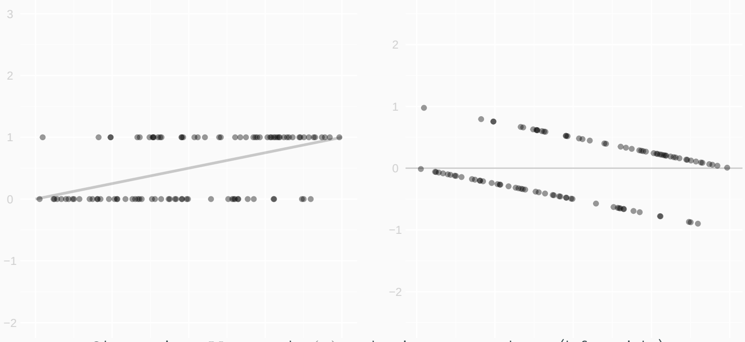
$$Y_i = X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \begin{cases} \text{Uniform}(0,2) & \text{with probability } 1/3 \\ \text{Uniform}(-1,0) & \text{with probability } 2/3 \end{cases}$$

Signal and Noise in Least Squares Regression

If everything goes right, we'll approximate the conditional mean.

$$\mu(x) = \mathrm{E}[\,Y_i \mid X_i = x]$$

That's the *signal* we're trying to recover.
The *noise* it hides in has mean zero at each X_i.
It can be very asymmetric.



Observations Y_i around $\mu(x)$ and noise ε_i around zero (left to right).

$$Y_i = X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \begin{cases} 1 - \mu(X_i) & \text{with probability } \mu(X_i) \\ -\mu(X_i) & \text{with probability } 1 - \mu(X_i) \end{cases}$$

What if we don't want to estimate the conditional mean?

We'll have to do something else.

There are many things to estimate and many ways to estimate them.
This week, we'll estimate *personalized treatment effects* using the *R-Learner*.

Personalized Treatment Effects

Context. Who did the NSW Job Training Program increase income for?

- The NSW program was implemented in the mid-1970s.
- It provided work experience and counseling for a period of 9-18 months.
- It enrolled people who tended to have difficulty with employment, e.g.,
  - People who'd been convicted of crimes
  - People who'd been addicted to drugs
  - People who'd not completed high school
- These participants were randomly assigned to the control or treatment groups.
- Both groups were interviewed, only the treated were given these short-term jobs.
- We want know who the treatment helps.

Specifics

- We're looking at income in 1978, after the program ended.
- We're interested in the impact of treatment on this.
- And we want to estimate the average effect of this treatment *among participants with a given 1974 income*.

Identification

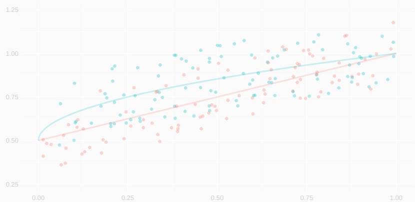
Due to randomization, this is conceptually simple.

- We want to compare each participant to an imaginary version of themself—one that got a different treatment—then average over folks with the same 1974 income.
- But given randomization, this is equivalent to a real comparison.
- If there's no difference, on average, between participants with identical 1974 incomes, we can swap in a real participant for our imaginary one.
- That's the case when all participants with the same '74 income receive treatment vs. control with the same probability.

What we want is to compare two conditional means.

$$\tau(x) = \mathrm{E}[\,Y_i(1) \mid X_i = x] - \mathrm{E}[\,Y_i(0) \mid X_i = x] \qquad \text{participant vs imaginary version of self}$$

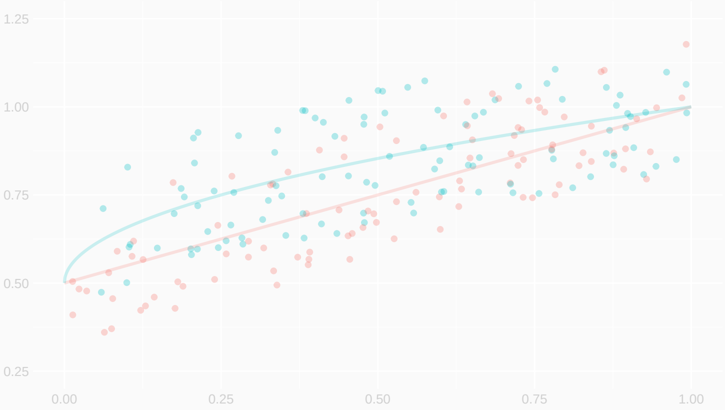
$$= \underset{\mu(0,x)}{\mathrm{E}[\,Y_i \mid W_i = 1, X_i = x]} - \underset{\mu(1,x)}{\mathrm{E}[\,Y_i \mid W_i = 0, X_i = x]} \qquad \text{participant vs one w/ same '74 income}$$



- $\mu(1, x)$, the mean for treated participants with 1974 income x
- $\mu(0, x)$, the mean for untreated participants with 1974 income x

Our income-specific treatment effect is a difference of two conditional means.

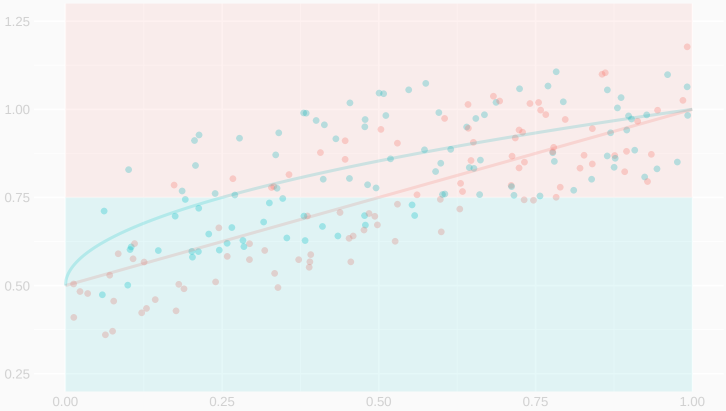
$$\tau(x) = \mu(1, x) - \mu(0, x)$$



In this fake data, all participants
receive treatment with probability $1/2$.

Our income-specific treatment effect is a difference of two conditional means.

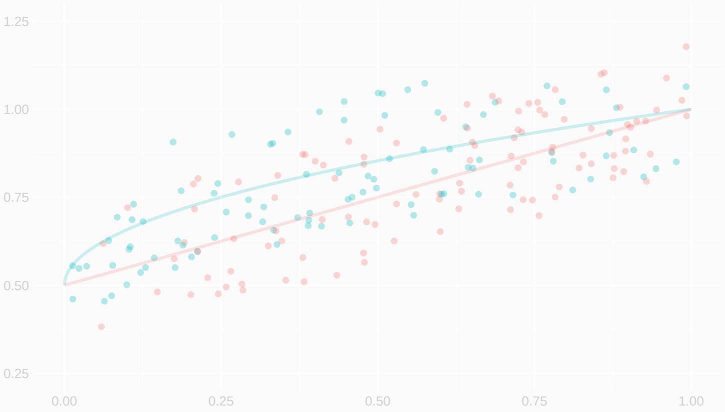
$$\tau(x) = \mu(1, x) - \mu(0, x)$$



In this fake data, all participants
receive treatment with probability $1/2$.

Our income-specific treatment effect is a difference of two conditional means.

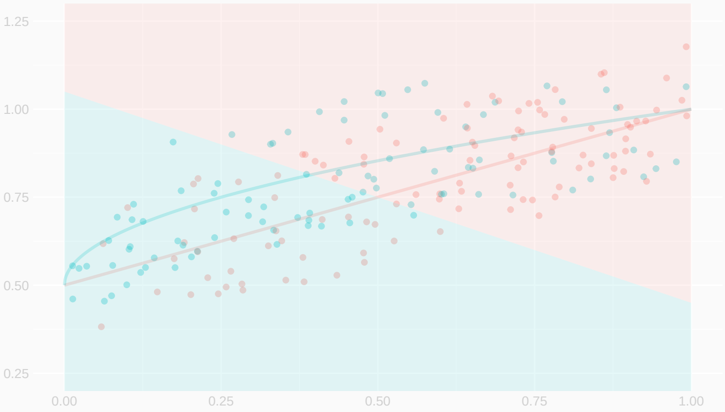
$$\tau(x) = \mu(1, x) - \mu(0, x)$$



In this fake data, participants with lower ’74 incomes
receive treatment more often than those with higher ’74 incomes.

Our income-specific treatment effect is a difference of two conditional means.

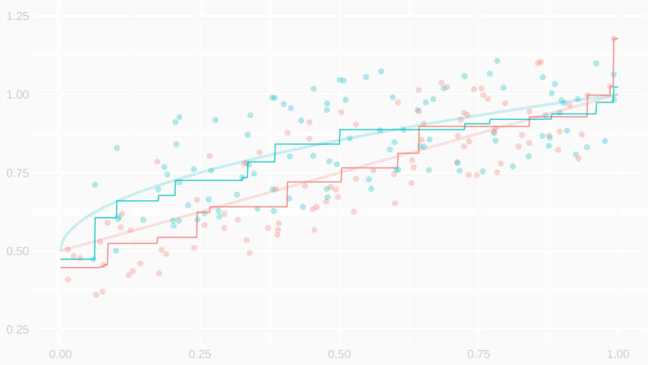
$$\tau(x) = \mu(1, x) - \mu(0, x)$$



In this fake data, participants with lower ’74 incomes
receive treatment more often than those with higher ’74 incomes.

We estimate the conditional mean for each treatment group, then subtract.

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



Here we've fit increasing curves to each group via least squares.

$$\hat{\mu}(w, \cdot) = \underset{\text{increasing } m}{\operatorname{argmin}} \sum_{i:W_i = w} \{Y_i - m(X_i)\}^2 \quad \text{for} \quad w \in \{0, 1\}.$$

What we don't like about it

- We have to estimate two treatment-specific conditional means.
- If we don't estimate one well, we tend to get a bad treatment effect estimate.
- It's hard to encode assumptions about the treatment effect itself in our model for these conditional means.
  - e.g. constancy, $\tau(x) = \tau$.
  - e.g. approximate constancy, $\rho_{TV}(\tau) \approx 0$.
  - e.g. decreasingness, $\tau'(x) \leq 0$.

Let's try to fix that.

Starting Point: Robinson's Decomposition

We express our treatment-specific conditional means in terms of a few other things.

$$\mu(W_i, X_i) = \beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i) \quad \text{where}$$

$$\beta(X_i) = \mathrm{E}[Y_i \mid X_i] \quad \text{is the (nonspecific) conditional mean,}$$

$$\pi(X_i) = P(W_i = 1 \mid X_i), \quad \text{is the conditional treatment probability,}$$

$$\tau(X_i) = \mathrm{E}[Y_i \mid W_i = 1, X_i] - \mathrm{E}[Y_i \mid W_i = 0, X_i] \quad \text{is the conditional treatment effect.}$$

Starting Point: Robinson's Decomposition

We express our treatment-specific conditional means in terms of a few other things.

$$\mu(W_i, X_i) = \beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i)$$ where

$$\beta(X_i) = \mathrm{E}[Y_i \mid X_i]$$ is the (nonspecific) conditional mean,

$$\pi(X_i) = P(W_i = 1 \mid X_i),$$ is the conditional treatment probability,

$$\tau(X_i) = \mathrm{E}[Y_i \mid W_i = 1, X_i] - \mathrm{E}[Y_i \mid W_i = 0, X_i]$$ is the conditional treatment effect.

Derivation. We start with a characterization of $\beta(X_i)$ as a marginal of $\mu(W_i, X_i)$.

Then we plug it in.

Starting Point: Robinson's Decomposition

We express our treatment-specific conditional means in terms of a few other things.

$$\mu(W_i, X_i) = \beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i) \quad \text{where}$$

$$\beta(X_i) = \mathrm{E}[Y_i \mid X_i] \quad \text{is the (nonspecific) conditional mean,}$$

$$\pi(X_i) = P(W_i = 1 \mid X_i), \quad \text{is the conditional treatment probability,}$$

$$\tau(X_i) = \mathrm{E}[Y_i \mid W_i = 1, X_i] - \mathrm{E}[Y_i \mid W_i = 0, X_i] \quad \text{is the conditional treatment effect.}$$

Derivation. We start with a characterization of $\beta(X_i)$ as a marginal of $\mu(W_i, X_i)$.

$$\begin{aligned}
\beta(X_i) &= \mathrm{E}[Y_i \mid X_i] \\
&= \mathrm{E}[Y_i \mid W_i = 1, X_i]P(W_i = 1 \mid X_i) + \mathrm{E}[Y_i \mid W_i = 0, X_i]P(W_i = 0 \mid X_i) \\
&= \mu(1, X_i)\pi(X_i) + \mu(0, X_i)\{1 - \pi(X_i)\}.
\end{aligned}$$

Then we plug it in.

Starting Point: Robinson's Decomposition

We express our treatment-specific conditional means in terms of a few other things.

$$
\begin{aligned}
\mu(W_i, X_i) &= \beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i) && \text{where} \\
\beta(X_i) &= \mathrm{E}[Y_i \mid X_i] && \text{is the (nonspecific) conditional mean,} \\
\pi(X_i) &= P(W_i = 1 \mid X_i), && \text{is the conditional treatment probability,} \\
\tau(X_i) &= \mathrm{E}[Y_i \mid W_i = 1, X_i] - \mathrm{E}[Y_i \mid W_i = 0, X_i] && \text{is the conditional treatment effect.}
\end{aligned}
$$

Derivation. We start with a characterization of $\beta(X_i)$ as a marginal of $\mu(W_i, X_i)$.

$$
\begin{aligned}
\beta(X_i) &= \mathrm{E}[Y_i \mid X_i] \\
&= \mathrm{E}[Y_i \mid W_i = 1, X_i] P(W_i = 1 \mid X_i) + \mathrm{E}[Y_i \mid W_i = 0, X_i] P(W_i = 0 \mid X_i) \\
&= \mu(1, X_i)\pi(X_i) + \mu(0, X_i)\{1 - \pi(X_i)\}.
\end{aligned}
$$

Then we plug it in.

$$
\begin{aligned}
&\beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i) \\
&= \mu(1, X_i)\pi(X_i) + \mu(0, X_i)\{1 - \pi(X_i)\} + \{W_i - \pi(X_i)\}\{\mu(1, X_i) - \mu(0, X_i)\} \\
&= \mu(1, X_i)\{\pi(X_i) + W_i - \pi(X_i)\} + \mu(X_i, 0)[\{1 - \pi(X_i)\} - \{W_i - \pi(X_i)\}] \\
&= \mu(1, X_i) W_i + \mu(0, X_i)(1 - W_i) = \mu(X_i, W_i).
\end{aligned}
$$

The R-Learner idea

$$\mu(W_i, X_i) = \beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i)$$ where

$$\beta(X_i) = \mathrm{E}[Y_i \mid X_i]$$ is the (nonspecific) conditional mean,

$$\pi(X_i) = P(W_i = 1 \mid X_i),$$ is the conditional treatment probability,

$$\tau(X_i) = \mathrm{E}[Y_i \mid W_i = 1, X_i] - \mathrm{E}[Y_i \mid W_i = 0, X_i]$$ is the conditional treatment effect.

If we knew the functions β and π, we could estimate τ using a special model.

$$\hat{\tau} = \underset{t \in \mathcal{M}_\tau}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m_t(W_i, X_i)\}^2 \text{ where } m_t(w, x) = \beta(x) + [w - \pi(x)]t(x).$$

This is a *weighted least squares* estimate of τ based on a *pseudo-outcome* Y_i^τ.

To show this, let's decompose the error $Y_i - m_t(W_i, X_i)$ in terms of the functions above, then plug the result into our least squares loss.

The R-Learner idea

$$\mu(W_i, X_i) = \beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i)$$ where

$$\beta(X_i) = \mathrm{E}[Y_i \mid X_i]$$ is the (nonspecific) conditional mean,

$$\pi(X_i) = P(W_i = 1 \mid X_i),$$ is the conditional treatment probability,

$$\tau(X_i) = \mathrm{E}[Y_i \mid W_i = 1, X_i] - \mathrm{E}[Y_i \mid W_i = 0, X_i]$$ is the conditional treatment effect.

If we knew the functions β and π, we could estimate τ using a special model.

$$\hat{\tau} = \operatorname*{argmin}_{t \in \mathcal{M}_\tau} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m_t(W_i, X_i)\}^2 \quad \text{where} \quad m_t(w, x) = \beta(x) + [w - \pi(x)]t(x).$$

This is a *weighted least squares* estimate of τ based on a *pseudo-outcome* Y_i^τ.

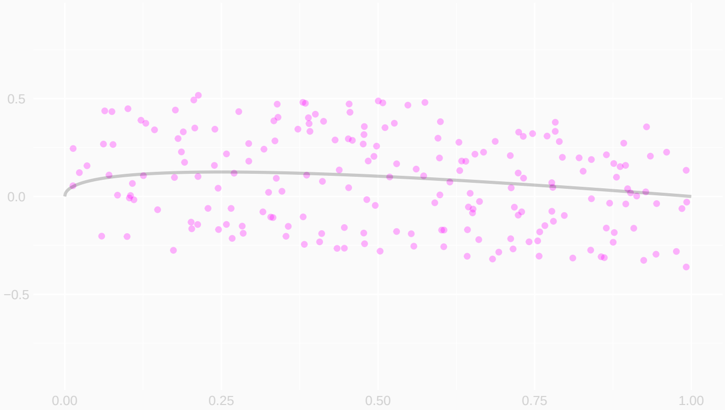
$$\begin{aligned} Y_i - m_t(W_i, X_i) &= [\beta(X_i) + \{W_i - \pi(X_i)\}\tau(X_i) + \varepsilon_i] - [\beta(X_i) + \{W_i - \pi(X_i)\}t(X_i)] \\ &= \{W_i - \pi(X_i)\}\{\tau(X_i) - t(X_i)\} + \varepsilon_i \\ &= \{W_i - \pi(X_i)\}\{\tau(X_i) + \varepsilon_i^\tau - t(X_i)\} \quad \text{where} \quad \varepsilon_i^\tau = \frac{\varepsilon_i}{W_i - \pi(X_i)}. \end{aligned}$$

so what we'd minimize is weighted squared error for predicting Y_i^τ.

$$\hat{\tau} = \operatorname*{argmin}_{t \in \mathcal{M}_\tau} \frac{1}{n} \sum_{i=1}^{n} \{W_i - \pi(X_i)\}^2 \{Y_i^\tau - t(X_i)\}^2 \quad \text{where} \quad Y_i^\tau = \tau(X_i) + \varepsilon_i^\tau.$$

Our Pseudo-Outcomes

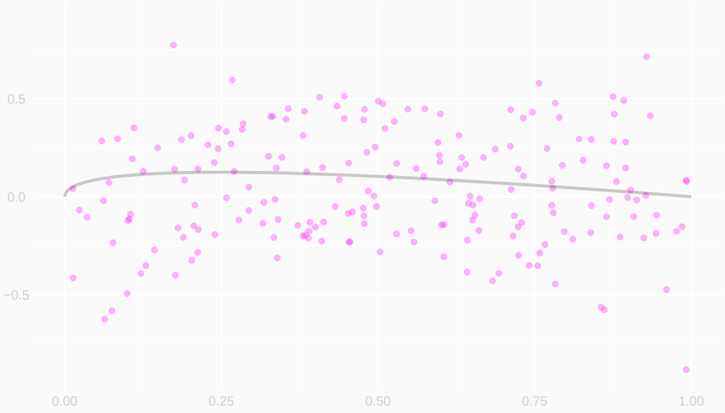
$$\hat{\tau} = \underset{t \in \mathcal{M}_\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{W_i - \pi(X_i)\}^2 \{Y_i^\tau - t(X_i)\}^2 \quad \text{where} \quad Y_i^\tau = \tau(X_i) + \varepsilon_i^\tau.$$



Pseudo-outcomes Y_i^τ when all participants receive treatment with probability $1/2$.

Our Pseudo-Outcomes

$$\hat{\tau} = \underset{t \in \mathcal{M}_\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{W_i - \pi(X_i)\}^2 \{Y_i^\tau - t(X_i)\}^2 \quad \text{where} \quad Y_i^\tau = \tau(X_i) + \varepsilon_i^\tau.$$



Pseudo-outcomes Y_i^τ when participants with lower '74 incomes receive treatment more often than those with higher '74 incomes.

The real R-Learner

$$\hat{\tau}_{\star} = \operatorname*{argmin}_{t \in \mathcal{M}_\tau} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m_t(W_i, X_i)\}^2 \quad \text{where} \quad m_t(w,x) = \beta(x) + [w - \pi(x)]t(x)$$

- This is what we've been talking about doing.
- But we can't really do it because we don't know the nuisance function β.
- That's why it's a nuisance. We need to know it, even if we're not interested in it.
- To actually use the R-Learner, we'll have to substitute an estimate.

$$\hat{\tau} = \operatorname*{argmin}_{t \in \mathcal{M}_\tau} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m_t(W_i, X_i)\}^2 \quad \text{where} \quad m_t(w,x) = \hat{\beta}(x) + [w - \pi(x)]t(x)$$

$$\text{and} \quad \hat{\beta} = \operatorname*{argmin}_{b \in \mathcal{M}_\beta} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - b(X_i)\}^2$$

The real R-Learner as least squares

$$\hat{\tau} = \operatorname*{argmin}_{t \in \mathcal{M}_\tau} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m_t(W_i, X_i) \}^2 \quad \text{where} \quad m_t(w, x) = \hat{\beta}(x) + [w - \pi(x)] t(x)$$

This is another *weighted least squares* estimate of τ.

The real R-Learner as least squares

$$\hat{\tau} = \underset{t \in \mathcal{M}_\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m_t(W_i, X_i) \}^2 \quad \text{where} \quad m_t(w, x) = \hat{\beta}(x) + [w - \pi(x)] t(x)$$

This is another *weighted least squares* estimate of τ.

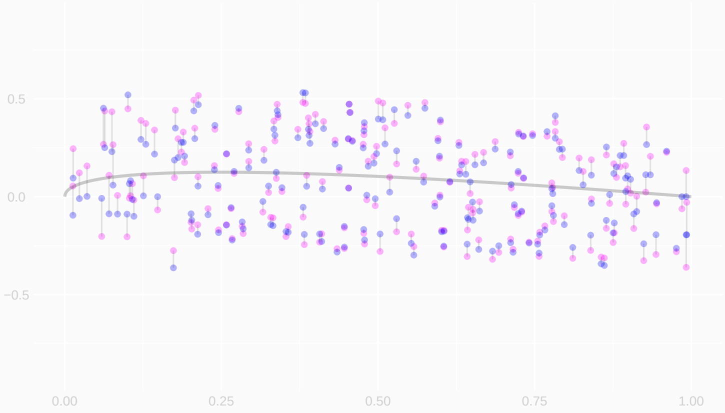
$$\begin{aligned} Y_i - m_t(W_i, X_i) &= [\beta(X_i) + \{ W_i - \pi(X_i) \} \tau(X_i) + \varepsilon_i] - \left[\hat{\beta}(X_i) + \{ W_i - \pi(X_i) \} t(X_i) \right] \\ &= \{ W_i - \pi(X_i) \} \{ \tau(X_i) - t(X_i) \} + \{ \beta(X_i) - \hat{\beta}(X_i) \} + \varepsilon_i \\ &= \{ W_i - \pi(X_i) \} \{ \tau(X_i) + \varepsilon_i^\tau + \delta_i - t(X_i) \} \quad \text{where} \quad \delta_i = \frac{\beta(X_i) - \hat{\beta}(X_i)}{W_i - \pi(X_i)} \end{aligned}$$

we're minimizing weighted squared error for predicting a corrupted pseudo-outcome.

$$\hat{\tau} = \underset{t \in \mathcal{M}_\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ W_i - \pi(X_i) \}^2 \{ Y_i^\tau + \delta_i - t(X_i) \}^2 \quad \text{where} \quad Y_i^\tau = \tau(X_i) + \varepsilon_i^\tau .$$

The corrupted pseudo-outcome

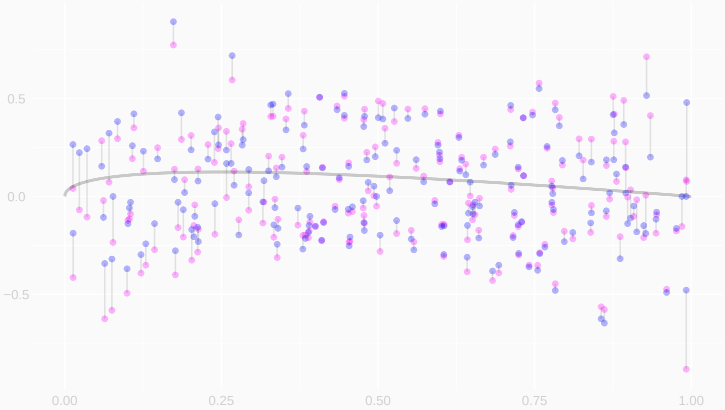
$$\hat{\tau} = \underset{t \in \mathcal{M}_\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{W_i - \pi(X_i)\}^2 \{Y_i^\tau + \delta_i - t(X_i)\}^2 \quad \text{where} \quad Y_i^\tau = \tau(X_i) + \varepsilon_i^\tau.$$



When all participants receive treatment with probability $1/2$.

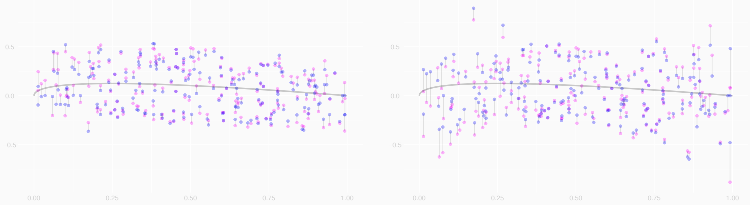
The corrupted pseudo-outcome

$$\hat{\tau} = \underset{t \in \mathcal{M}_\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{W_i - \pi(X_i)\}^2 \{Y_i^\tau + \delta_i - t(X_i)\}^2 \quad \text{where} \quad Y_i^\tau = \tau(X_i) + \varepsilon_i^\tau.$$



When participants with lower '74 incomes receive treatment more often than those with higher '74 incomes.

Robustness



- One very interesting property of the R-Learner is that it's insensitive to $\hat{\beta}$.
  - That is, it works well even if $\hat{\beta}$ is a pretty bad estimate.
  - Or, at least, it works almost as well as a version using β itself.
- That is, we estimate τ essentially as if we were doing weighted least squares prediction of the pseudo-outcomes.
  - The 'corruption' of the pseudo-outcomes we really learn to predict isn't a big deal.
  - We're using our knowledge about the treatment probability $\pi(x)$ to help us.
- Let's look at how this works for a very simple treatment effect model $\mathcal{M}_\tau$.

An Exercise

Show that, in the case that we use the constant treatment effect model $\mathcal{M}_\tau = \{t(x) = c : c \in \mathbb{R}\}$, these two versions of the R-learner differ by a term that's small relative to $1/\sqrt{n}$ as long as $\hat{\beta} \to \beta$. That is, show that

$$\sqrt{n}(\hat{\tau}_{\hat{\beta}} - \hat{\tau}_{\beta}) \to 0 \quad \text{if} \quad \hat{\beta} \to \beta$$

for

$$\hat{\tau}_{\hat{\beta}} = \operatorname*{argmin}_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m_t(W_i, X_i)\}^2 \ \text{ where } \ m_t(w, x) = \hat{\beta}(x) + [w - \pi(x)]t$$
$$\hat{\tau}_{\beta} = \operatorname*{argmin}_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m_t(W_i, X_i)\}^2 \ \text{ where } \ m_t(w, x) = \beta(x) + [w - \pi(x)]t$$

You may treat $\hat{\beta}$ as a non-random function. In practice, we'll split our sample in two and estimate $\hat{\beta}$ and $\hat{\tau}$ on different halves, which allows us to justify this rigorously.

Hint. Solve for $\hat{\tau}_{\hat{\beta}}$ and $\hat{\tau}_{\beta}$ explicitly by setting derivatives to zero, then compare the results. When you do, pay attention to the mean and *standard deviation* of your terms.

Step 1. Solving for $\hat{\tau}$ as a function of β

$$\hat{\tau}_b = \underset{t \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m_t(W_i, X_i)\}^2 \ \text{ where } \ m_t(w,x) = b(x) + [w - \pi(x)]t$$

solves

$$\begin{aligned} 0 &= \frac{d}{dt}\bigg|_{t=\hat{\tau}_b} \ \frac{1}{n}\sum_{i=1}^{n}\{Y_i - b(X_i) - [W_i - \pi(X_i)]t\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} 2\{Y_i - b(X_i) - [W_i - \pi(X_i)]\hat{\tau}_b\} \times -[W_i - \pi(X_i)]. \end{aligned}$$

Rearranging,

$$\frac{2}{n}\sum_{i=1}^{n}[W_i - \pi(X_i)]\{Y_i - b(X_i)\} = \frac{2}{n}\sum_{i=1}^{n}[W_i - \pi(X_i)]^2\hat{\tau}_b$$

and therefore

$$\hat{\tau}_b = \frac{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}\{Y_i - b(X_i)\}}{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}^2}$$

Comparison

$$\hat{\tau}_b = \frac{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}\{Y_i - b(X_i)\}}{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}^2}$$

for $b = \hat{\beta}$ and $b = \beta$. Comparing,

$$\begin{aligned}\hat{\tau}_{\hat{\beta}} - \hat{\tau}_{\beta} &= \frac{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}[\{Y_i - \hat{\beta}(X_i)\} - \{Y_i - \beta(X_i)\}]}{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}\{\beta(X_i) - \hat{\beta}(X_i)\}}{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}^2}\end{aligned}$$

What this tells us about the difference $\hat{\tau}_{\hat{\beta}} - \hat{\tau}_{\beta}$.

1. It's *almost* an average of independent random variables with mean zero, as $\mathrm{E}\{W_i - \pi(X_i) | X_i\} = \pi(X_i) - \pi(X_i) = 0$.
2. It would be if we replaced the denominator with its expectation, which the law of large numbers more or less justifies.

$$\hat{\tau}_{\hat{\beta}} - \hat{\tau}_{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}\{\beta(X_i) - \hat{\beta}(X_i)\}}{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{W_i - \pi(X_i)\}^2} \times \frac{1}{Q} \quad \text{for} \quad Q = \underbrace{\frac{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{W_i - \pi(X_i)\}^2}}_{\to 1}$$

$$\hat{\tau}_{\hat{\beta}} - \hat{\tau}_{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}\{\beta(X_i) - \hat{\beta}(X_i)\}}{\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{W_i - \pi(X_i)\}^2} \times \frac{1}{Q} \quad \text{for} \quad Q \to 1$$

If we ignore the largely irrelevant factor $1/Q$ (see Slutsky's Theorem), then ...

1. This difference is an average of independent mean-zero random variables.
2. So it's approximately normal with variance $\frac{1}{n}\times$ the average of their variances.

What is this variance?

$$\begin{aligned} n \times V &= \frac{\mathrm{E}\ \frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}^2\{\beta(X_i) - \hat{\beta}(X_i)\}^2}{\left[\ \mathrm{E}\ \frac{1}{n}\sum_{i=1}^{n}\{W_i - \pi(X_i)\}^2\ \right]^2} \\ &= \frac{\mathrm{E}\langle u, v\rangle_{L_2(\mathrm{P_n})}}{\left[\ \mathrm{E}\lVert u\rVert_{L_1(\mathrm{P_n})}\ \right]^2} \quad \text{for} \quad u(w,x) = \{w - \pi(x)\}^2 \\ &\qquad\qquad\qquad\qquad \text{and} \quad v(w,x) = \{\beta(x) - \hat{\beta}(x)\}^2. \end{aligned}$$

We can bound this using Hölder's inequality.

Bounding the Variance (Option 1)

$$n \times V = \frac{\mathrm{E}\langle u, v\rangle_{L_2(\mathrm{P_n})}}{\left[\ \mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \right]^2} \quad \text{for} \quad u(w,x) = \{w - \pi(x)\}^2$$

$$\text{and} \quad v(w,x) = \{\beta(x) - \hat{\beta}(x)\}^2.$$

Idea: $\langle u, v\rangle_{L_2(\mathrm{P_n})} \le \underbrace{\|u\|_{L_\infty(\mathrm{P_n})}}_{\le \|u\|_\infty} \|v\|_{L_1(\mathrm{P_n})}$

Bounding the Variance (Option 1)

$$n \times V = \frac{\mathrm{E}\langle u, v\rangle_{L_2(\mathrm{P_n})}}{\left[\ \mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \right]^2} \quad \text{for} \quad u(w,x) = \{w - \pi(x)\}^2$$

$$\text{and} \quad v(w,x) = \{\beta(x) - \hat\beta(x)\}^2.$$

Idea: $\langle u, v\rangle_{L_2(\mathrm{P_n})} \le \underbrace{\|u\|_{L_\infty(\mathrm{P_n})}}_{\le \|u\|_\infty} \|v\|_{L_1(\mathrm{P_n})}$

$$\begin{aligned} n \times V &\le \frac{\mathrm{E}[\|u\|_\infty \ \|v\|_{L_1(\mathrm{P_n})}]}{\left[\ \mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \right]^2} \\ &\le \frac{\mathrm{E}[1 \cdot \|v\|_{L_1(\mathrm{P_n})}]}{\left[\ \mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \right]^2} \\ &= \frac{\|\beta - \hat\beta\|_{L_2(P)}}{\left[\ \mathrm{E}\ \frac{1}{n}\sum_i \{W_i - \pi(X_i)\}^2\ \right]^2} \end{aligned}$$

Note. $\mathrm{E}\ \|u\|_{L_1(\mathrm{P_n})} = \frac{1}{n}\sum_i \mathrm{E}\{W_i - \pi(X_i)\}^2 = \frac{1}{n}\sum_i \mathrm{E}\,\mathrm{Var}[W_i \mid X_i]$

Bounding the Variance (Option 2)

$$n \times V = \frac{\mathrm{E}\langle u, v\rangle_{L_2(\mathrm{P_n})}}{\left[\, \mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\,\right]^2} \quad \text{for} \quad u(w,x) = \{w - \pi(x)\}^2$$

$$\text{and} \quad v(w,x) = \{\beta(x) - \hat\beta(x)\}^2.$$

Idea: $\langle u, v\rangle_{L_2(\mathrm{P_n})} \le \|u\|_{L_1(\mathrm{P_n})} \underbrace{\|v\|_{L_\infty(\mathrm{P_n})}}_{\le \|v\|_\infty}$

This approach gives us a bound that blows up less when you have a not-all-that-random treatment assignment, i.e. small $\mathrm{Var}[W_i \mid X_i]$, but involves a norm $\|\cdot\|_\infty$ on $\beta - \hat\beta$ that's both bigger and harder to analyze than the two-norm that we had in our first bound.

Bounding the Variance (Option 2)

$$n \times V = \frac{\mathrm{E}\langle u, v\rangle_{L_2(\mathrm{P_n})}}{\left[\ \mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \right]^2} \quad \text{for} \quad u(w,x) = \{w - \pi(x)\}^2$$

$$\text{and} \quad v(w,x) = \{\beta(x) - \hat\beta(x)\}^2.$$

Idea: $\langle u, v\rangle_{L_2(\mathrm{P_n})} \le \|u\|_{L_1(\mathrm{P_n})} \underbrace{\|v\|_{L_\infty(\mathrm{P_n})}}_{\le \|v\|_\infty}$

$$\begin{aligned} n \times V &\le \frac{\mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \|v\|_{L_2(\mathrm{P_n})}}{\left[\ \mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \right]^2} \\ &\le \frac{\mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \|v\|_\infty}{\left[\mathrm{E}\|u\|_{L_1(\mathrm{P_n})}\ \right]^2} \\ &\le \frac{\|v\|_\infty}{\mathrm{E}\|u\|_{L_1(\mathrm{P_n})}} \\ &= \frac{\|\beta - \hat\beta\|_\infty^2}{\frac{1}{n}\sum_i \mathrm{E}\,\mathrm{Var}[W_i \mid X_i]} \end{aligned}$$

This approach gives us a bound that blows up less when you have a not-all-that-random treatment assignment, i.e. small $\mathrm{Var}[W_i \mid X_i]$, but involves a norm $\|\cdot\|_\infty$ on $\beta - \hat\beta$ that's both bigger and harder to analyze than the two-norm that we had in our first bound.

Conclusion

The difference $\hat{\tau}_{\hat{\beta}} - \hat{\tau}_{\beta}$ (more or less, i.e. ignoring Q) has mean zero and ...

standard deviation $\sqrt{V} \le \frac{\|\beta - \hat{\beta}\|_{L_2(\mathrm{P})}}{\sqrt{n}} \times \frac{1}{\frac{1}{n}\sum_i \mathrm{E}\,\mathrm{Var}[W_i \mid X_i]}$

Therefore, if we have a consistent estimate of β, i.e. if $\|\beta - \hat{\beta}\|_{L_2(\mathrm{P})} \to 0$ ...

- this difference is negligible relative to $1/\sqrt{n}$, i.e., $(\hat{\tau}_{\hat{\beta}} - \hat{\tau}_{\beta})/(1/\sqrt{n}) \to 0$
- and therefore negligible relative to the random variation of the oracle estimator $\hat{\tau}_{\beta}$, which has standard deviation proportional to $1/\sqrt{n}$.

One implication is that, in large samples, it doesn't matter whether you use the actual estimator $\hat{\tau}_{\hat{\beta}}$ or the oracle estimator $\hat{\tau}_{\beta}$. They have the same asymptotic distribution.