Part 2: Confounding and regression adjustments.

We want to estimate average treatment effects in a setting where **treatment assignment** may be associated with pre-treatment covariates X.

- How can we flexibly **"control"** for X?
- Under what **conditions** is controlling for X enough?

We want to estimate average treatment effects in a setting where **treatment assignment** may be associated with pre-treatment covariates X.

- How can we flexibly **"control"** for X?
- Under what **conditions** is controlling for X enough?

The Assumption: Controlling for X is enough if treatment is **as good as random** conditionally on X.

The Question: What methods enable **inference** about the average treatment effect given this assumption?

Covariates and unconfoundedness

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** $W_i \in \{0, 1\}$, with **potential outcomes** $Y_i(0)$ and $Y_i(1)$ such that $Y_i = Y_i(W_i)$.

Controlling for X_i is sufficient for identifying average treatment effects if W_i is **as good as random** once we condition on X_i:

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

This assumption is commonly referred to as **unconfoundedness**, or selection on observables (Rosenbaum & Rubin, 1983).

Covariates and unconfoundedness

We've assumed that W_i is **as good as random** once we condition on X_i:

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

We want to estimate the **average treatment effect**

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

How should we proceed?

Regression adjustments under unconfoundedness

Given **unconfoundedness**

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

we can express the ATE in terms of conditional response surfaces,

$$\begin{aligned}
\tau &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i \,\middle|\, X_i,\, W_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, X_i,\, W_i = 0\right]\right] \\
&= \mathbb{E}\left[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)\right],
\end{aligned}$$

where $\mu_{(w)}(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, W_i = w\right]$.

Regression adjustments and unconfoundedness

Given **unconfoundedness**, we know that

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

This suggests an estimation strategy:

1. Learn $\hat{\mu}_{(0)}(x)$ by predicting Y from X on controls.
2. Learn $\hat{\mu}_{(1)}(x)$ by predicting Y from X on treated units.
3. Estimate $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right)$.

This is "obviously" **consistent** if $\hat{\mu}_{(w)}(x)$ is consistent for $\mu_{(w)}(x)$.

But is this any good?

Regression adjustments: The classical approach

A classical approach to the ATE involves estimating $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$ via **ordinary least-squares regression** (OLS).

We first **posit a linear model**, $\mu_{(w)}(x) = x\beta_{(w)}$. We then **fit the model** as follows (using `R` notation):

$$\hat{\beta}_{(0)} \leftarrow \texttt{lm}\left(Y_i \sim X_i,\ \texttt{subset}\ \ W_i = 0\right),$$
$$\hat{\beta}_{(1)} \leftarrow \texttt{lm}\left(Y_i \sim X_i,\ \texttt{subset}\ \ W_i = 1\right).$$

Finally, we make **predictions** $\hat{\mu}_{(w)}(x) = \hat{\beta}_{(w)}x$, and obtain a treatment effect estimate as

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right) = \left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right)\overline{X},$$

where $\overline{X} = \sum_{i=1}^{n} X_i$. Note that, X implicitly includes an **intercept**.

The classical approach: Pros and cons

The OLS approach hinges on having a **well specified linear model**

$$\mu_{(w)}(x) = x\beta_{(w)}.$$

Pro: The method is simple, familiar, and well justified when the above linear model holds.

Con: No guarantees if the linear model doesn't hold.

Regression adjustments: The machine learning approach

A modern, non-parametric approach seeks to **avoid extraneous assumptions** on the regression functions $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$.

Recall: Given our setting, the **optimal prediction** of Y_i given $X_i = x$ and $W_i = w$ is $\mu_{(w)}(x)$. Idea:

- Pick your favorite machine learning method, and use it to predict Y_i from X_i and W_i.
- Use these predictions as estimates for $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$.
- Estimate $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \right)$.

In many settings, machine learning methods enable accurate prediction without needing to model the shape of $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$.

Regression adjustments and unconfoundedness

We're interested in estimating the ATE by fitting $\hat{\mu}_{(w)}(x)$:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \right).$$

Let's try this out! Consider two estimators,

1. Fit $\hat{\mu}_{(w)}(x)$ via linear regression, or
2. Fit $\hat{\mu}_{(w)}(x)$ via a random forest,

and two simulation settings,

1. The functions $\mu_{(w)}(x)$ are linear, or
2. The functions $\mu_{(w)}(x)$ are non-linear.

Approach #1: Use OLS for estimation

```
library(sandwich) # for robust standard errors

# First center the X, then run OLS with full W:X
# interactions. With this construction, the
# W-coefficient can be interpreted as ATE.
X.centered = scale(X, center = TRUE, scale = FALSE)
ols.fit = lm(Y ~ W * X.centered)

# Use robust standard errors
tau.hat = coef(ols.fit)["W"]
tau.se = sqrt(sandwich::vcovHC(ols.fit)["W", "W"])
print(paste0("95% CI: ", round(tau.hat),
             " +/- ", round(1.96 * tau.se)))
```

Approach #2: Use random forests for estimation

Fit $\hat{\mu}_{(w)}(x)$ using random forests, and set

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \right).$$

The following example uses out-of-bag predictions when relevant.

```
library(grf)
rf.0 = regression_forest(X[W==0,], Y[W==0])
mu.hat.0 = predict(rf.0, X)$predictions
mu.hat.0[W==0] = predict(rf.0)$predictions

rf.1 = regression_forest(X[W==1,], Y[W==1])
mu.hat.1 = predict(rf.1, X)$predictions
mu.hat.1[W==1] = predict(rf.1)$predictions

tau.hat.rf = mean(mu.hat.1 - mu.hat.0)
```

A simulation comparison

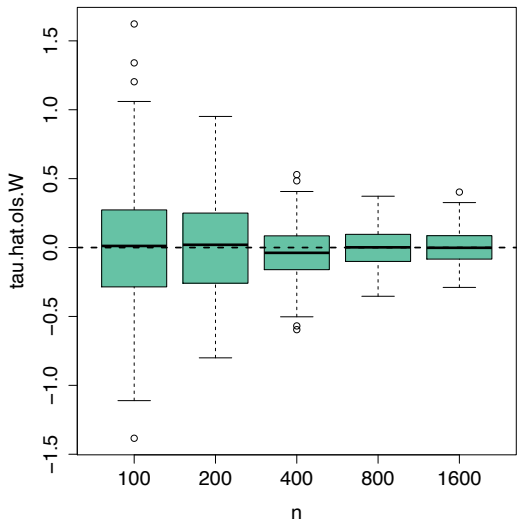
Linear setting:

$$
\begin{aligned}
&X \sim \mathcal{N}\left(0,\, I_{20\times 20}\right)\\
&\mathbb{P}\left[W=1 \,\middle|\, X\right] = 1/\left(1+e^{-X_1}\right)\\
&Y(0) = X_1 + X_2 + \mathcal{N}\left(0,\, 4\right),\\
&Y(1) = X_1 + X_3 + \mathcal{N}\left(0,\, 4\right)
\end{aligned}
$$

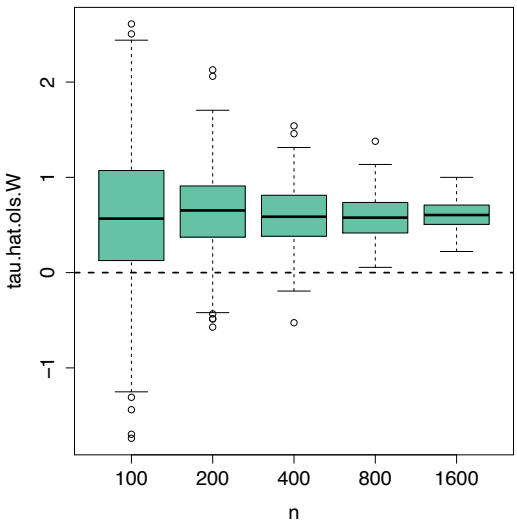
Non-linear setting:

$$
\begin{aligned}
&X \sim \mathcal{N}\left(0,\, I_{20\times 20}\right)\\
&\mathbb{P}\left[W=1 \,\middle|\, X\right] = \left(1+\sin\left(X_1\right)\right)/2\\
&Y(0) = 4\cdot 1\left(\{X_1>0\}\right) + X_2^2/2 + \mathcal{N}\left(0,\, 4\right),\\
&Y(1) = 4\cdot 1\left(\{X_1>0\}\right) + X_3^2/2 + \mathcal{N}\left(0,\, 4\right)
\end{aligned}
$$

Evaluating OLS

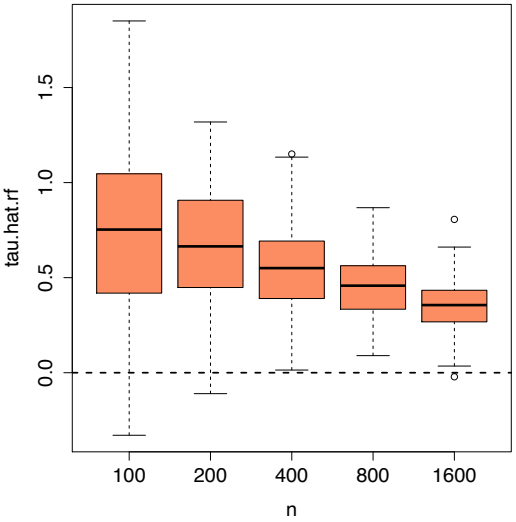


linear setting

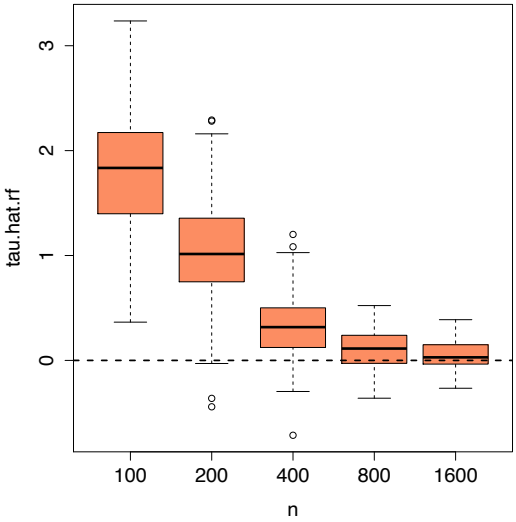


non-linear setting

Evaluating Random Forests



linear setting



non-linear setting

Linear Regression vs Random Forests

Linear Regression (OLS) bets everything on $\hat{\mu}_{(w)}(x)$ being linear:

- If this assumption is valid, everything works out perfectly. Parametric reasoning applies. No need to worry.
- If this assumption fails, everything is wrong. Nothing to do.
- In either case, not much subtlety in how one should do inference

Machine learning methods, such as **random forests**, seek to fit potentially complicated functions $\hat{\mu}_{(w)}(x)$:

- Never completely wrong: expect consistency as $n \to \infty$.
- But even when a simple model is correct, converge slowly.
- It is possible to extract useful insights, but one must be robust to suboptimal finite sample performance.

Linear Regression vs Random Forests

We've considered estimating the **average treatment effect** as

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \right),$$

with $\hat{\mu}_{(w)}(x)$ estimated using either OLS or random forests.

Neither method so far is particularly good:

- Averaging with **OLS** is not robust, because it requires **linearity**.
- Averaging with **RF** is not robust, because it does not account for **finite sample errors** of RF.