Social Data Science

Soc 114
Winter 2025

Supervised Machine Learning
Illustration with Trees

Learning goals for today

By the end of class, you will be able to

- understand the notion of supervised machine learning
  - an input-output machine
  - learned on some learning cases
  - used to predict for new cases
- apply that notion to the specific case of regression trees

Prediction function

A **prediction function** is an input-output function:

- input a vector of predictors $\vec{x}$
- output a predicted outcome $\hat{y} = \hat{f}(\vec{x})$

	Input	**Output**
	$\vec{x} \longrightarrow$	$\hat{y}$
Example:	{Sex,Age}	Probability of Employment Given Sex and Age

	Age	Sex	Employed
cases for learning	26	F	1
	40	M	1
	61	M	0
	32	F	1
case to predict	63	F	?

OLS is a prediction function

Input $\vec{x} \rightarrow$ Output $\hat{y}$

$$\hat{y} = \hat{f}(\vec{x}) = \hat{\beta}_0 + \hat{\beta}_1(\text{Sex} = \text{Male}) + \hat{\beta}_2(\text{Age})$$

- Learn $\hat{f}$ in a **learning sample** with $\{\vec{x}_i, y_i\}_{i=1}^n$
  - Computer finds $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ that predict well in the learning sample
- At a new $\vec{x}$ value, predict $\hat{f}(\vec{x})$

Logistic regression is a prediction function

Input $\vec{x} \rightarrow$ Output $\hat{y}$

$$\hat{y} = \hat{f}(\vec{x}) = \text{logit}^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1(\text{Sex} = \text{Male}) + \hat{\beta}_2(\text{Age})\right)$$

- Learn $\hat{f}$ in a **learning sample** with $\{\vec{x}_i, y_i\}_{i=1}^n$
  - Computer finds $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ that predict well in the learning sample
- At a new $\vec{x}$ value, predict $\hat{f}(\vec{x})$

Matching is a prediction function

Input $\vec{x} \rightarrow$ Output $\hat{y}$

$$\hat{y} = \hat{f}(\vec{x}) = y_j$$

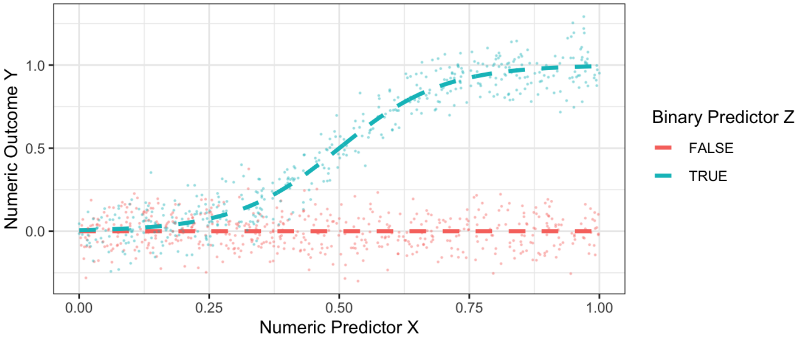
where unit j is the best match among the learning sample, which minimizes a distance from the case to predict: $d(\vec{x}, \vec{x}_j)$ is small

- Learn $\hat{f}$ in a **learning sample** with $\{\vec{x}_i, y_i\}_{i=1}^n$
  - Computer finds j with $\vec{x}_j$ most similar to $\vec{x}$
- At a new $\vec{x}$ value, predict $\hat{f}(\vec{x})$

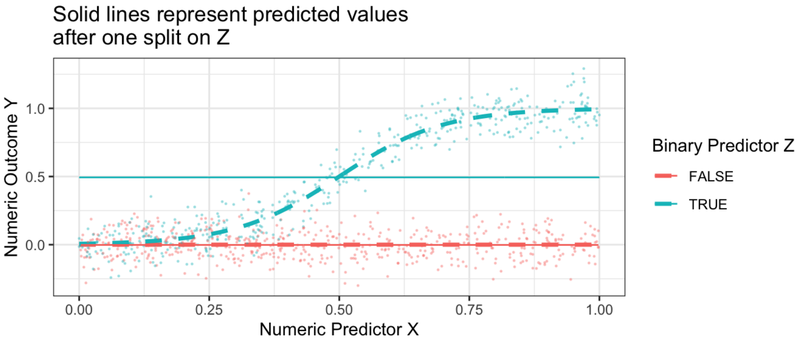
There are many prediction functions

- input a vector of predictors $\vec{x}$
- output a predicted outcome $\hat{y} = \hat{f}(\vec{x})$

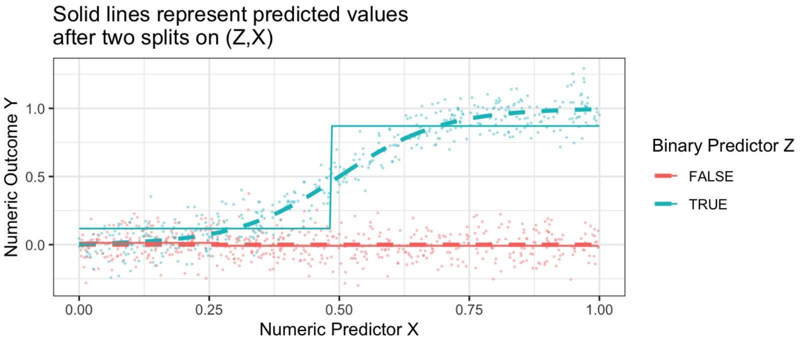
Trees as a prediction function



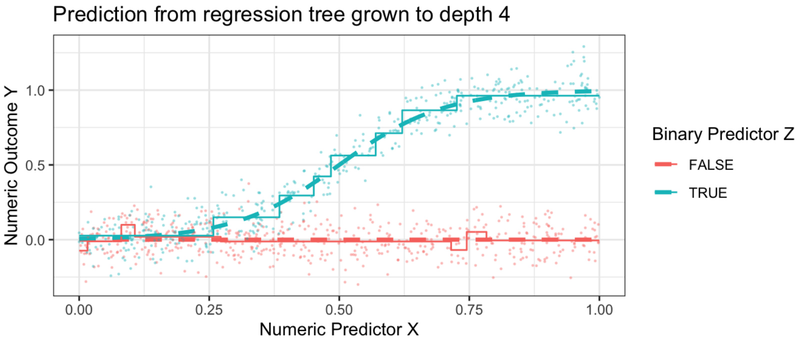
Trees as a prediction function



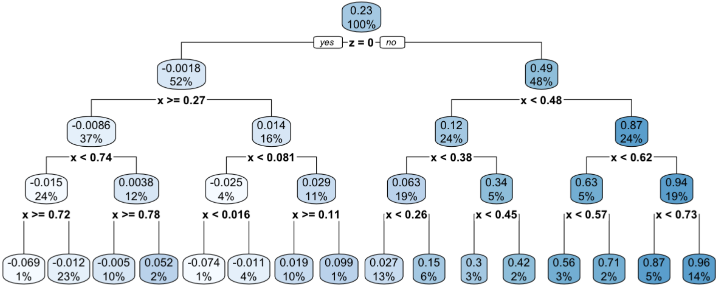
Trees as a prediction function



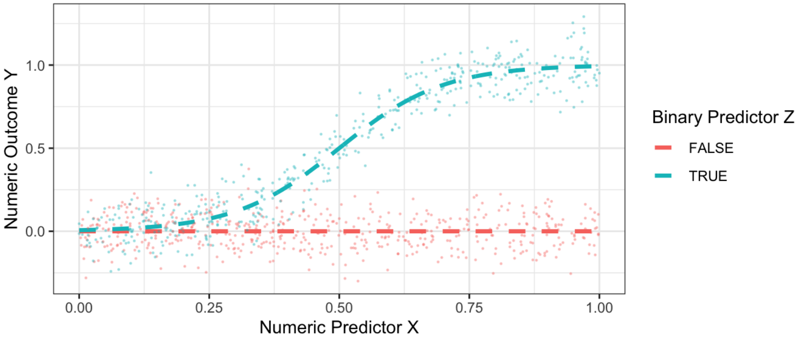
Trees as a prediction function



Trees as a prediction function

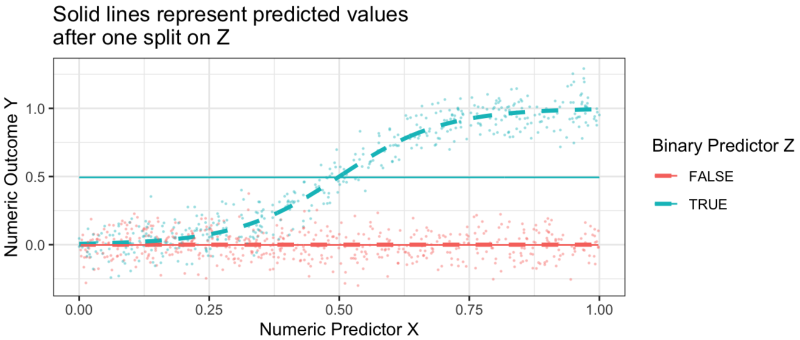


Trees as a prediction function: How that worked

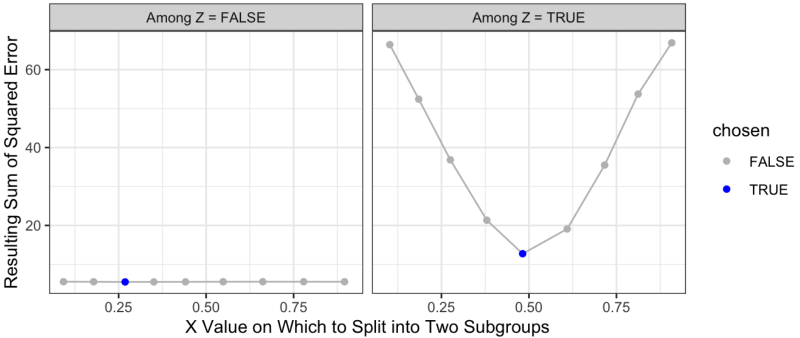


Trees as a prediction function: How that worked

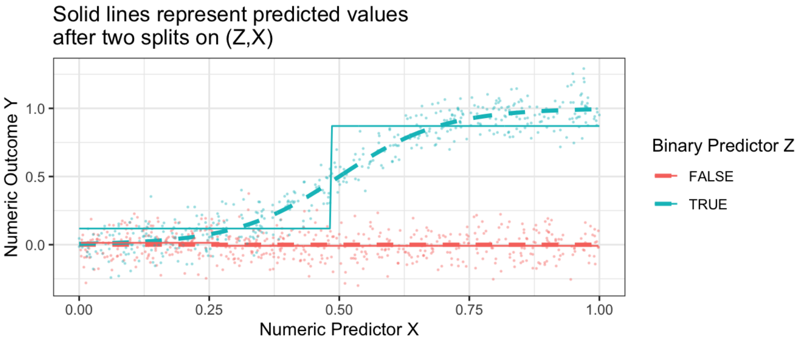
Trees as a prediction function: How that worked



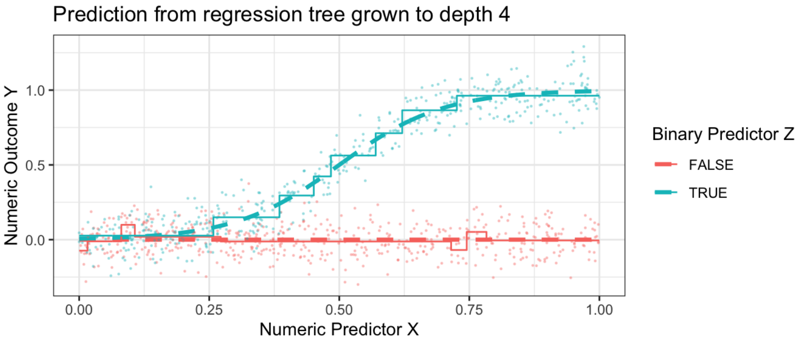
Trees as a prediction function: How that worked



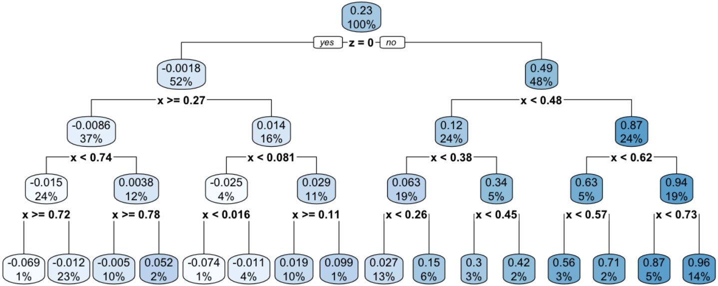
Trees as a prediction function: How that worked



Trees as a prediction function: How that worked



Trees as a prediction function: How that worked



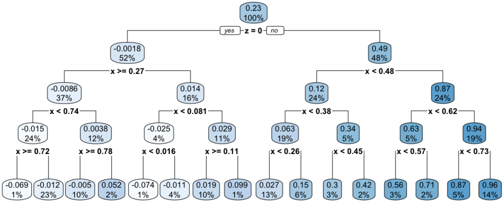
Trees as a prediction function: How that worked. **Summary.**

1. Begin with all data
2. Consider many ways to partition into two parts
3. Estimate the mean squared prediction error for each: $\mathrm{E}((\hat{Y} - Y)^2)$
4. Choose the split that minimizes mean squared prediction error

Repeatedly, apply steps (1–4) to each subgroup.
Stop by a data-driven rule.

Trees: Some terminology

- Branch = one direction of a split
- Leaf = terminal node at the bottom



When presented with a new case, find its leaf.
Predict the mean of Y among learning cases in that leaf.

A tree can be interpretable: Realistic example

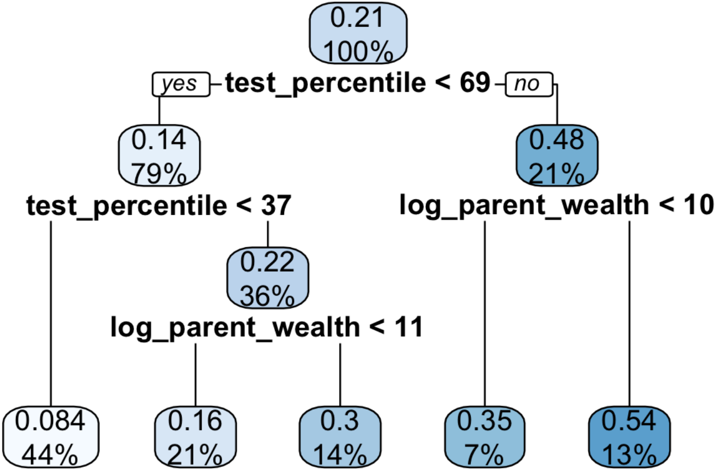
- Outcome: Has spouse or partner with BA degree at age 35
- Predictors: Demographics and measures of family background

A tree can be interpretable: Realistic example

```r
library(tidyverse)
library(rpart)
library(rpart.plot)

all_cases <- read_csv("https://soc114.github.io/data/nlsy97_simulated.csv")

rpart.out <- rpart(
  y ~ sex + race + mom_educ + dad_educ + log_parent_income +
    log_parent_wealth + test_percentile,
  data = all_cases
)

rpart.plot(rpart.out)
```

A tree can be interpretable: Realistic example

Y = has spouse or partner with BA degree at age 35



Pruning a tree

Sometimes you want a simpler decision rule

- you worry you are fitting to noise
- you want to explain predictions more easily

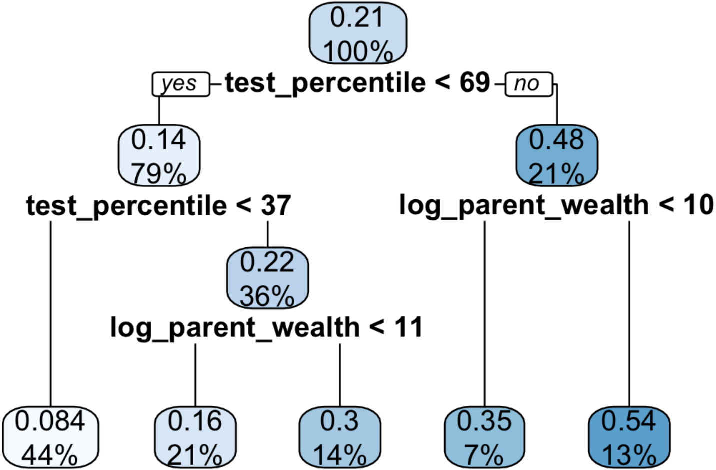
Pruning a tree

Sometimes you want a simpler decision rule

- you worry you are fitting to noise
- you want to explain predictions more easily

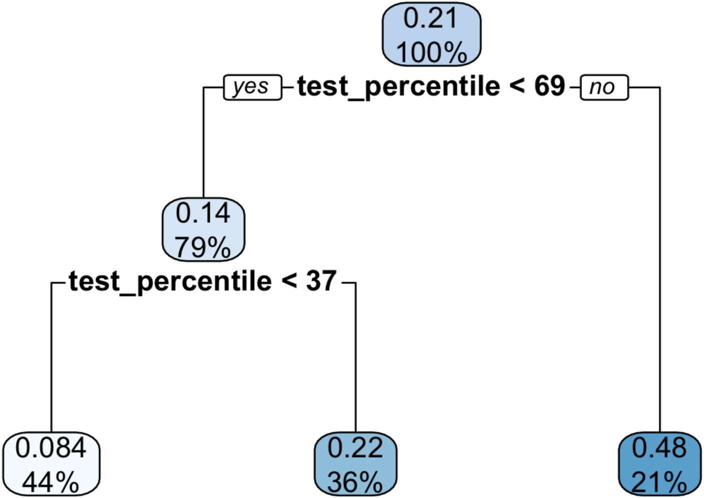
Then you prune the tree: Trim back some branches

Pruning a tree: Original tree



Pruning a tree: Pruned tree

pruned <- prune(rpart.out, cp = .02)



Discussion: Why prefer a tree vs OLS?

- Reasons to prefer a tree

- Reasons to prefer OLS

Discussion: Why prefer a tree vs OLS?

- Reasons to prefer a tree
  - No need to assume a functional form
  - Easy to explain how a prediction is made: follow the decision branches
- Reasons to prefer OLS

Discussion: Why prefer a tree vs OLS?

- Reasons to prefer a tree
  - No need to assume a functional form
  - Easy to explain how a prediction is made: follow the decision branches
- Reasons to prefer OLS
  - More widely known in social science
  - Better if the functional form is correct

From regression to causal trees

What step would change if our goal was to discover heterogeneous causal effects?

Regression Trees

1. Begin with all data.
2. Split to two sides with very different average value of Y.
3. Repeat 1–2 on each leaf until a stopping rule is reached.

From regression to causal trees

What step would change if our goal was to discover heterogeneous causal effects?

Regression Trees

1. Begin with all data.
2. Split to two sides with very different average value of Y.
3. Repeat 1–2 on each leaf until a stopping rule is reached.

Causal Trees

1. Begin with all data.
2. Split to two sides with very different average value of $Y^1 - Y^0$.
3. Repeat 1–2 on each leaf until a stopping rule is reached.

Athey, S. & G. Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS*.

Causal trees in randomized experiments

Setting:

- Many pre-treatment variables $\vec{X}$
- Randomized treatment A

Procedure:

- In sample 1, partition into leaves.
- In sample 2, estimate effects within leaves by difference in means.

Causal trees in observational studies

Setting:

- Many pre-treatment variables $\vec{X}$
- Non-randomized treatment A

Procedure:

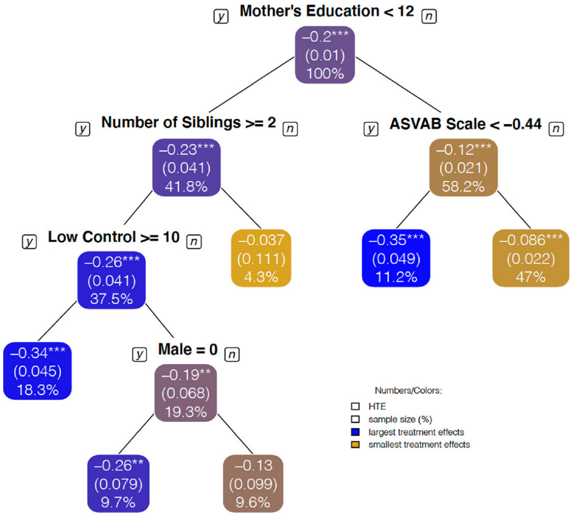
- In sample 1, partition into leaves.
- In sample 2, estimate effects within leaves by difference in means, adjusted for confounding by IPW or matching.

Brand, Xu, Koch, & Geraldo. 2021. "Uncovering sociological effect heterogeneity using tree-based machine learning." Sociological Methodology, 51(2), 189-223.

Causal trees in observational studies

Brand, Xu, Koch, & Geraldo (2021)

Causal question: Effect of college completion on the proportion of time in low-wage work.



Causal trees in observational studies

The setting:

- Many pre-treatment variables $\vec{X}$
- Non-randomized treatment A
- Conditional exchanngeability holds

The procedure

- One sample: Learn the tree
- Learn propensity score function
- New sample: Inverse-probability-weighted or matching estimates in each leaf

Recap: Machine learning as an input-output function

Input		**Output**
$\vec{x}$	$\longrightarrow$	$\hat{y}$

Example: {Sex,Age} — Probability of Employment Given Sex and Age

cases for learning

Age	Sex	Employed
26	F	1
40	M	1
61	M	0
32	F	1

case to predict

Age	Sex	Employed
63	F	?

Learning goals for today

By the end of class, you will be able to

- understand the notion of supervised machine learning
  - an input-output machine
  - learned on some learning cases
  - used to predict for new cases
- apply that notion to the specific case of regression trees