“Location, location, location!”

Regression Trees for Housing Data

Boston

- Capital of the state of Massachusetts, USA
- First settled in 1630
- 5 million people in greater Boston area, some of the highest population densities in America.

Boston

Housing Data

- A paper was written on the relationship between **house prices** and **clean air** in the late 1970s by David Harrison of Harvard and Daniel Rubinfeld of U. of Michigan.
- "Hedonic Housing Prices and the Demand for Clean Air" has been cited ~1000 times
- Data set widely used to evaluate algorithms.

The R in CART

- In the lecture we mostly discussed **classification trees** – the output is a factor/category

- Trees can also be used for **regression** – the output at each leaf of the tree is no longer a category, but a number

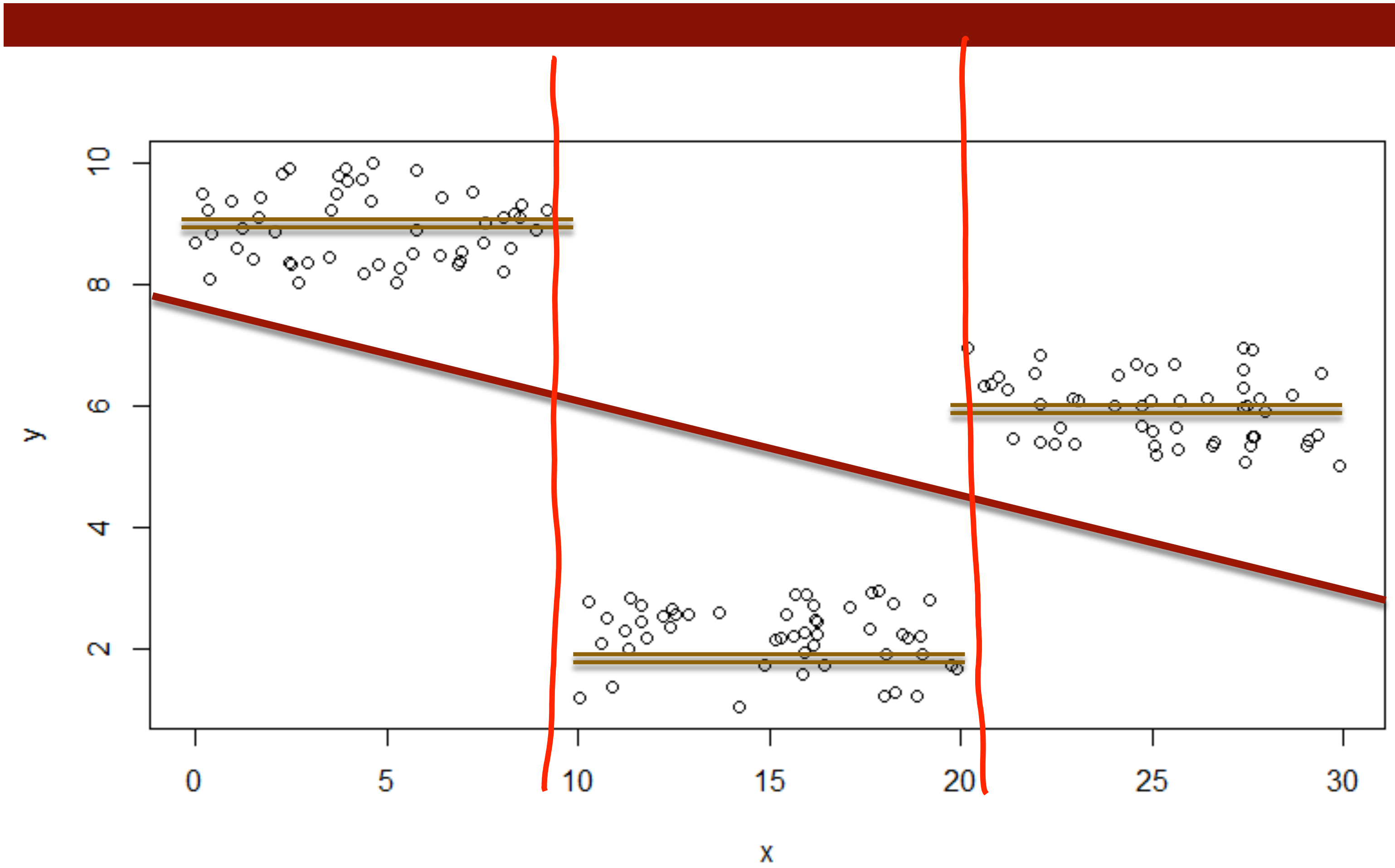
- Just like classification trees, **regression trees** can capture **nonlinearities** that linear regression can't.

Regression Trees

- With Classification Trees we report the average outcome at each leaf of our tree, e.g. if the outcome is "true" 15 times, and "false" 5 times, the value at that leaf is: $\frac{15}{15+5} = 0.75 \geq 0.5 \rightarrow \text{true}$

- With Regression Trees, we have continuous variables, so we simply report the average of the values at that leaf: $3, 4, 5 = 4$

Example



Housing Data

- We will explore the dataset with the aid of trees.
- Compare linear regression with regression trees.
- Discussing what the "cp" parameter means.
- Apply cross-validation to regression trees.

Understanding the data

- Each entry corresponds to a census **tract**, a statistical division of the area that is used by researchers to break down towns and cities.
- There will usually be multiple census tracts per **town**.
- **LON** and **LAT** are the longitude and latitude of the center of the census tract.
- **MEDV** is the median value of owner-occupied homes, in thousands of dollars.

Understanding the data

- **CRIM** is the per capita crime rate
- **ZN** is related to how much of the land is zoned for large residential properties
- **INDUS** is proportion of area used for industry
- **CHAS** is 1 if the census tract is next to the Charles River
- **NOX** is the concentration of nitrous oxides in the air
- **RM** is the average number of rooms per dwelling

Understanding the data

- **AGE** is the proportion of owner-occupied units built before 1940
- **DIS** is a measure of how far the tract is from centers of employment in Boston
- **RAD** is a measure of closeness to important highways
- **TAX** is the property tax rate per $10,000 of value
- **PTRATIO** is the pupil-teacher ratio by town

The "cp" parameter

- "cp" stands for "**complexity parameter**"
- Recall the first tree we made using LAT/LON had many splits, but we were able to trim it without losing much accuracy.
- Intuition: having too many splits is bad for generalization, so we should penalize the **complexity**

The "cp" parameter

- Define **RSS**, the **residual sum of squares**, the sum of the square differences
$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$
- Our goal when building the tree is to minimize the RSS by making splits, but we want to penalize too many splits. Define **S** to be the number of splits, and λ (lambda) to be our penalty. Our goal is to find the tree that minimizes
$$\sum_{Leaves} (\text{RSS at each leaf}) + \lambda S$$

The “cp” parameter

- λ (lambda) = 0.5

Splits	RSS	Total Penalty
0	5	5
1	2 + 2 = 4	4 + 0.5*1 = 4.5
2	1+0.8+2 = 3.8	3.8 + 0.5*2 = 4.8

The “cp” parameter

$$\sum_{Leaves} (\text{RSS at each leaf}) + \lambda S$$

- If pick a large value of λ, we won’t make many splits because we pay a big price for every additional split that outweighs the decrease in “error”

- If we pick a small (or zero) value of λ, we’ll make splits until it no longer decreases error.

The “cp” parameter

- The definition of “cp” is closely related to λ

- Consider a tree with no splits – we simply take the average of the data. Calculate RSS for that tree, let us call it **RSS(no splits)**

$$c_p = \frac{\lambda}{\text{RSS(no splits)}}$$