Specification and Estimation of CART

$$\hat{f}(x) = \sum_{m=1}^{|T|} \mathbb{I}_{R_m}(x) \cdot \hat{y}_m$$

- $|T|$ is the number of leaves in the tree
- R_m is the set of values of x in the m'th leaf
- $\mathbb{I}_{R_m}(x) = \begin{cases} 1 & \text{if } x \text{ is in region } R_m \\ 0 & \text{otherwise} \end{cases}$
- $\hat{y}_m$ is the estimated function value for leaf m.

Temp. vs. O3 example

Parameters to Estimate:

Split Points: where do we make splits? Determines R_m

Regression Constants: In each leaf, what is $\hat{y}_m$?

Optimization Target for Regression:

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (\hat{y}_{im} - y_i)^2$$

Optimization Target for Classification:

Often use Gini Index:

$1 - \sum_{k=1}^{K} p_{km}$ where p_{km} is proportion of obs. in region m that are in class k.

Top down Estimation Algorithm:

1. Initialize tree with no splits
   - $\hat{y}$ is mean of all observations
   - calculate RSS for this "tree"

2. Repeat until a stopping criteria is met:
   - For every leaf, try every possible split at the midpoint of values of x in that leaf;
   - calculate RSS based on that split
   - select the split that gives the largest → reduction in RSS

Possible stopping criteria:

- all leaves have 5 or fewer obs.
  ↑ or some other #
- a maximum # of leaves has been reached
- a max. depth has been reached
- No reduction in RSS larger than λ can be achieved.

Regularization / Penalization:

minimize $\text{RSS} + \lambda |T|$ ↑ # of leaves

R package minimizes

$$\approx -R^2 + \lambda |T| = -\left(1 - \frac{\text{RSS}}{\text{TSS}}\right) + \lambda |T|$$

$$= \frac{\text{RSS}}{\text{TSS}} - 1 + \lambda |T|$$