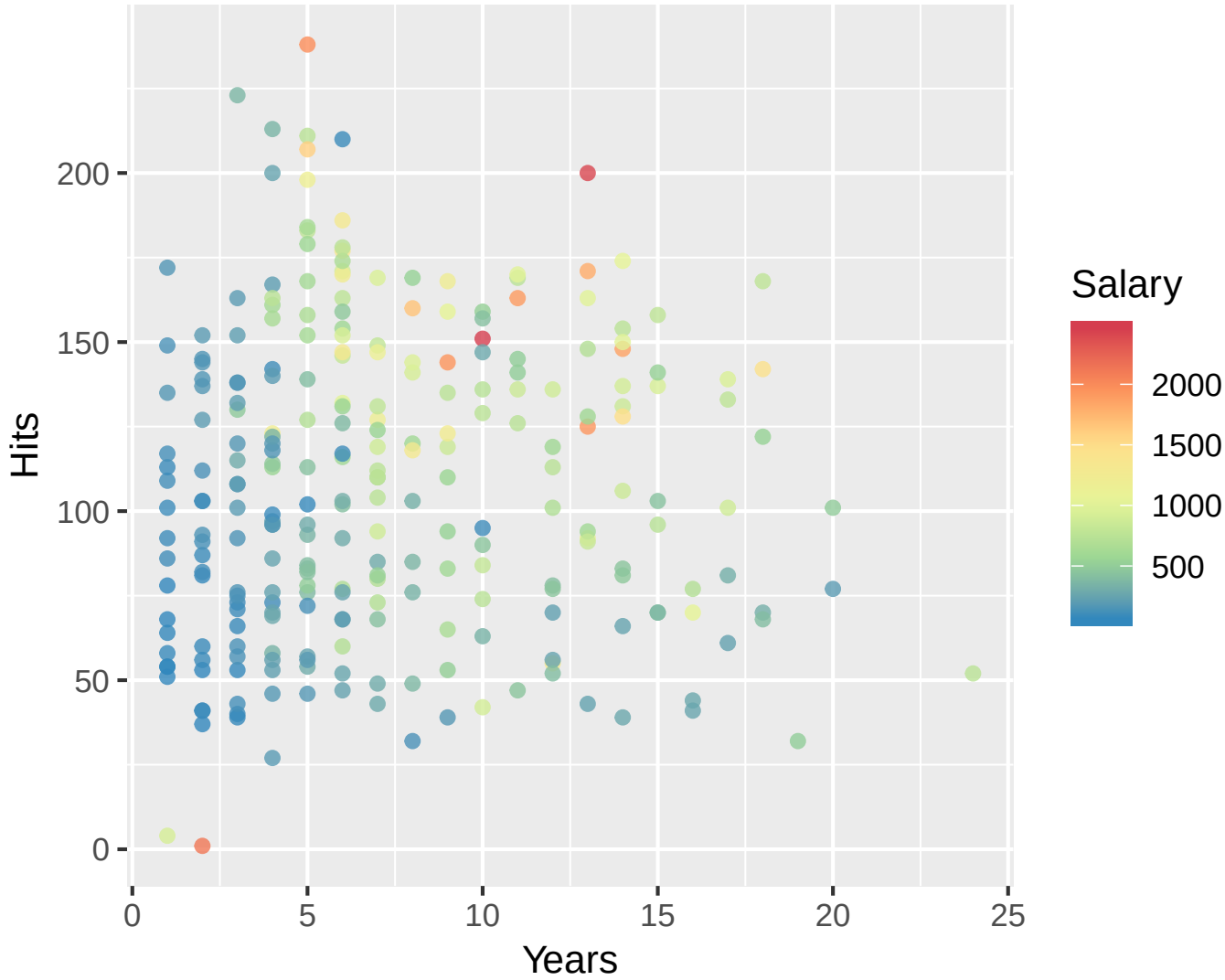
Lecture 14: Regression Trees
STAT 452, Spring 2021

Tree-Based Methods

- Here we describe *tree-based* methods. These involve *stratifying* or *segmenting* the predictor space into a number of simple regions.
- Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these types of approaches are known as *decision-tree* methods.
- Decision trees can be applied to both regression and classification problems. We will first consider regression problems, and then move on to classification.

Example: Baseball Salary Data



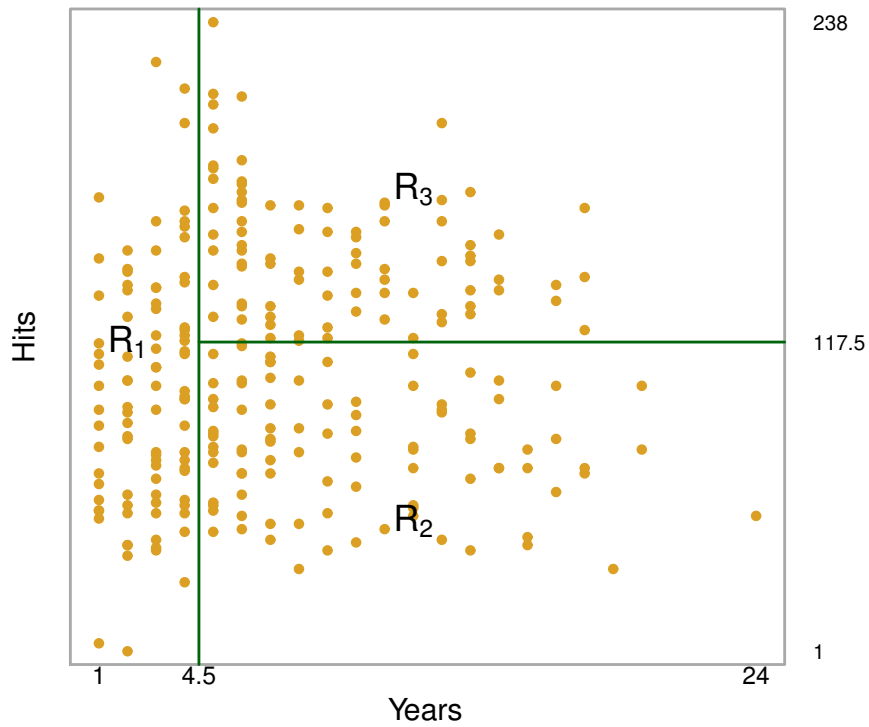
Example: Baseball Salary Data

Example: Baseball Salary Data

- The previous slide shows a regression tree for predicting the log salary of a baseball player, based on `Years`, the number of years that he has played in the major league, and `Hits`, the number of hits that he made in the previous year.
- The split at the top of the tree results in two large branches. The left-hand branch corresponds to `Years` $<$ `4.5`, and the right-hand branch corresponds to `Years` ≥ 4.5.
- The tree has two internal nodes and three terminal nodes, or leaves.
- The number in each leaf is the mean of the response for the observations that fall there.

Example: Baseball Salary Data

The tree segments the players into three regions of the predictor space:
$R_1 = \{X | \texttt{Years} < 4.5\}$
$R_2 = \{X | \texttt{Years} \geq 4.5, \texttt{Hits} < 117.5\}$
$R_3 = \{X | \texttt{Years} \geq 4.5, \texttt{Hits} \geq 117.5\}$



Terminology for Trees

- The regions R_1, R_2, and R_3 are known as the *terminal nodes* or *leaves* of the tree.
- The points along the tree where the predictor space is split are referred to as *internal nodes*.
- In the hitters tree, the two internal nodes are indicated by the text `Years < 4.5` and `Hits < 117.5`.
- We refer to the segments of the trees that connect the nodes as *branches*.

Interpretation of Results

- `Years` is the most important factor in determining `Salary`, and players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced, the number of `Hits` that he made in the previous year seems to play little role in his `Salary`.
- But among players who have been in the major leagues for five or more years, the number of `Hits` made in the previous year does affect `Salary`, and players who made more `Hits` last year tend to have higher salaries.
- Surely an over-simplification, but compared to a regression model, it is easy to display, interpret and explain

Details of Tree Building Process

- The algorithm used to build, or estimate, a regression tree is called *recursive binary splitting*.
- The main idea of this approach is that at each iteration the algorithm searches over all predictors and possible cutpoints of those predictors in each subregion. The algorithm then selects the predictor and cutpoint that results in the lowest residual sum of squares (RSS) on the training data.
- For a details see Chapter 8, pp. 306 - 307, of *An Introduction to Statistical Learning*

Details of Tree Building Process

- How large of a tree should we grow?
- A very large tree might overfit the data, while a small tree might not capture important structure.
- One strategy to prevent over-fitting, is to grow a very large tree, and then *prune* it back to obtain an optimal subtree.
- This pruning method is implemented by the `rpart` package.