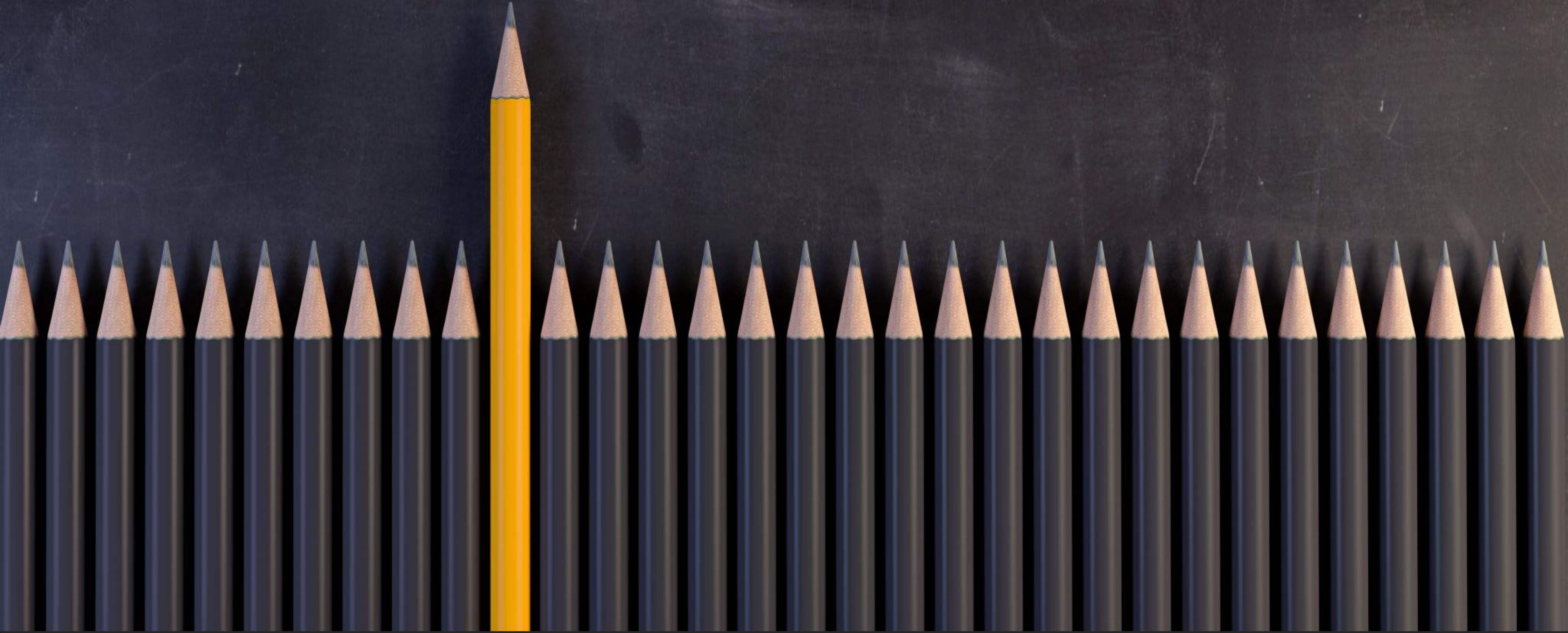


Regression Tree

ECON 258 DATA ANALYTICS WITH R

Working with Algorithms

The type of algorithm that you have previously seen:

- **Regression**
- **K-nearest Neighbor**
- **Today, we will learn about Trees!**

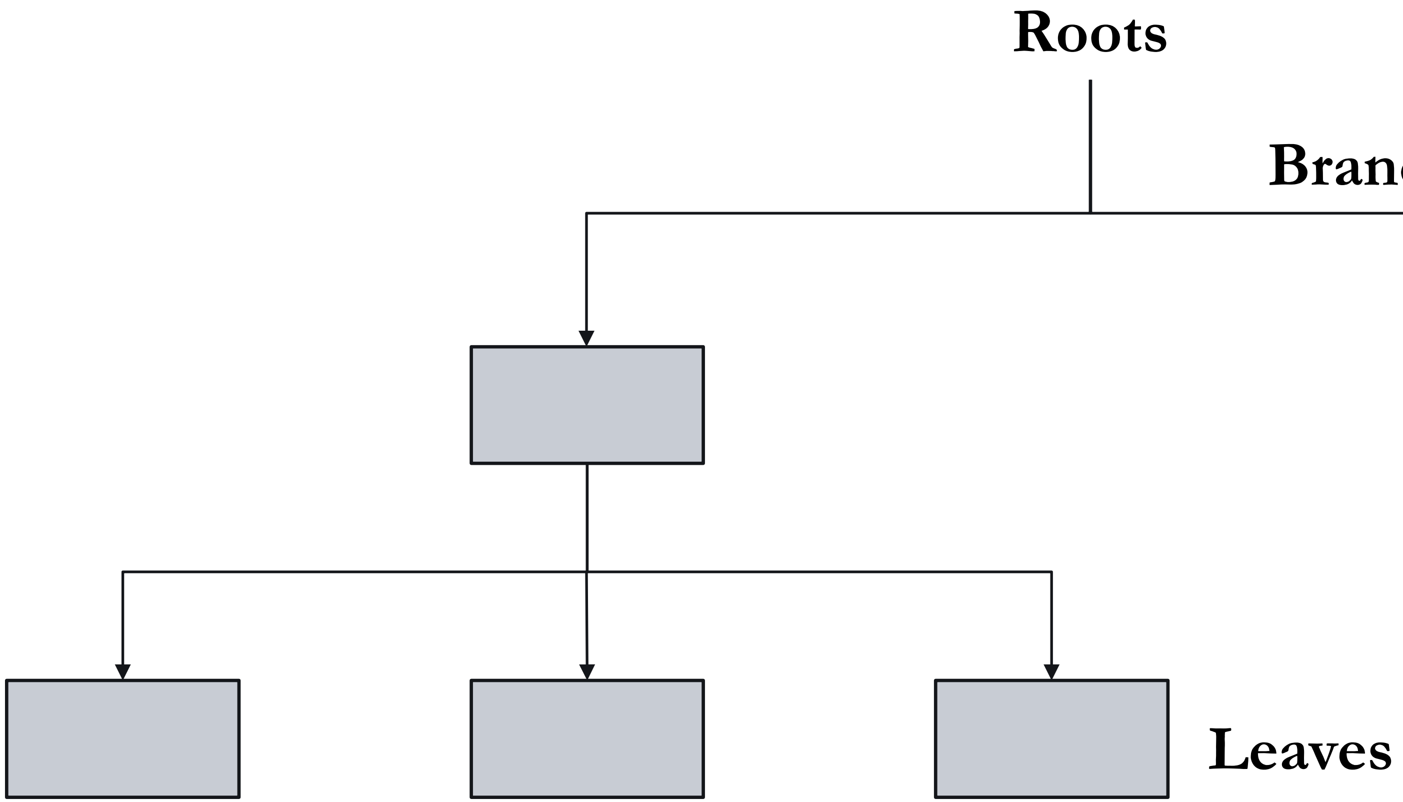
1.1 Two Types of Trees

	Regression Tree	Classification Tree
• Outcome of Interest:	Continuous variable Y	Binary variable Y
	Given the characteristic of observation i, what would we predict the y to be?	Given the characteristic of observation i, what would we predict the y (a categorical variable) to be?

We will focus first on regression tree.

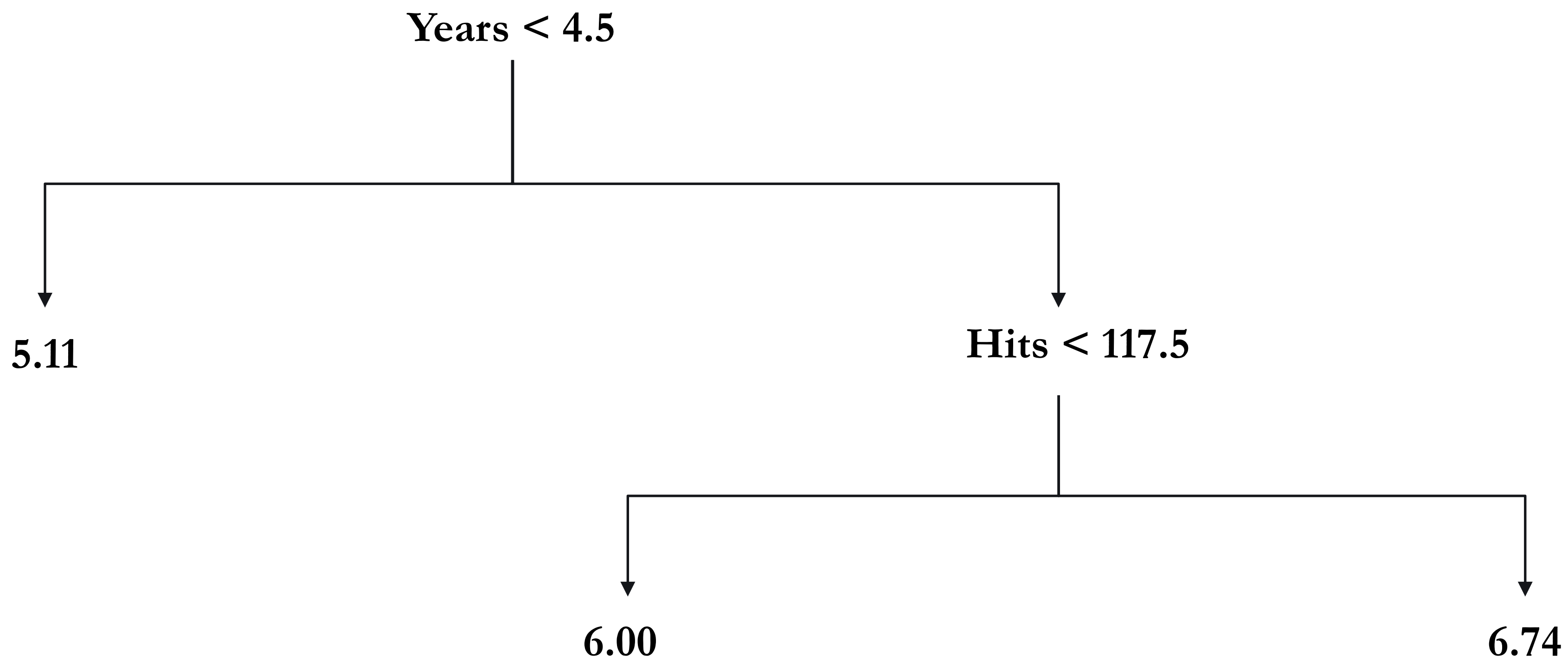
1.2 What is a Tree?

- **The tree is an upside-down tree with roots on top and leaves on the bottom.**



1.3 Regression Tree Example

Let's use a tree to predict the salary of a baseball player.



1.4 Algorithm of Tree

1) Divide predictor space into J distinct and non-overlapping regions: $\boldsymbol{R_1}, \boldsymbol{R_2}, \boldsymbol{R_3}, \boldsymbol{R_4}$, etc.

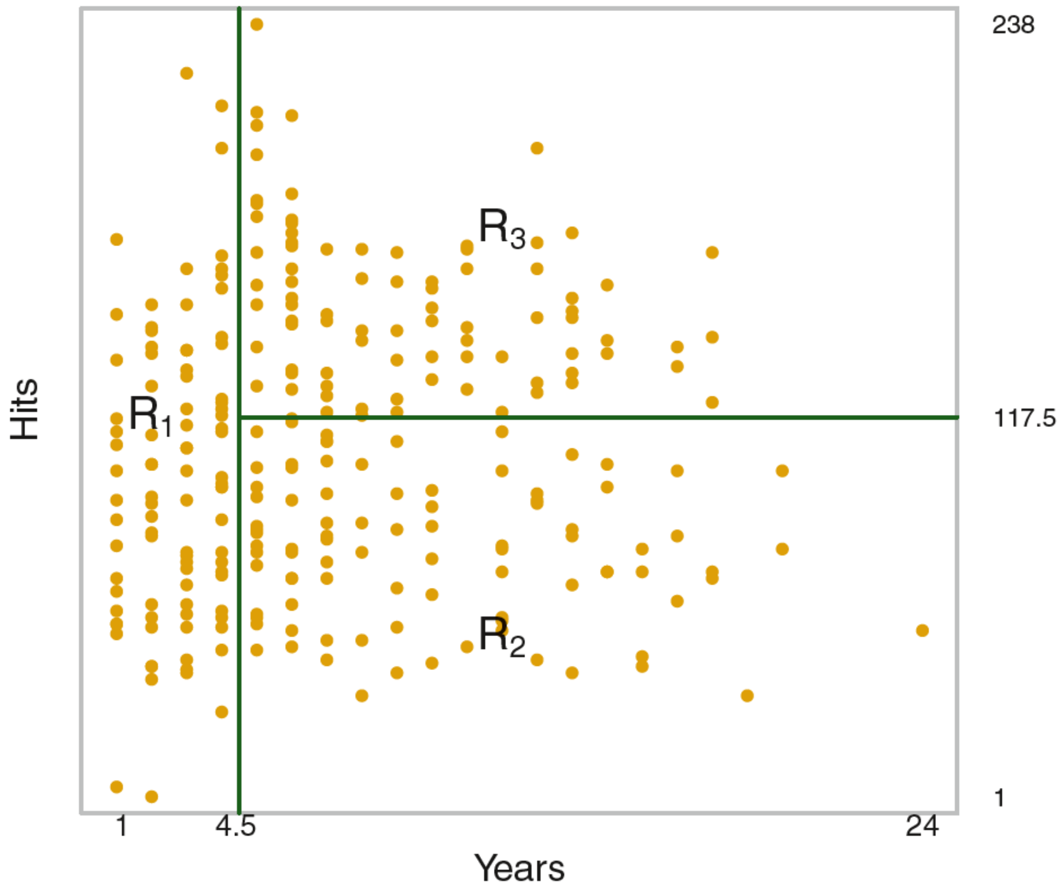
2) For every observation that falls into the region $\boldsymbol{R_j}$, it will have the same prediction, which is the mean of y for the training observations located in region $\boldsymbol{R_j}$.

3) Need to find boxes or regions that minimize residual sum of squares (RSS) given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

This is a measure of performance!

1.5 Algorithm in Pictures



From James et al (2017)

1.6 How are regions created?

Regions are created through *binary splitting.*

1) The algorithm goes through all the different covariates available.

2) For each covariate, check different cutoff s and calculate RSS.

3) Choose the covariate and S that minimizes RSS below:

$$\sum_{i:y_i \in R_1(s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:y_i \in R_2(s)} (y_i - \hat{y}_{R_2})^2$$

Then repeat process again until stopping criteria reached.

- **Notice that the split is done at a particular step and does not look ahead.**

1.7 Overfitting

- **With non-parametric methods, including decision trees, *overfitting* might be an issue. Overfitting means you fit the data too well.**
  - **Since it fits the training data extremely well, very likely it will not fit as well when we used the test data.**
  - **Error for the training data is very low, but when overfitting happens, test error are not the lowest and can be much larger.**

What can we then do?

- **Make sure tree not too long? Keep only branches that reduces RSS by a lot?**
- **Better Option: Tree Pruning**

1.8 Brief Overview on Model Selection

Those interested further should read section 6.1.3 James et al (2017).

- Usually, the goal of these exercises is to find a model that can best predict the test data.
  - aka smallest test error.
  - If you have such models, then you can make a lot of predictions which can influence business decisions or policy.

Thus, running one non-parametric model is not enough, usually the goal is to keep tweaking until you find the best model!

1.9 Pruning a Tree

- **Let the tree grow long, then want to find subtree with the lowest test error rate.**

One possible method: *cost complexity pruning*

Add a penalty for complexity. |T| specifies number of nodes.

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

α can be chosen using cross-validation method.

Note: Cross validation is a way to resample the test sample to find different α. Once a good α is chosen, use this on the training sample.

Coding Exercise

See "Lesson 14 Regression Tree Example.R"

Reference

- Basuchoudhary, Bang, Sen (2017) Chapter 3 Machine-learning Techniques in Economics: New Tools for Predicting Economic Growth, SpringerBriefs in Economics
- James, Witten, Hastie, and Tibshirani (2017) Chapter 8 An Introduction to Statistical Learning with Applications in R, Springer Texts in Statistics