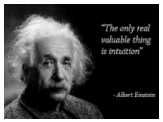
Regression Trees

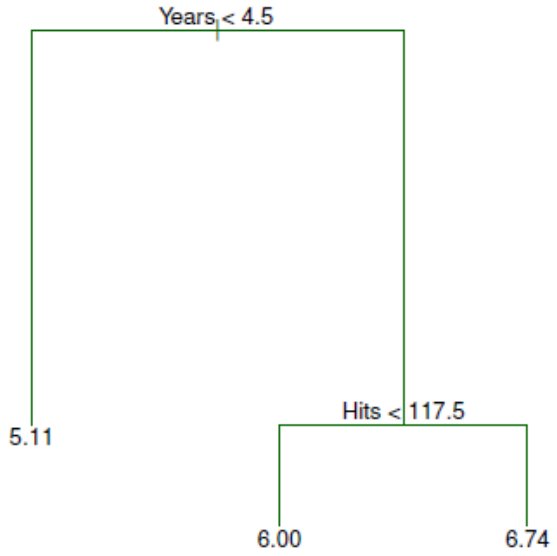
Analytics Modeling Options

	Modeling Method		
	Structured		Visual, Text, Unstructured, etc.
Descriptive	Cluster analysis, correlation, market basket analysis, sample statistics, ANOVA		Bubble charts, network diagrams, natural language processing, clustering dendograms, etc.
Predictive	Association	Decision Tree	Charts
Quantitative Value	Regression	Regression Trees	Regression plots, scatter plots, Tableau diagrams, trend charts, etc.
Classification	Logistic Regression; Other Categorical Regression Models	Classification Trees	Tree maps, interactive diagrams,
Prescriptive	Operations research, decision modeling, optimization, linear programming		Simulations, etc.



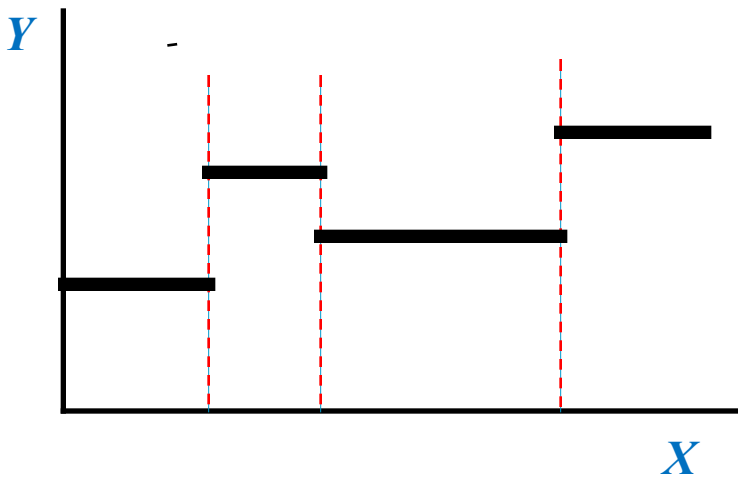
Regression Trees: Intuition

- Regression tree is **NOT** a classification method, but a **regression method** with a **quantitative** outcome. It is covered here for **consistency** with classification tree methods.
- The intuition is simple. If we have a **quantitative** outcome Y, we find the predictor X_1 that can **separate** the outcomes the **farthest** in the **training** data and **split** it at that **"node"**, creating **2** tree **branches**.
- We then find the predictor X_2 (or X_1 again) that can separate the outcomes the farthest within each branch; and **so on**
- We **predict** an outcome using the mean of the training observation in the region they belong
- In the tree **illustration** (see textbook) we are predicting baseball **player** Log(**salaries**) based on the player's number of years in the major leagues and number of **hits** that year.
- So a player with more than **4.5 years** and more than **117.5 hits** will make a Log(salary) of 6.74 or a **salary** of **$845.6K**



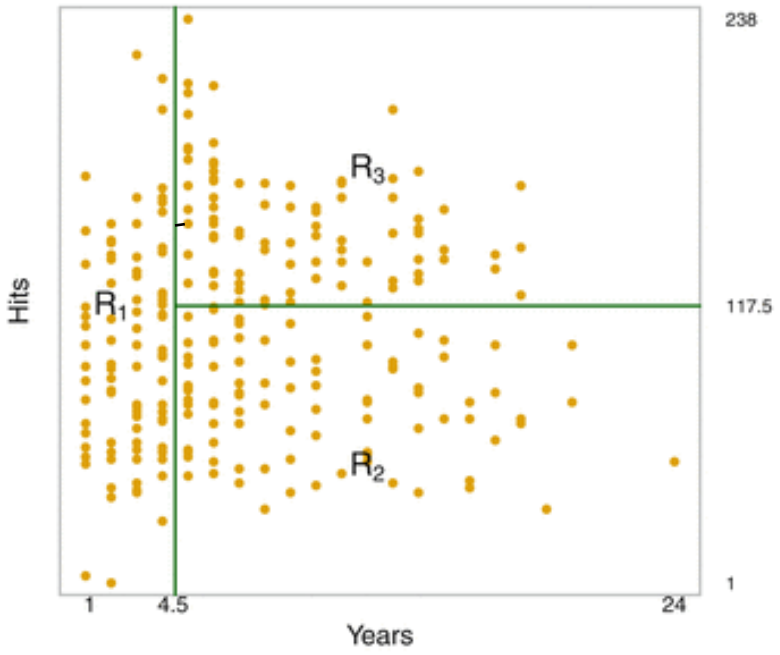
Regression Trees: Explained

- Essentially, we are **partitioning** the data into **"regions"** such that the distance **within** regions is **minimal** and the distance **between** regions is **largest**.
- **"Nodes"** the points where the branches **split**
- Any region can be **further subdivided** into more regions
- All observations within a **region** are assigned the same prediction equal to the **mean** of all training outcomes in the region
- The red **dotted** lines in the graph show the **nodes** that separate regions, where branches **split** in a tree with a single predictor X
- All predictions **within** a region use the the **mean** value for that region
- In essence regression trees are a **hybrid** between **step regressions** and **K Nearest Neighbors**



Tree Regions

- The diagram shows two partitions in the baseball salary example, with the first partition at $Years < 4.5$ and the region for Years >= 4.5 further sub-partitioned at $Hits < 117.5$
- This **partitioning** is also called **"Stratification"** or **"Segmenting"**
- With P predictors, the regression tree method finds the specific predictor x_i and **cutoff node** within that predictor which minimizes the training **ESS**
- It then find which of the **resulting partitions** to split, one at a time, and **where**, such that **ESS** is further reduced and minimized again
- We then **repeat** the process
- We can continue until **each branch** has exactly **one data point**, but **where** do we **stop**?



KOGOD SCHOOL
of
BUSINESS