# MATH 4322 - Lecture 15

## Tree Based Methods

Dr. Cathy Poliak, cpoliak@uh.edu

University of Houston

# Tree Based Models

So far we have covered such parametric models as:

- Linear Regression
- Logistic Regression
- Linear Discriminant Analysis

and such resampling techniques as

- Cross-Validation
- Bootstrap

we are ready for another model class:

- **Tree-based models**. CART

Tree-based models can be applied to **both** regression and classification problems.

# Motivating example

**Example**. For *Hitters* data on baseball hitters, we'd like to predict baseball player's *Salary* based on

- *Years* - # years he played in major leagues, and
- *Hits* - # hits made in the previous year.

```
library(ISLR)
head(Hitters[,c("Salary","Years","Hits")])
```

```
                  Salary Years Hits
-Andy Allanson        NA     1   66
-Alan Ashby        475.0    14   81
-Alvin Davis       480.0     3  130
-Andre Dawson      500.0    11  141
-Andres Galarraga   91.5     2   87
-Alfredo Griffin   750.0    11  169
```

1. From our models that we have previous talked about, which would be best to use in this example?

a) Linear Regression

Regression

b) Logistic Regression

Classification

# Clean the Data First

- Notice we have missing values

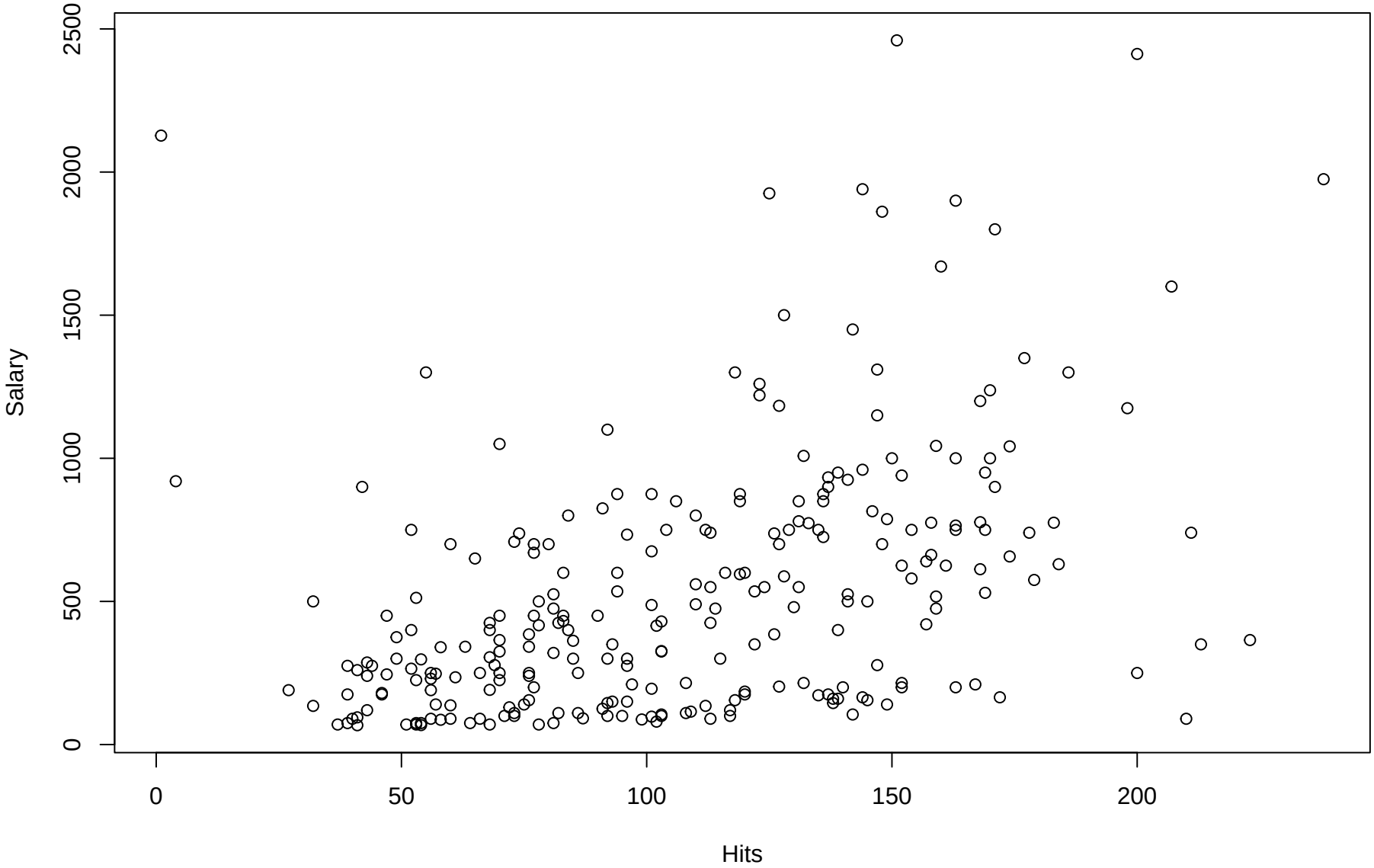
```
library(ISLR)
summary(Hitters$Salary)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   67.5   190.0   425.0   535.9   750.0  2460.0      59
```

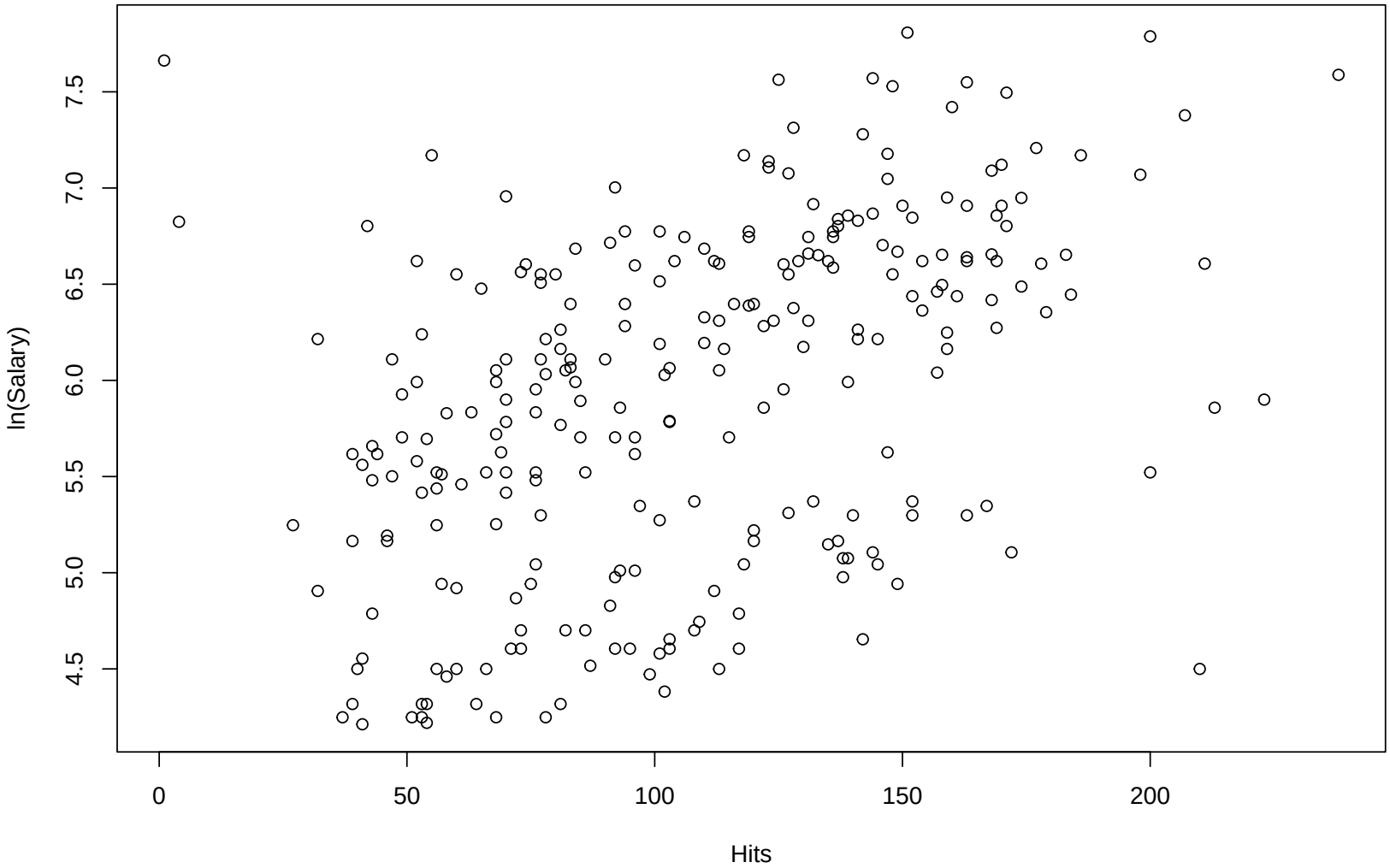
- We will clean this up and create a new data frame.

```
Hitters2 = na.omit(Hitters)
```

- In order to use the linear regression we have to assume
  - Linear relationship
  - Independence between observations
  - Normal distribution
  - Equal variance


Salary
0
500
1000
1500
2000
2500
0
50
100
150
200
Hits

# Log of Salary



# Testing and Training Data

We will split the data in half for training and testing.

```
set.seed(100)
sample = sample(1:nrow(Hitters2),nrow(Hitters2)/2)
train.hitters = Hitters2[sample,]
test.hitters = Hitters2[-sample,]
```

# Linear Regression

We will use all of the variables initially as the inputs.

```
lm.hitters = lm(log(Salary) ~ Hits + HmRun + Runs + RBI + Walks + Years + PutOuts +
                data = train.hitters)
hitters.coef = round(coef(summary(lm.hitters)),4)
print(knitr::kable(hitters.coef, escape = FALSE))
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.1819 | 0.1746 | 23.9550 | 0.0000 |
| Hits | 0.0093 | 0.0039 | 2.3734 | 0.0192 |
| HmRun | 0.0054 | 0.0139 | 0.3874 | 0.6992 |
| Runs | -0.0020 | 0.0061 | -0.3197 | 0.7497 |
| RBI | -0.0049 | 0.0062 | -0.7845 | 0.4343 |
| Walks | 0.0065 | 0.0036 | 1.8158 | 0.0719 |
| Years | 0.1059 | 0.0121 | 8.7538 | 0.0000 |
| PutOuts | 0.0001 | 0.0002 | 0.4057 | 0.6857 |
| Assists | 0.0001 | 0.0006 | 0.2451 | 0.8068 |
| Errors | -0.0065 | 0.0113 | -0.5690 | 0.5704 |

Which variables appear to be significant for predicting salary?

# Keeping on the Significant Variables

After using the step() function only Hits, Walks, and Years are significant.

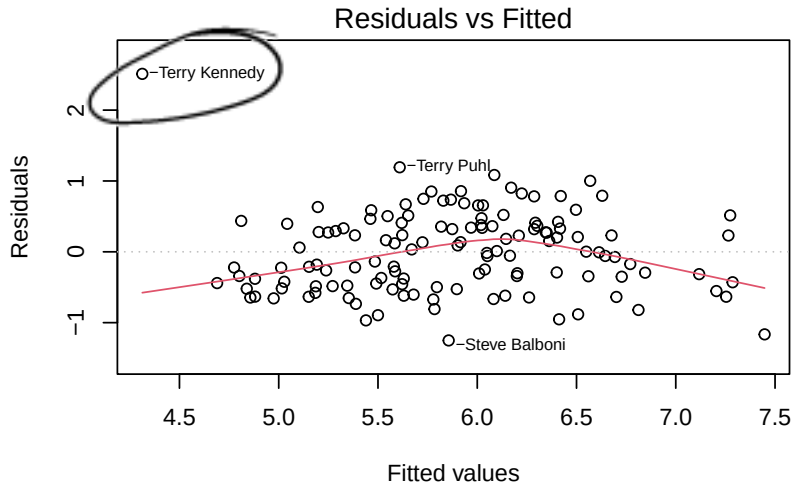
```
lm.hitters = lm(log(Salary) ~ Hits + Walks + Years, data = train.hitters)
summary(lm.hitters)
```

```
Call:
lm(formula = log(Salary) ~ Hits + Walks + Years, data = train.hitters)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25086 -0.45243 -0.00694  0.37243  2.51158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.173483   0.158230  26.376  < 2e-16 ***
Hits        0.006773   0.001375   4.926 2.56e-06 ***
Walks       0.005860   0.002797   2.095   0.0381 *
Years       0.106356   0.011169   9.522  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5801 on 127 degrees of freedom
Multiple R-squared:  0.5497,    Adjusted R-squared:  0.5391
F-statistic: 51.69 on 3 and 127 DF,  p-value: < 2.2e-16
```

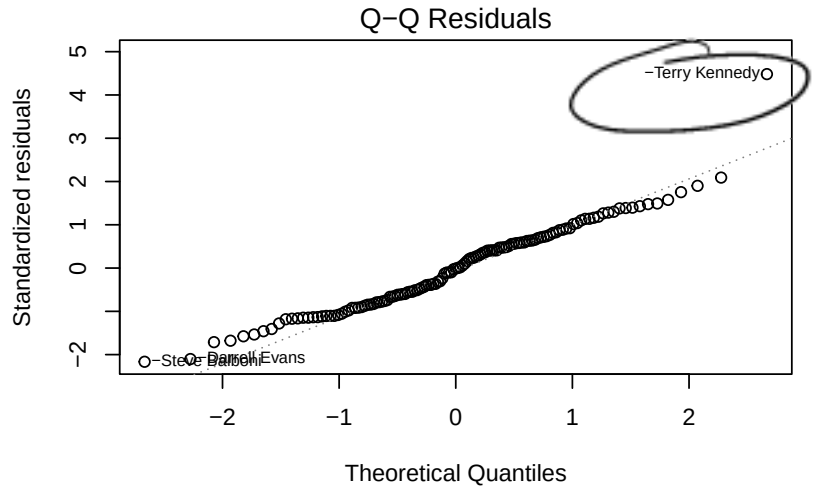
$\Rightarrow \widehat{\log(\text{salary})} = 4.1735 + 0.0068 * \text{hits} + 0.0059 * \text{Walks} + 0.1064 * \text{Years}$

# Are the Assumptions Met?

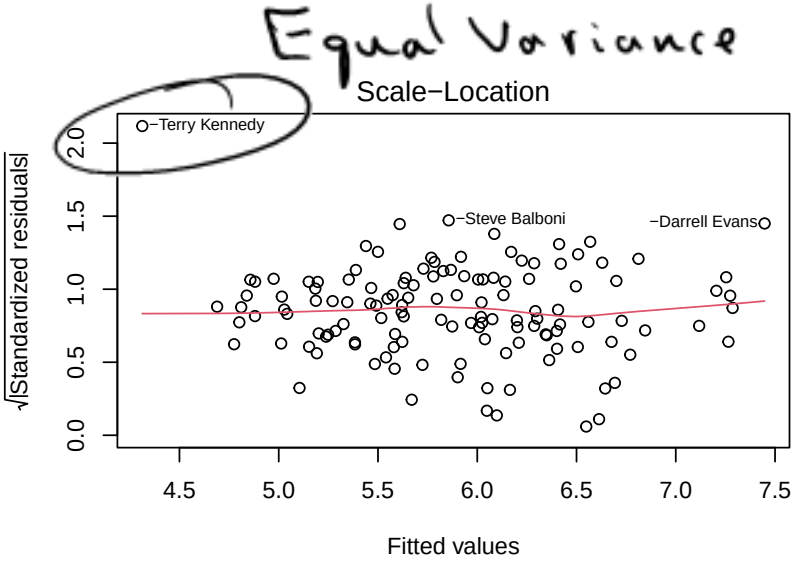
Linear



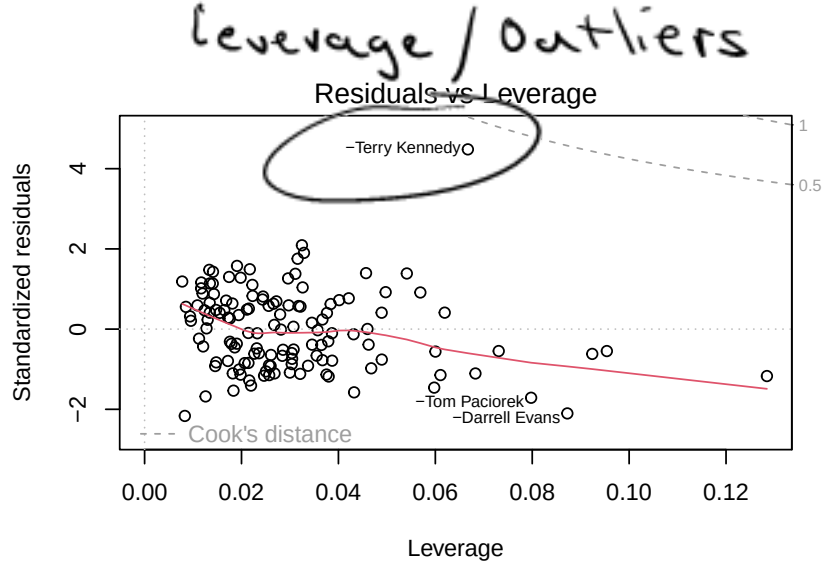

Normal




Equal Variance




Leverage / Outliers



# How Well are We Predicting?

- Remember that we desire to have a small **MSE** for both training and testing set.
- MSE from the training set

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

```
pred.train = predict(lm.hitters)
(mse.hitter.train = mean((log(train.hitters$Salary) - pred.train)^2))
```

```
[1] 0.3262283
```

```
sqrt(exp(mse.hitter.train))
```

```
[1] 1.177171
```

← training MSE

- MSE from the test set

```
pred.test = predict(lm.hitters,newdata = test.hitters)
(mse.hitter.test = mean((log(test.hitters$Salary) - pred.test)^2))
```

```
[1] 0.4696672
```

```
sqrt(exp(mse.hitter.test))
```

```
[1] 1.264698
```

- This states that the estimates of the salary may be off by about $1177.17 based on the training set and about $1264.7 based on the test set.
- Can we do better?

# 10-Fold CV

```
library(boot)
glm.hitters = glm(log(Salary) ~ Hits + Walks + Years, data = Hitters2)
set.seed(2)
(cv.hitters = cv.glm(Hitters2,glmfit = glm.hitters, K =10)$delta[1])
```

```
[1] 0.4114788
```

```
sqrt(exp(cv.hitters))
```

```
[1] 1.228433
```

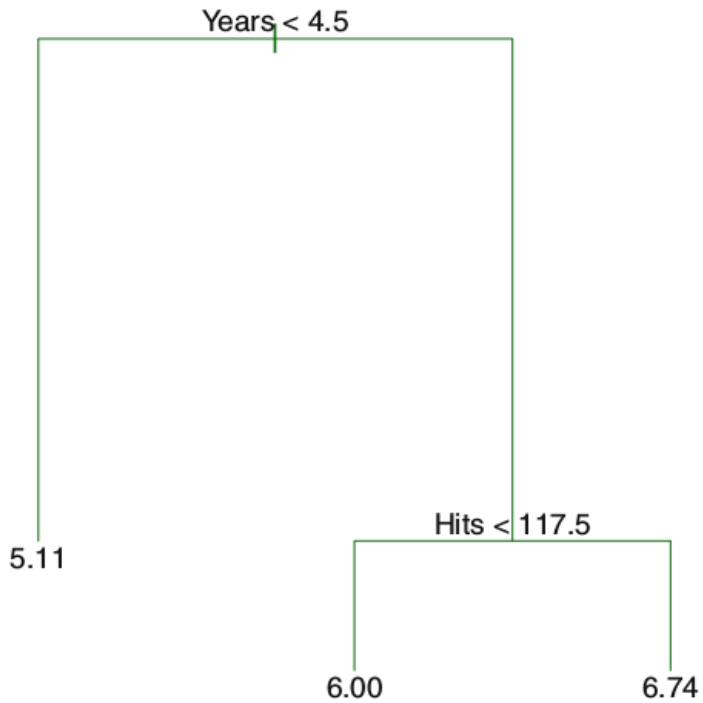
When we split the data 10 times, we get an average MSE of \$1228.43.

# Use A Decision Tree Instead

After certain data pre-processing steps, such as:

- dropping observations with missing values
- applying log-transformation to response variable *Salary*

an example of a fitted decision tree looks as follows



# Regression Trees: *Hitters* example

Explanation of the tree.

– Predictor space (range of values for *Years* and *Hits* variables) got segmented into **3** regions (terminal nodes):

- $R_1 = (Years < 4.5)$
- $R_2 = (Years \geq 4.5) \ \ \& \ \ (Hits < 117.5)$,
- $R_3 = (Years \geq 4.5) \ \ \& \ \ (Hits \geq 117.5)$

# Regression Trees: *Hitters* example

Explanation of the tree.

– Predictor space (range of values for *Years* and *Hits* variables) got segmented into **3** regions (terminal nodes):

- $R_1 = (Years < 4.5)$
- $R_2 = (Years \geq 4.5)$ & $(Hits < 117.5)$,
- $R_3 = (Years \geq 4.5)$ & $(Hits \geq 117.5)$

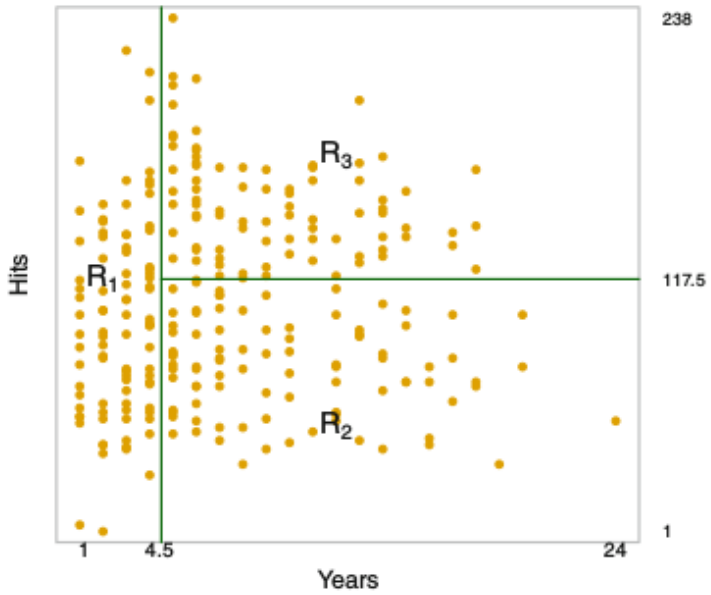
– *Salary* prediction in region $R_i$ = (**mean** *Salary* of **all** hitters $\in \boldsymbol{R_i}$):

- Any hitter in $R_1$ ($< 4.5$ years of experience) is predicted to make $e^{5.11} \approx 165k$.
- For hitters with over 4.5 years of experience:
  - if hitter $\in R_2$ ($< 117.5$ hits), he is projected to make $e^{6.00} \approx 403k$,
  - while for $R_3$ ($\geq 117.5$ hits) - $e^{6.74} \approx 845k$

# Regression Trees: General Idea

Decision trees for regression is also known as **regression trees**, are built via:

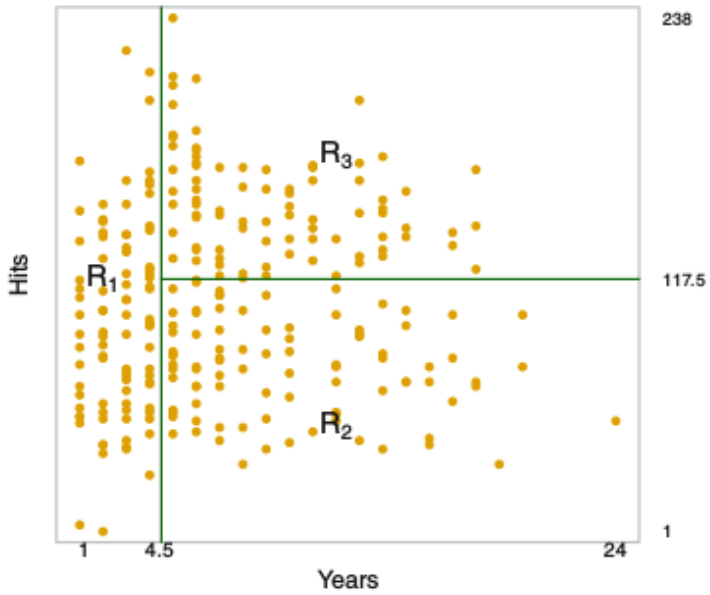
- Segmenting predictor space ($\leftrightarrow$ set of all possible values for $X_1, \ldots, X_p$) into simple non-overlapping regions $R_1, \ldots, R_J$.
- For every observation falling into region $R_j$, the predicted response is the **mean response** of *all* training observations in $R_j$.



# Back to *Hitters*: How do we segment?

How do we construct the regions $R_1, \ldots, R_J$ ?

1. Regions are built via splitting range of a variable. E.g. range of *Years* is split **at value** 4.5 into $(Years < 4.5)$ and $(Years \geq 4.5)$.




Results are boxes, or high-dimensional rectangles, for the sake of

- simplicity,
- ease of interpretation

# How do we segment? Recursive binary splitting.

2. We need a partition $R_1, \ldots, R_J$ that minimizes

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

It is computationally in-feasible to go through all possible partitions $R_1, \ldots, R_J$

**Recursive binary splitting** is an approach that's

- top-down (begins with a full region and then successively splits the predictor space)
- greedy (at each step the best split is made regardless of next steps)

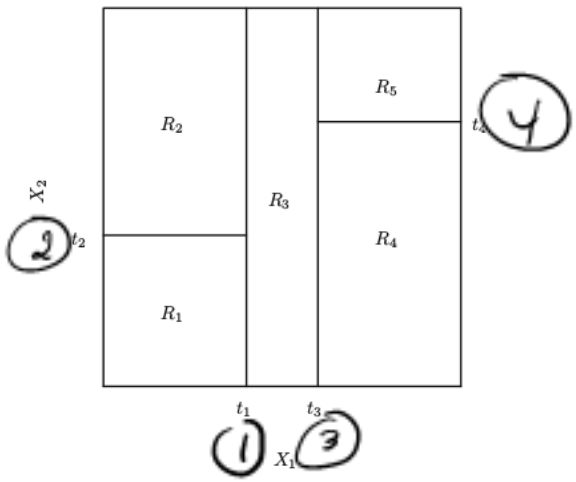
# Recursive Binary Splitting: Steps

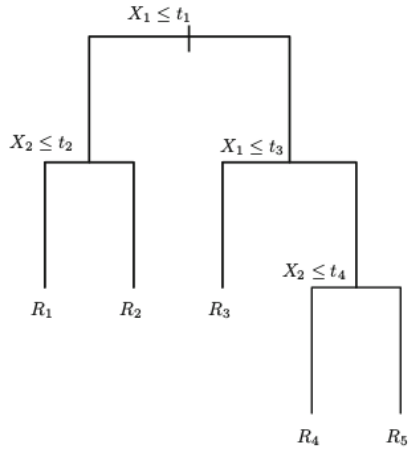
**Notation**: $R(j, s) = \{X \mid X_j < s\}$ is the region of predictor space $(X_1, \ldots, X_p)$ where $X_j < s$ ($j^{th}$ predictor is less than value $s$).
**Steps** of recursive binary splitting:

1. Start with full predictor space $R = \{(X_1, \ldots, X_p)\}$.
2. For any $j$ and $s$, we define the pair of half-planes
$$R_1(j, s) = \{X | X_j < s\}, \quad R_2(j, s) = \{X | X_j \geq s\},$$
3. We seek for values of $j$ and $s$ that minimize
$$RSS = \sum_{i:\ x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\ x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$
4. We repeat steps 2 and 3 trying to split the data further by minimizing the RSS, until a stopping criterion (e.g. "no region contains more than five observations") is reached .
5. We get a final set of regions $R_1, \ldots, R_J$, and later predict the **response** for a test observation from region $R_j$ via the **mean** of training observations $\in R_j$, $j = 1, \ldots, J$.

# Recursive Binary Splitting: Simulated Example

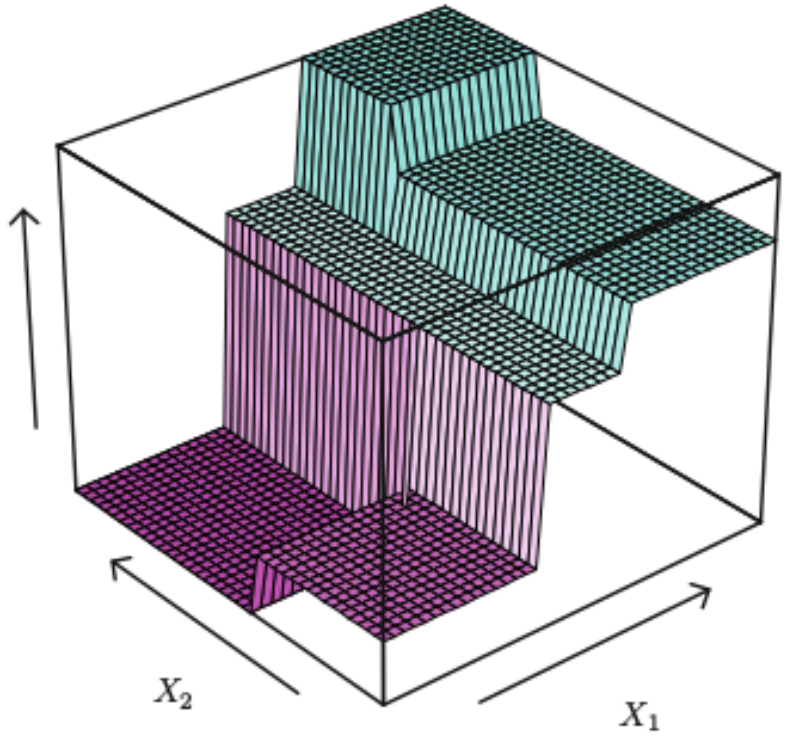
**Example.** Below is the output of recursive binary splitting applied to a simulated two-dimensional example with full predictor space $\{(X_1, X_2)\}$. With corresponding tree on right.



# Recursive Binary Splitting: Simulated Example

Prediction **surface** corresponding to that tree is depicted below.




$z$-axis contains predicted response value for the respective region $R_j$.

# Overfitting Issue: Tree Pruning

Recursive binary splitting usually results into large trees that

- produce good predictions on training data, but
- tends to overfit and perform poorly on test data.

A smaller tree with fewer splits ($\leftrightarrow$ fewer regions $R_1, \ldots, R_J$) might

- lead to lower variance and better testing set performance,
- better interpretation,
- at the cost of a little extra bias.

**Solution**:

1. Grow a large tree $T_0$, but then
2. **Prune** it back in order to obtain a subtree.

For step 2 our goal is to select a subtree that leads to the lowest **test** error rate (which could be estimated via cross-validation).

# Overfitting Issue: Cost Complexity Pruning

In **cost complexity** pruning, instead of all possible subtrees, we consider a sequence of trees indexed by a tuning parameter $\alpha \geq 0$.

For each value of $\alpha$ there's a subtree $T \subset T_0$ that minimizes

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

where $|T|$ - number of terminal nodes of tree $T$; $R_m$ - region for $m^{th}$ terminal node; $\hat{y}_{R_m}$ - predicted response for $R_m$ (mean of obs. $\in R_m$).

Tuning parameter $\alpha$ controls the trade-off between

- quality of fit the subtree $T$ provides to training data (left part), and
- the complexity ($\leftrightarrow$ # of terminal nodes) of subtree $T$ .

As $\alpha$ increases, there's a penalty for trees with too many terminal nodes $\Longrightarrow$ quantity (1) tends to be minimized for a smaller subtree.

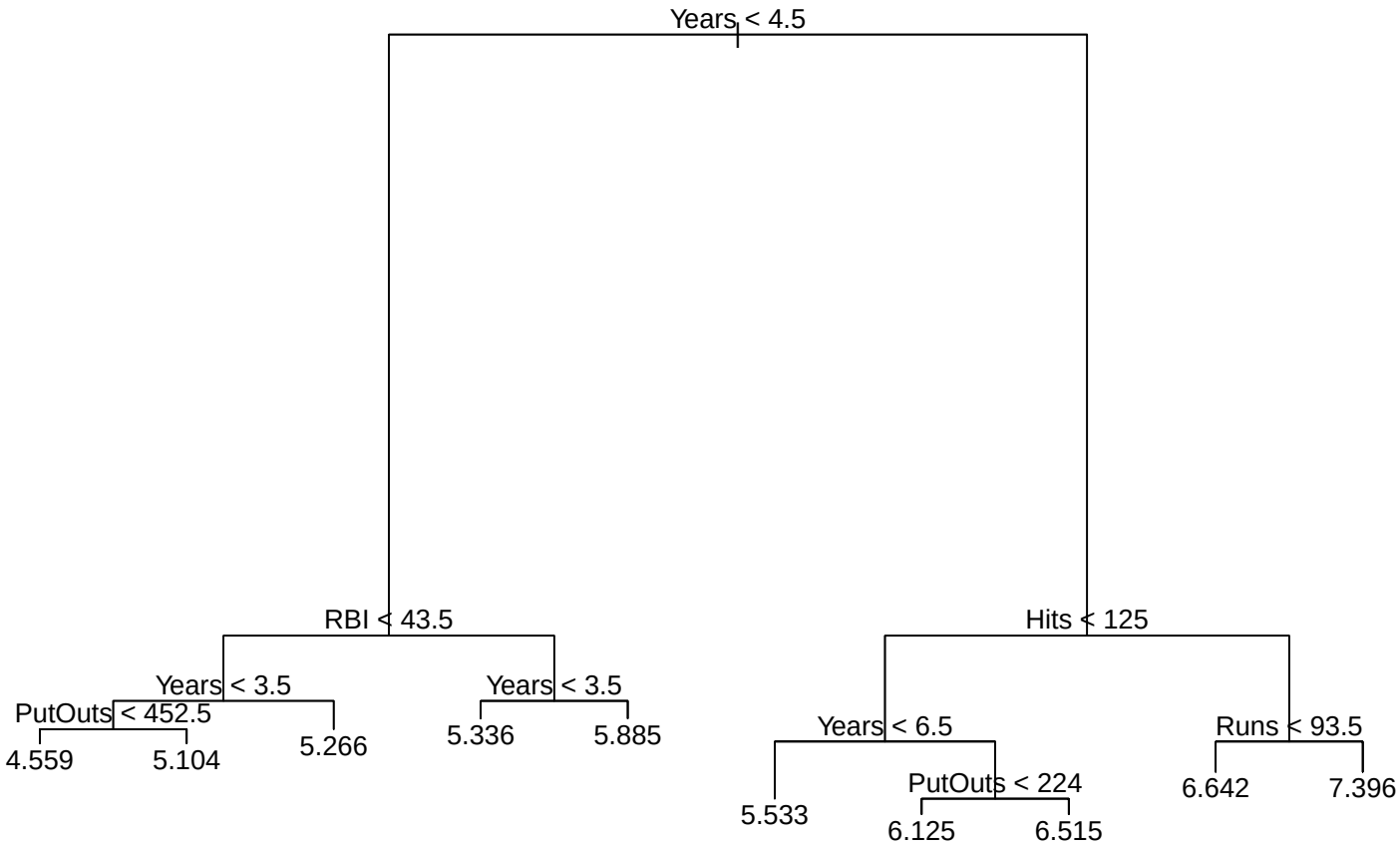
# Decision Tree: Full Algorithm

Below are the steps of building a decision tree (ISLR, page 309):

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of $\alpha$.
3. Use $K$-fold cross validation to choose $\alpha$. That is, divide the training observations into $K$ folds. For each $k = 1, \ldots, K$;
   (a) Repeat steps 1 and 2 on all but the $k$th fold of the training data.
   (b) Evaluate the mean squared prediction error on the data in the left-out $k$th fold, as a function of $\alpha$.
4. Average the results for each value of $\alpha$, and pick $\alpha$ to minimize the average error.
5. Return the subtree from Step 2 that corresponds to the chosen value of $\alpha$.

# *Hitters* example: Large tree

**Example**. We randomly divide *Hitters* data into 132 training and 131 testing observations, and build a large regression tree on 9 features:



# R Code for Decision Trees

This introduces a new package called tree. Type in R and run the following to answer the following questions.

```
library(tree)
tree.hitters = tree(log(Salary) ~ Hits + HmRun + Runs + RBI + Walks +
                      Years + PutOuts + Assists + Errors, data = train.hitters)
summary(tree.hitters)
```

2. How many "nodes" did this produce?

   a) 10 b) 9 c) 4 d) 3

3. Type and Run the following in R What is the predicted salary for a player that has between 3.5 and 4.5 *Years* and *RBI* greater than 43.5?

   a) \$5.885
   b) \$5,885
   c) \$359.60
   d) \$359,602.80 $= e^{5.881} * 1000$

```
plot(tree.hitters)
text(tree.hitters)
```

# Tree Summary

```
Regression tree:
tree(formula = log(Salary) ~ Hits + HmRun + Runs + RBI + Walks +
    Years + PutOuts + Assists + Errors, data = train.hitters)
Variables actually used in tree construction:
[1] "Years"   "RBI"     "PutOuts" "Hits"    "Runs"
Number of terminal nodes:  10
Residual mean deviance:  0.1691 = 20.46 / 121
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.01600 -0.24880 -0.01925  0.00000  0.21450  1.72000
```

"Residual mean deviance" is the "Total residual deviance" divided by the "Number of observations" - "Number of Terminal Nodes".

```
dim(train.hitters)
```

```
[1] 131  20
```

```
(tot.resid.dev = sum((log(train.hitters$Salary) - predict(tree.hitters))^2))
```

```
[1] 20.45564
```

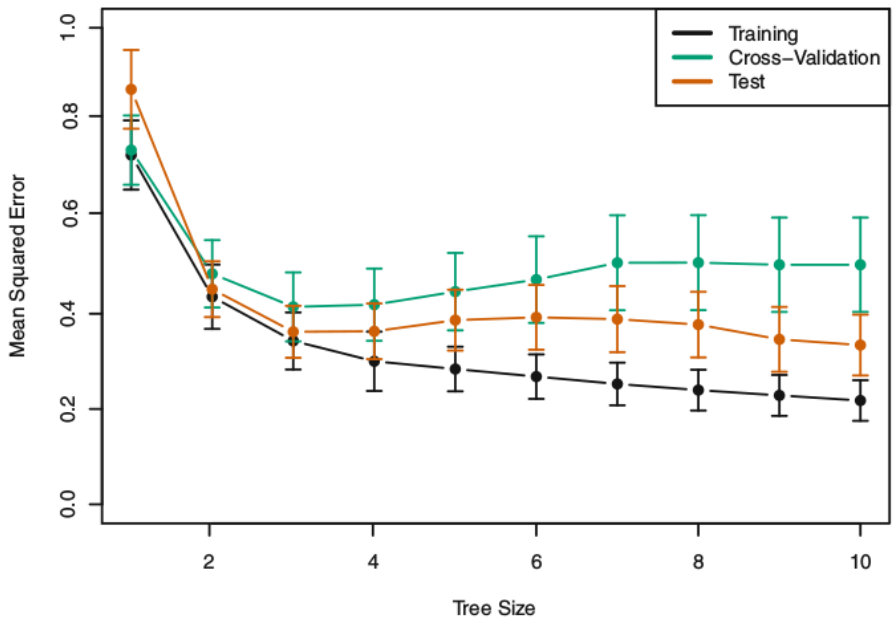
```
tot.resid.dev/(nrow(train.hitters) - 10)
```

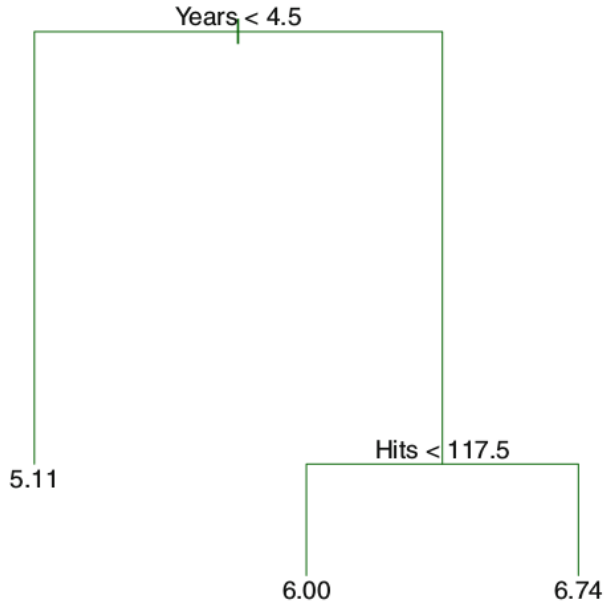
```
[1] 0.1690549
```

MSE

# *Hitters* example: Pruning

Then we perform pruning, and record **training error**, CV error & actual test set error for all resulting subtree sizes.




CV yields a subtree with $|T| = 3$ terminal nodes as the optimal one.

# Pruning the Tree

We think that 10 regions are too many. We can do the following to see how many regions we need.

```
cv.hitters = cv.tree(tree.hitters)
plot(cv.hitters$size,cv.hitters$dev,type = "b")
```

Where there is an apparent "elbow" determines how many regions we need.

4. How many regions (nodes) do we need?

   a) 11 b) 9 c) 5 d) 3

5. Type and Run the following in R. What is the predicted salary for a player that has played for less than 4.5 *Years*?

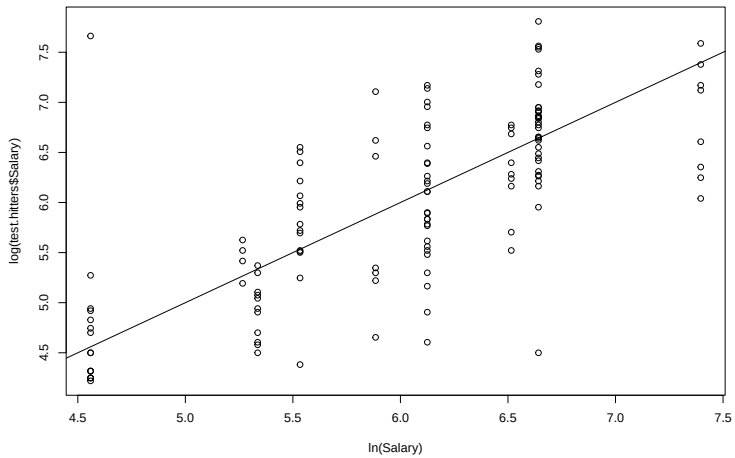
   a) \$5.135
   b) \$5,135
   c) \$169.86
   d) \$169,864.30 $x\ e^{5.135} * 1000$

```
prune.hitters = prune.tree(tree.hitters,best = 3)
plot(prune.hitters)
text(prune.hitters)
```

# How Well Are We Predicting?

- We can use the test set to determine the MSE of both trees.
- For the original tree with 10 regions we get.

```
yhat = predict(tree.hitters,newdata=test.hitters)
plot(yhat,log(test.hitters$Salary),xlab = "ln(Salary)")
abline(0,1)
```

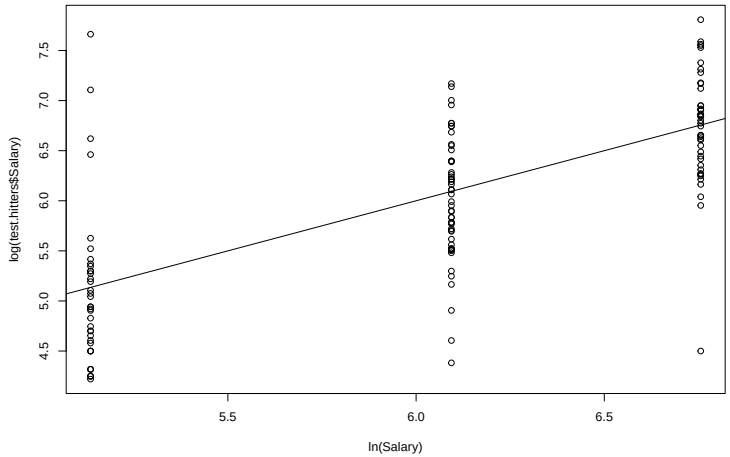



```
(mse.tree = mean((yhat-log(test.hitters$Salary))^2))
```

```
[1] 0.4180251
```

- For the pruned tree we get

```
yhat.prune = predict(prune.hitters,newdata = test.hitters)
plot(yhat.prune,log(test.hitters$Salary),xlab = "ln(Salary)")
abline(0,1)
```



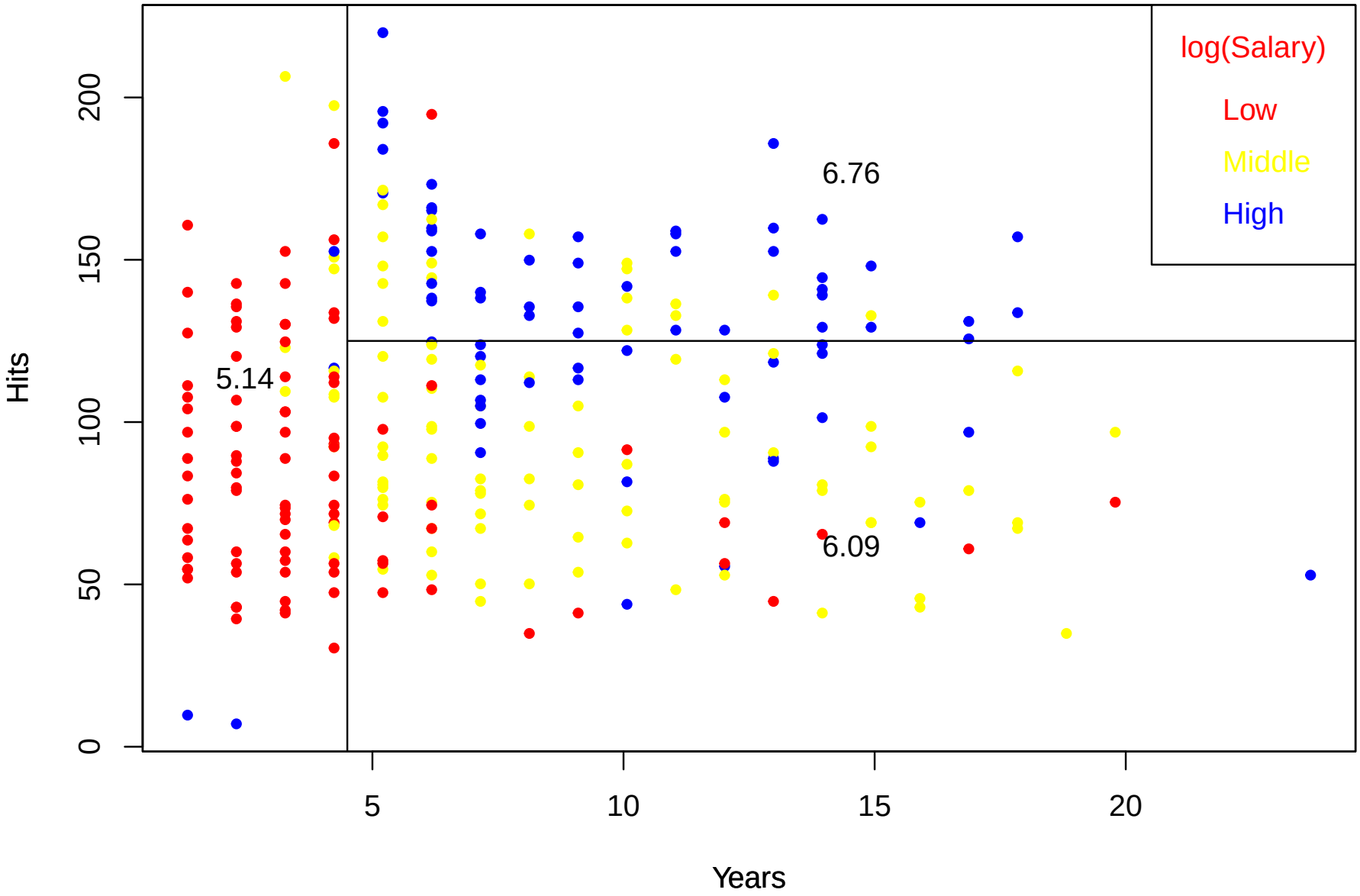

```
(mse.prune = mean((yhat.prune - log(test.hitters$Salary))^2))
```

```
[1] 0.4242652
```

- Taking the square root and exponential, we get
  - For the larger tree this model leads to test predictions that are within around $1232.46 of the true salary for a player.
  - For the pruned tree this model leads to test predictions that are within around $1236.31 of the true salary for a player.

# MSE Compared

| | Training Set | Test Set |
|---|---|---|
| Linear Regression | 1.386 | 1.599 |
| Regression Tree | 1.169 | 1.519 |
| Pruned Tree | 1.343 | 1.528 |


log(Salary)
Low
Middle
High
Hits
Years
0
50
100
150
200
5
10
15
20
5.14
6.76
6.09