Regression Trees

Nate Wells

Math 243: Stat Learning

November 2nd, 2020

Outline

In today's class, we will...

- Discuss regression trees as a non-parametric model

Section 1

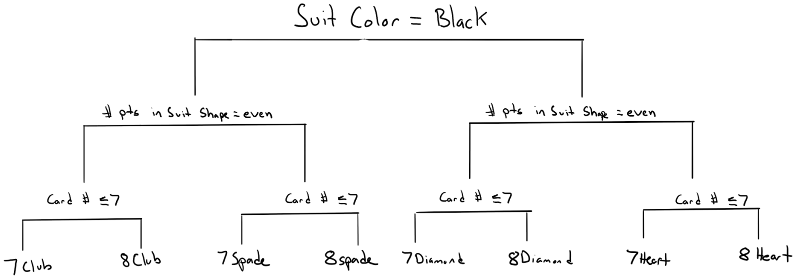
Regression Trees

Guess my card

I've chosen a card from a standard deck of 52 cards.

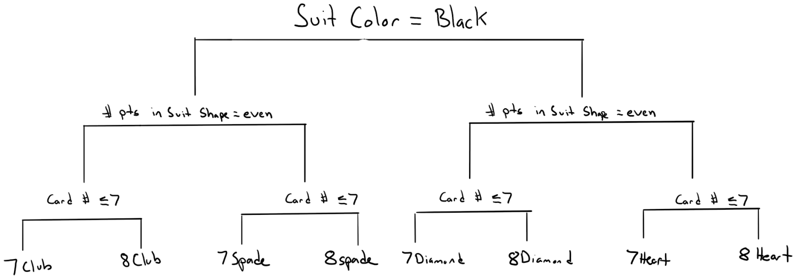
- You may ask me any 3 yes-no questions about the card, which I will answer truthfully.
- You you must try to guess the card I've chosen.
- You must decide your questions **before** you know any of the answers.

Decision tree for one strategy



Key: Proceed from top to bottom. If answer to question is yes, choose left branch.

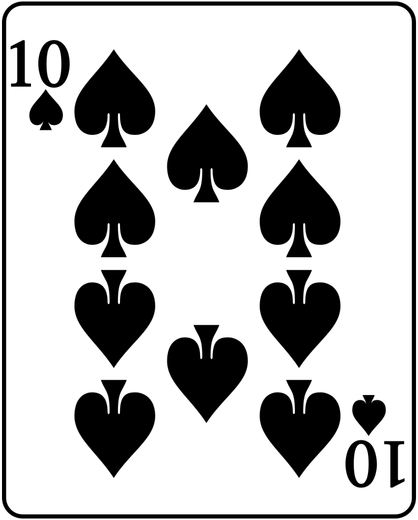
Decision tree for one strategy



Key: Proceed from top to bottom. If answer to question is yes, choose left branch.

Assuming card chosen is uniformly random, what is the success rate of this strategy?

My card



Regression Trees

Basic regression trees partition data into smaller groups that are homogenous with respect to predictors.

Regression Trees

Basic regression trees partition data into smaller groups that are homogenous with respect to predictors.

- They then make predictions based on average value of response in each group

Regression Trees

Basic regression trees partition data into smaller groups that are homogenous with respect to predictors.

- They then make predictions based on average value of response in each group

The most common technique is the Classification and Regression Tree (CART) method.

Regression Trees

Basic regression trees partition data into smaller groups that are homogenous with respect to predictors.

- They then make predictions based on average value of response in each group

The most common technique is the Classification and Regression Tree (CART) method.

1. The method begins with the entire data set S and searches every value of every predictor to cut S into two groups S_1 and S_2 that minimizes sum of squred error:

$$\mathrm{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2$$

Regression Trees

Basic regression trees partition data into smaller groups that are homogenous with respect to predictors.

- They then make predictions based on average value of response in each group

The most common technique is the Classification and Regression Tree (CART) method.

1. The method begins with the entire data set S and searches every value of every predictor to cut S into two groups S_1 and S_2 that minimizes sum of squred error:

$$\text{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2$$

2. The method then repeats step 1 for each of the two groups S_1 and S_2.

Regression Trees

Basic regression trees partition data into smaller groups that are homogenous with respect to predictors.

- They then make predictions based on average value of response in each group

The most common technique is the Classification and Regression Tree (CART) method.

1. The method begins with the entire data set S and searches every value of every predictor to cut S into two groups S_1 and S_2 that minimizes sum of squred error:

$$\mathrm{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2$$

2. The method then repeats step 1 for each of the two groups S_1 and S_2.
3. The method continues splitting groups until each subdivision has few observation (or another predetermined stopping condition is met)

Trees on Trees

We use a subset of the `pdxTrees` dataset from the `pdxTrees` repo (maintined by K. McConville, I. Caldwell, and N. Horton)

Trees on Trees

We use a subset of the `pdxTrees` dataset from the `pdxTrees` repo (maintined by K. McConville, I. Caldwell, and N. Horton)

- The data was collected by the Portland Parks and Recreation's Urban Forestry Tree Inventory Project.
- The Tree Inventory Project has gathered data on Portland trees since 2010, collecting this data in the summer months with a team of over 1,300 volunteers and city employees.

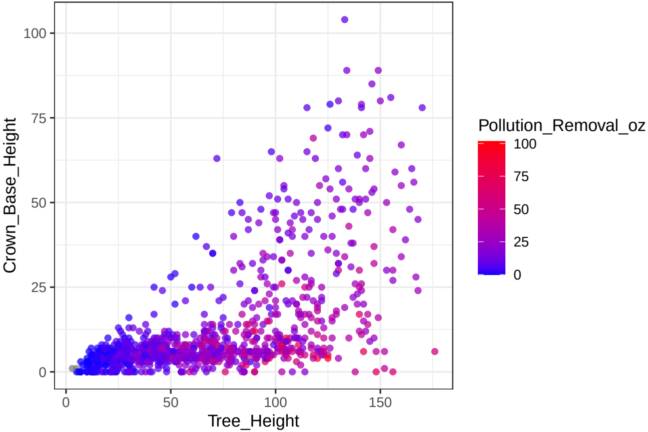
Trees on Trees

We use a subset of the pdxTrees dataset from the pdxTrees repo (maintined by K. McConville, I. Caldwell, and N. Horton)

- The data was collected by the Portland Parks and Recreation's Urban Forestry Tree Inventory Project.
- The Tree Inventory Project has gathered data on Portland trees since 2010, collecting this data in the summer months with a team of over 1,300 volunteers and city employees.

```
## Rows: 1,000
## Columns: 10
## $ Species              <fct> PSME, CAJA, QUMU, CADE, PSME, CPSP, PRAV, PSME...
## $ Condition            <fct> Fair, Fair, Fair, Fair, Fair, Fair, Poor, Fair...
## $ Tree_Height          <int> 102, 23, 18, 78, 123, 85, 11, 145, 16, 72, 88,...
## $ Crown_Width_NS       <int> 52, 36, 6, 17, 52, 36, 9, 36, 10, 86, 25, 12, ...
## $ Crown_Width_EW       <int> 43, 40, 6, 18, 38, 52, 11, 35, 10, 86, 10, 16,...
## $ Crown_Base_Height    <int> 63, 5, 5, 6, 13, 5, 6, 9, 5, 8, 6, 4, 4, 3, 2,...
## $ Structural_Value     <dbl> 6694.04, 2444.75, 71.28, 4162.43, 6159.02, 113...
## $ Carbon_Storage_lb    <dbl> 1992.9, 917.5, 5.3, 1428.7, 1901.4, 11071.6, 2...
## $ Stormwater_ft        <dbl> 78.9, 43.9, 1.0, 19.8, 117.6, 52.0, 4.1, 80.1,...
## $ Pollution_Removal_oz <dbl> 21.2, 11.8, 0.3, 5.3, 31.6, 14.0, 1.1, 21.5, 1...
```

Pollution Removal



An Old Friend

This seems like a good time to implement linear regression:

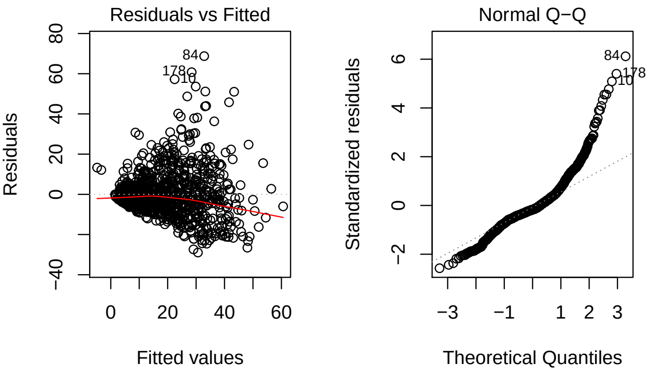
An Old Friend

This seems like a good time to implement linear regression:

```
tree_lm<-lm(Pollution_Removal_oz ~., data=small_pdxTrees)
summary(tree_lm)
```

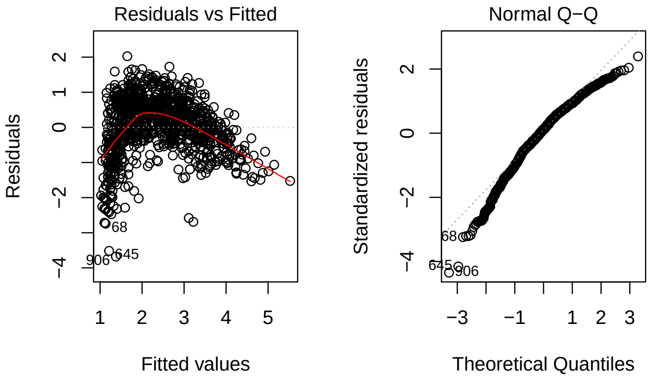
```
##
## Call:
## lm(formula = Pollution_Removal_oz ~ ., data = small_pdxTrees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.946  -5.591  -1.920   3.940  68.752
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.68725    0.69781   0.985    0.325
## Tree_Height        0.35734    0.01117  31.997   <2e-16 ***
## Crown_Base_Height -0.49555    0.02831 -17.506   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.25 on 984 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.5116, Adjusted R-squared:  0.5106
## F-statistic: 515.4 on 2 and 984 DF,  p-value: < 2.2e-16
```

Diagnostic Plots



Quick Fix

```
log_tree_lm<-lm(log(Pollution_Removal_oz) ~., data=small_pdxTrees)
par(mfrow = c(1, 2))
plot(log_tree_lm, 1:2)
```



Conclusion

```
##                       Estimate Std. Error     t value      Pr(>|t|)
## (Intercept)          0.6872486 0.69780598   0.9848706  3.249297e-01
## Tree_Height          0.3573381 0.01116780  31.9971761 1.459606e-154
## Crown_Base_Height -0.4955457 0.02830717 -17.5060118  6.071709e-60
```

- Increasing `Tree_Height` while holding `Crown_Base_Height` constant corresponds to an increase in `Pollution_Removal_oz` of about 0.36.
- Increasing `Crown_Base_Height` while holding `Tree_Height` constant corresponds to a decrease in `Pollution_Removal_oz` of about 0.5.

Conclusion

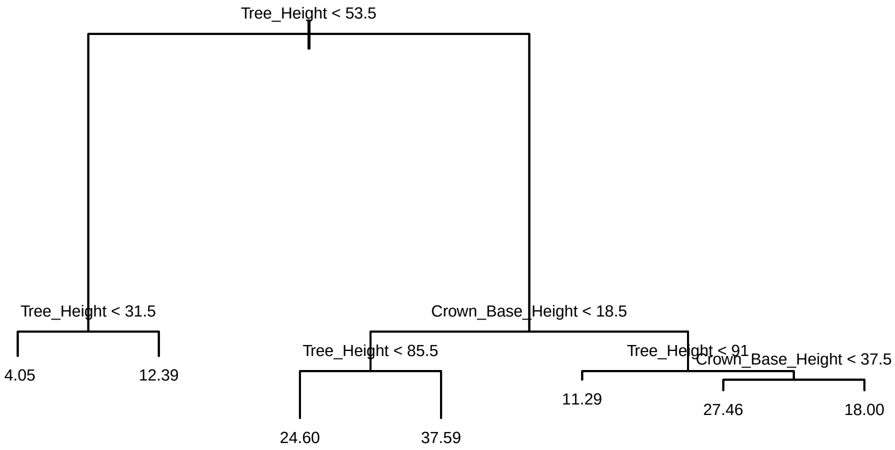
```
##                     Estimate Std. Error     t value      Pr(>|t|)
## (Intercept)        0.6872486 0.69780598   0.9848706  3.249297e-01
## Tree_Height        0.3573381 0.01116780  31.9971761 1.459606e-154
## Crown_Base_Height -0.4955457 0.02830717 -17.5060118  6.071709e-60
```

- Increasing Tree_Height while holding Crown_Base_Height constant corresponds to an increase in Pollution_Removal_oz of about 0.36.
- Increasing Crown_Base_Height while holding Tree_Height constant corresponds to a decrease in Pollution_Removal_oz of about 0.5.

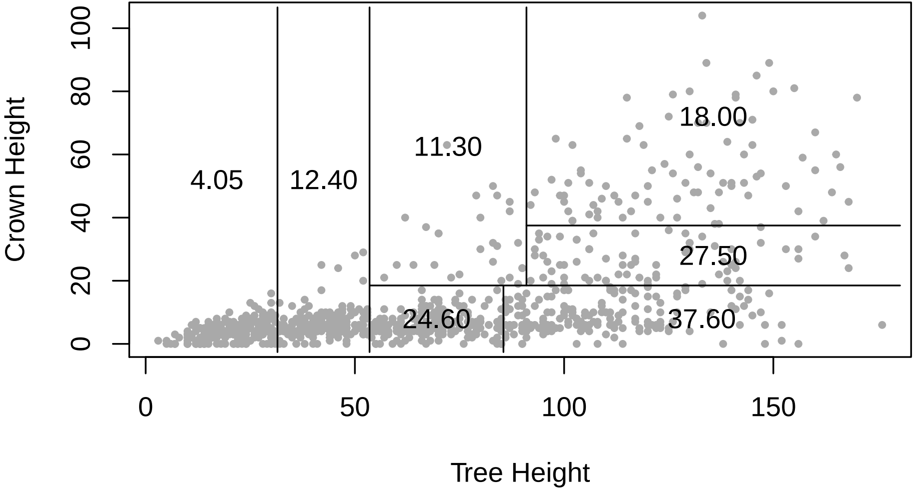
```
##                      Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)        0.91609650 0.0525205492  17.44263  1.417732e-59
## Tree_Height        0.02711289 0.0008405474  32.25623 2.503021e-156
## Crown_Base_Height -0.02720067 0.0021305466 -12.76699  1.189929e-34
```

- Increasing Tree_Height while holding Crown_Base_Height constant corresponds to a proportional increase of Pollution_Removal_oz by about 2.7%.
- Increasing Crown_Base_Height while holding Tree_Height constant corresponds to a decrease in Pollution_Removal_oz by about 2.7%.

Regression Tree



Another Visualization



Interpretation

- `Tree_Height` is the most important factor contributing to `Pollution_Removal`, with larger trees removing more pollution.

Interpretation

- Tree_Height is the most important factor contributing to Pollution_Removal, with larger trees removing more pollution.
- Given a small tree, Crown_Base_Height has little impact on Pollution_Removal

Interpretation

- `Tree_Height` is the most important factor contributing to `Pollution_Removal`, with larger trees removing more pollution.
- Given a small tree, `Crown_Base_Height` has little impact on `Pollution_Removal`
- Given a large tree, those with lower `Crow_Base_Height` tend to have higher `Pollution_Removal`.

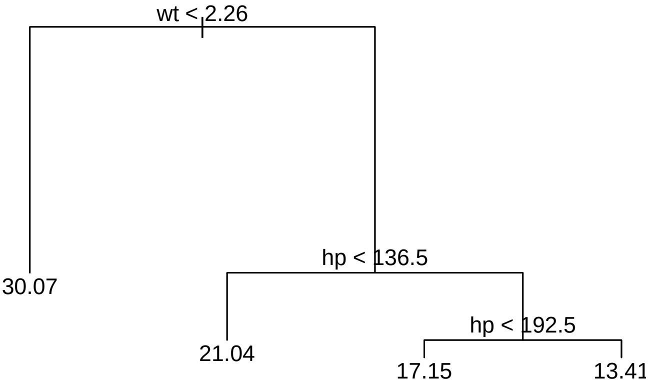
Extra Practice

The `mtcars` dataset gives the `mpg`, `hp`, and `wt` for 32 car models.

Extra Practice

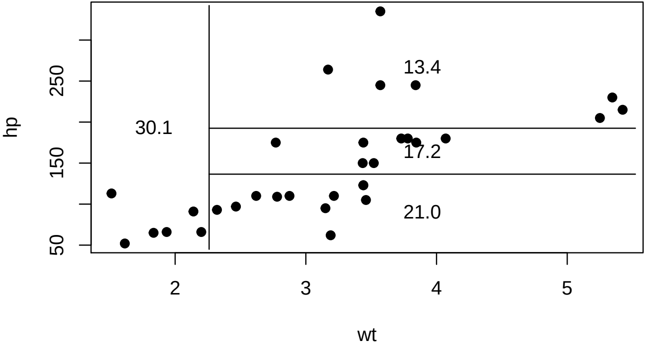
The `mtcars` dataset gives the `mpg`, `hp`, and `wt` for 32 car models.

Using the whiteboard in breakout rooms, draw the predictor space corresponding to the following tree, predicting `mpg` based on `wt` and `hp`.



What would you expect the signs of the corersponding regression slopes to be?

Results




```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.22727012 1.59878754 23.284689 2.565459e-20
## hp          -0.03177295 0.00902971 -3.518712 1.451229e-03
## wt          -3.87783074 0.63273349 -6.128695 1.119647e-06
```