Regression Trees

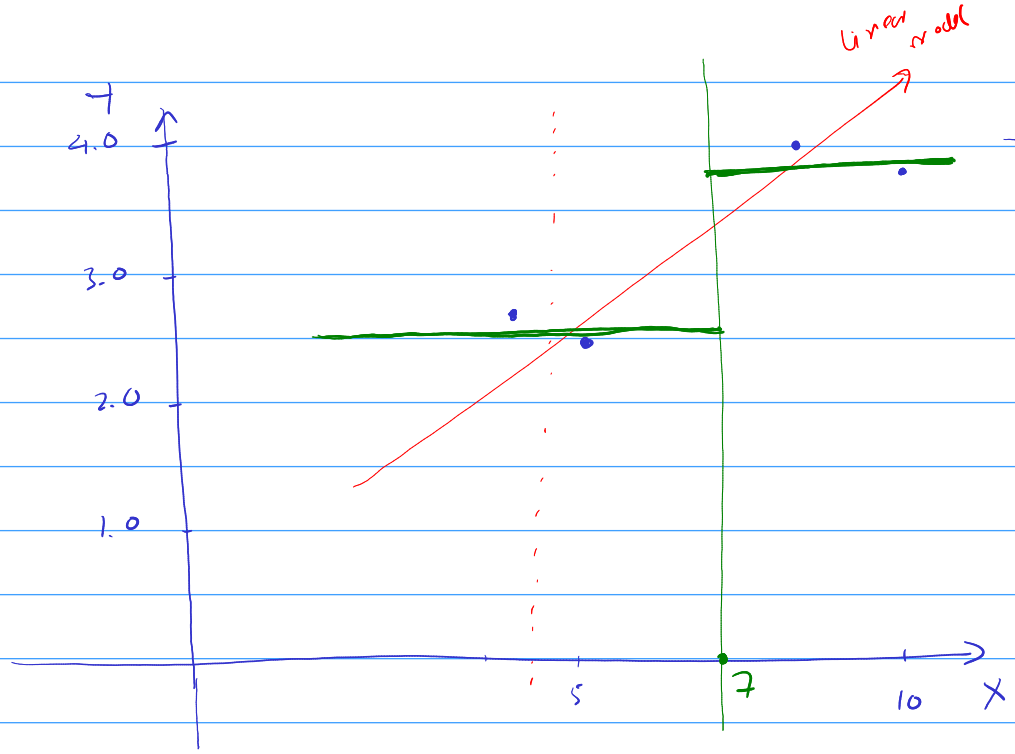
Regression Trees

Regression Trees

- The tree will search for all combination of predictors and cutoff value to decide the best split
- In Regression tree, the best split is the split that minimizes

$$\underbrace{\sum_{i:\mathbf{x}_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2}_{\text{RSS of obs. in left branch}} + \underbrace{\sum_{i:\mathbf{x}_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2}_{\text{RSS of obs. in right branch}}$$

- $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the means of the responses falling in to the left branch and right branch, respectively.



x	y
10	3.9
9	4.0
4	3.6
5	3.5

$X \geq 7$

False: $Y = 3.6$, $Y = 3.5$

predict: $\frac{3.6 + 3.5}{2} = 3.55$

True: $Y = 3.9$, $Y = 4.0$

predict $= \frac{3.9 + 4.0}{2} = 3.95$

Sum Square Errors of this Regression tree.

x	y	$\hat{Y}$ (tree predicts)	Squared errors
10	3.9	3.95	$(3.95 - 3.9)^2$
9	4.0	3.95	$(3.95 - 4)^2$
4	3.6	3.55	$(3.55 - 3.6)^2$
5	3.5	3.55	$(3.55 - 3.5)^2$

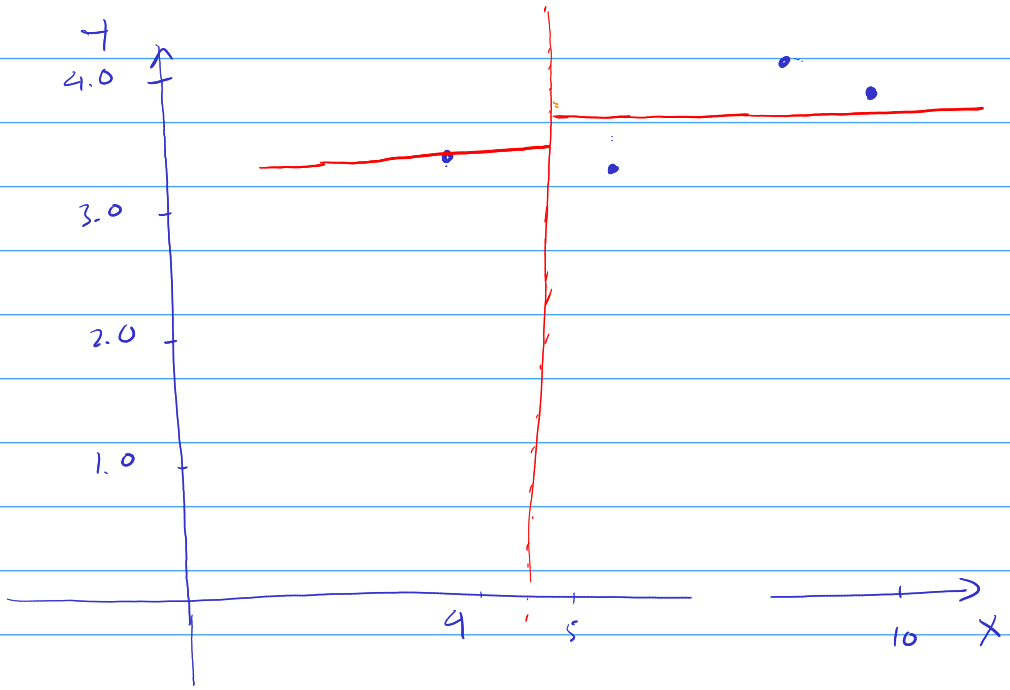
$SSE = .01$

$x \geq 4.5$

True

$y = 3.6$

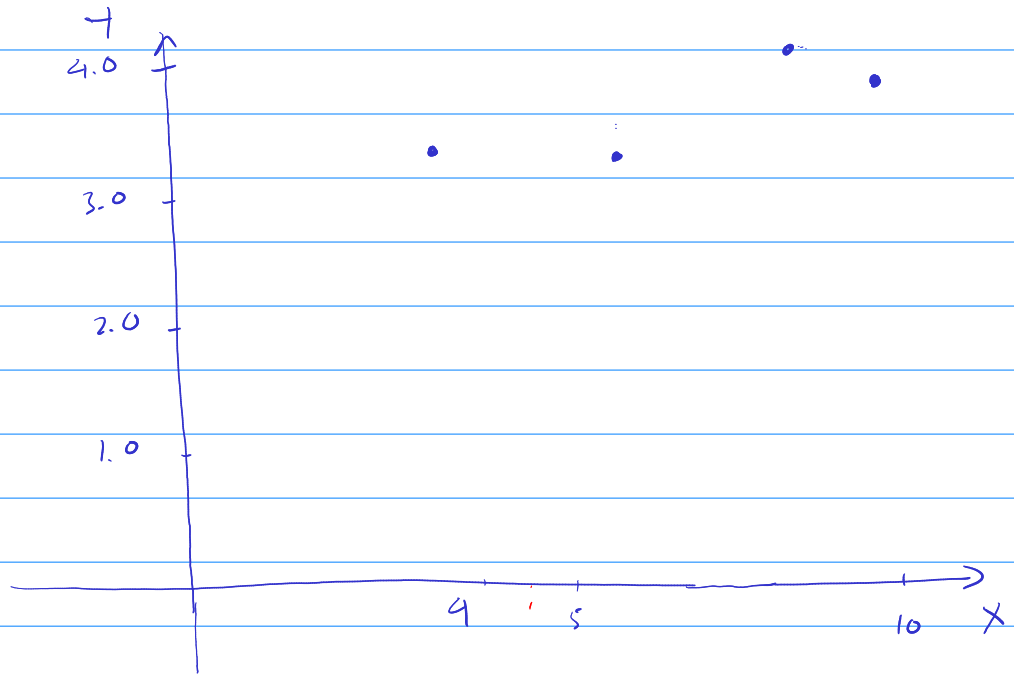
$y = 3.9, 4.0, 3.5 \Rightarrow \text{predict} = \frac{3.9 + 4.0 + 3.5}{3} = 3.8$



x	y	$\hat{y}$
10	3.9	3.8
9	4.0	3.8
4	3.6	3.6
5	3.5	3.8

$$SSE = (3.9 - 3.8)^2 + (4.0 - 3.8)^2 + (3.6 - 3.6)^2 + (3.5 - 3.8)^2$$

$$= .14$$



Example

X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

Using the RSS to decide the best split among

- Split 1: Region 1 $X_1 < 4$, Region 2 $X_1 \geq 4$
- Split 2: Region 1 $X_2 < 2$, Region 2 $X_2 \geq 2$

Example

X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

Using the RSS to decide the best split among

- Split 1: Region 1 $X_1 < 4$, Region 2 $X_1 \geq 4$
- Split 2: Region 1 $X_2 < 2$, Region 2 $X_2 \geq 2$

Example

X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

Using the RSS to decide the best split among

- Split 1: Region 1 $X_1 < 4$, Region 2 $X_1 \geq 4$
- Split 2: Region 1 $X_2 < 2$, Region 2 $X_2 \geq 2$