Regression Trees

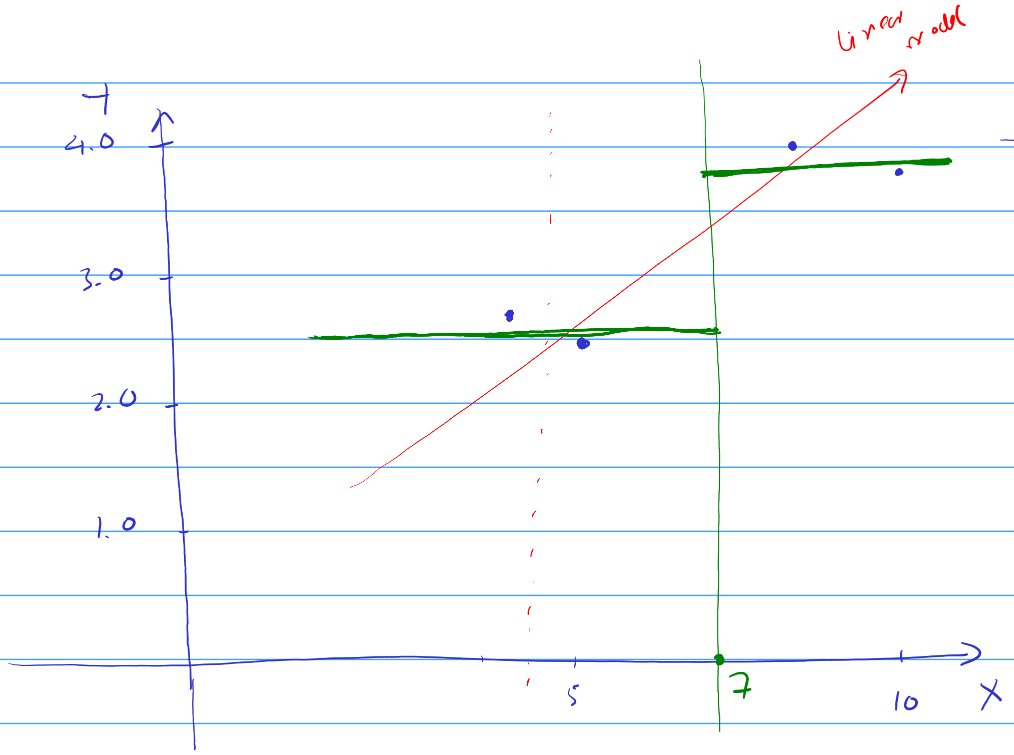
Regression Trees

Regression Trees

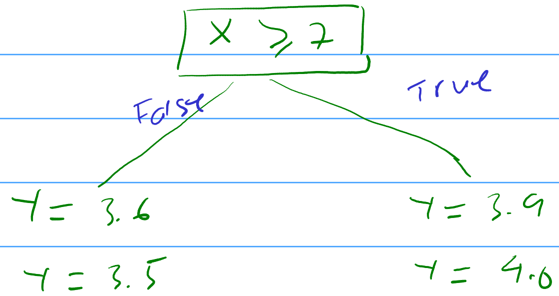
- The tree will search for all combination of predictors and cutoff value to decide the best split
- In Regression tree, the best split is the split that minimizes

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

- $\widehat{y}_{R_1}$ and $\widehat{y}_{R_2}$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



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

$= 3.55$

predict

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

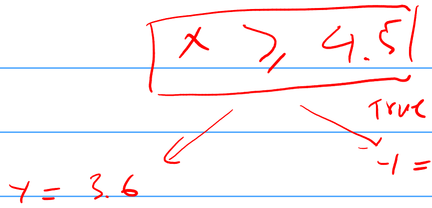
$= 3.95$

Sum Square Errors of this Regression tree.

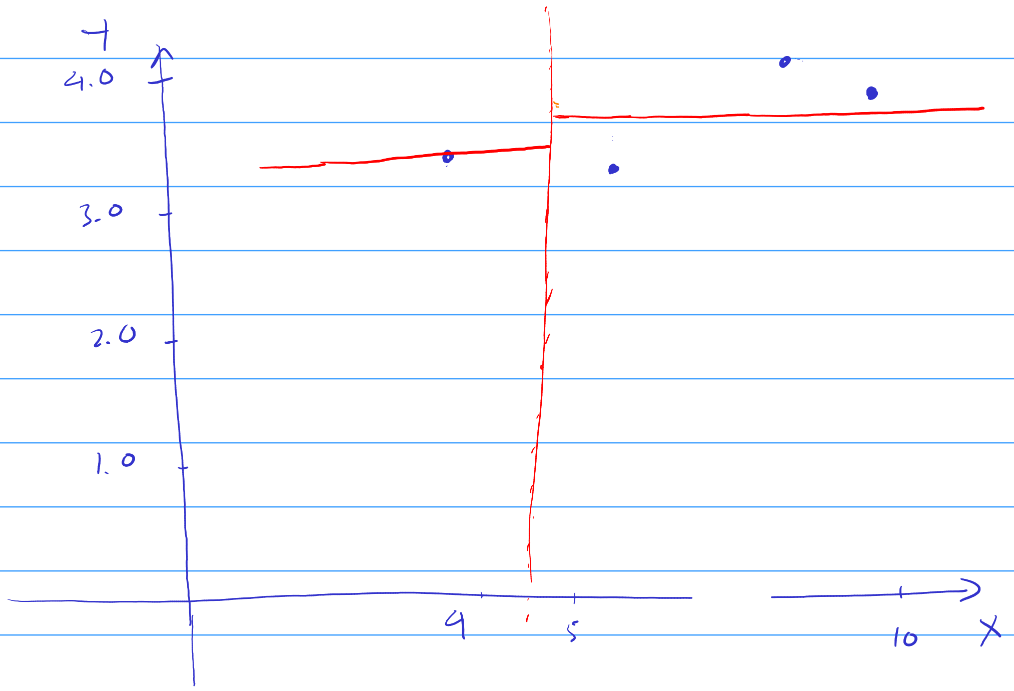
x	Y	$\hat{Y}$ (tree predicts)	Squared errors
10	3.9	3.95	$(3.95 - 3.9)^2$
9	4.0	3.95	$(3.95 - 4)^2$
4	3.6	3.55	$(3.55 - 3.6)^2$
5	3.5	3.55	$(3.55 - 3.5)^2$

+

SSE = .01



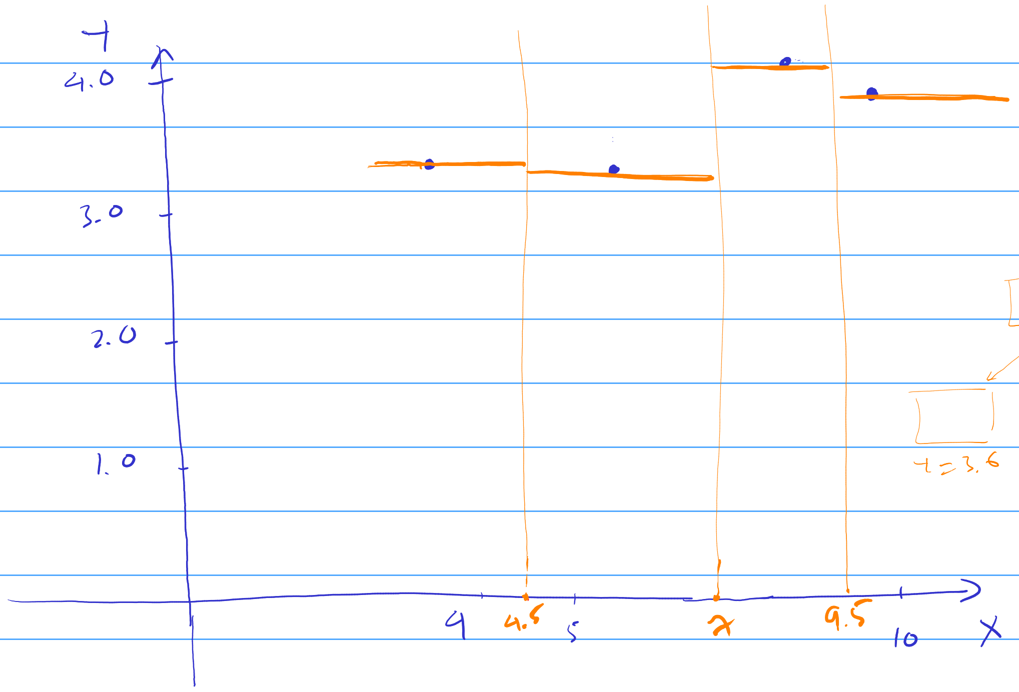
$y = 3.9,\ 4.0,\ 3.5 \Rightarrow \text{predict} = \dfrac{3.9 + 4.0 + 3.5}{3} = 3.8$

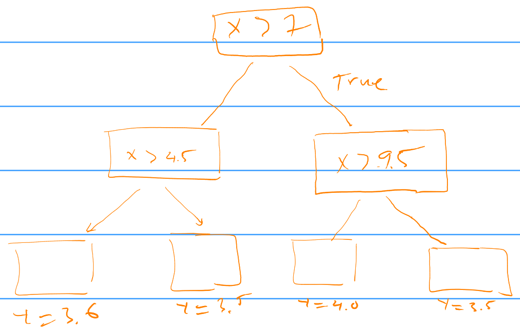


x	y	$\hat{y}$
10	3.9	3.8
9	4.0	3.8
4	3.6	3.6
5	3.5	3.8

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

$= .14$





$SSE = 0$

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

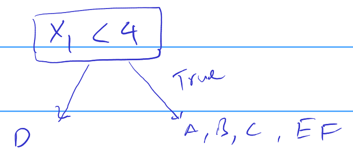
RSS = SSE

	X_1	X_2	Y
A	1	0	1.2
B	2	1	2.1
C	3	2	1.5
D	4	1	3.0
E	2	2	2.0
F	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 $X_2 \geq 2$

Split 1 :



	X_1	X_2	Y
A	1	0	1.2
B	2	1	2.1
C	3	2	1.5
D	4	1	3.0
E	2	2	2.0
F	1	1	1.6

$y = 3.0$

↓

$\hat{y} = 3.0$

$y = \{1.2, 2.1, 1.5, 2.0, 1.6\}$

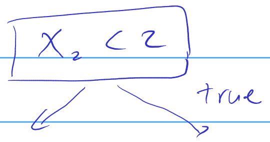
$$\hat{y} = \frac{1.2 + 2.1 + 1.5 + 2.0 + 1.6}{5} = 1.68$$

$$RSS = SSE = \underbrace{\sum (y - \hat{y})^2}_{\text{branch 1}} + \underbrace{\sum (y - \hat{y})^2}_{\text{branch 2}}$$

$$= (3 - 3.0)^2 + (1.2 - 1.68)^2 + (2.1 - 1.68)^2 + (1.5 - 1.68)^2 + (2.0 - 1.68)^2 + (1.6 - 1.68)^2$$

$$= .548$$

Split 2 :



C , E

$y = 1.5, 2.0$

$$\hat{y} = \frac{1.5 + 2.0}{2}$$

$= 1.75$

A, B, D , F

$y = 1.2, 2.1, 3.0, 1.6$

$$\hat{y} = \frac{1.2 + 2.1 + 3.0 + 1.6}{4}$$

$= 1.975$

	X_1	X_2	Y
A	1	0	1.2
B	2	1	2.1
C	3	2	1.5
D	4	1	3.0
E	2	2	2.0
F	1	1	1.6

$$SSE = \underbrace{(1.5 - 1.75)^2 + (2.0 - 1.75)^2}_{\text{left branch}} + \underbrace{(1.2 - 1.975)^2 + (2.1 - 1.975)^2 + (3.0 - 1.975)^2 + (1.6 - 1.975)^2}_{\text{right branch}}$$

$$\boxed{SSE = 1.9325}$$

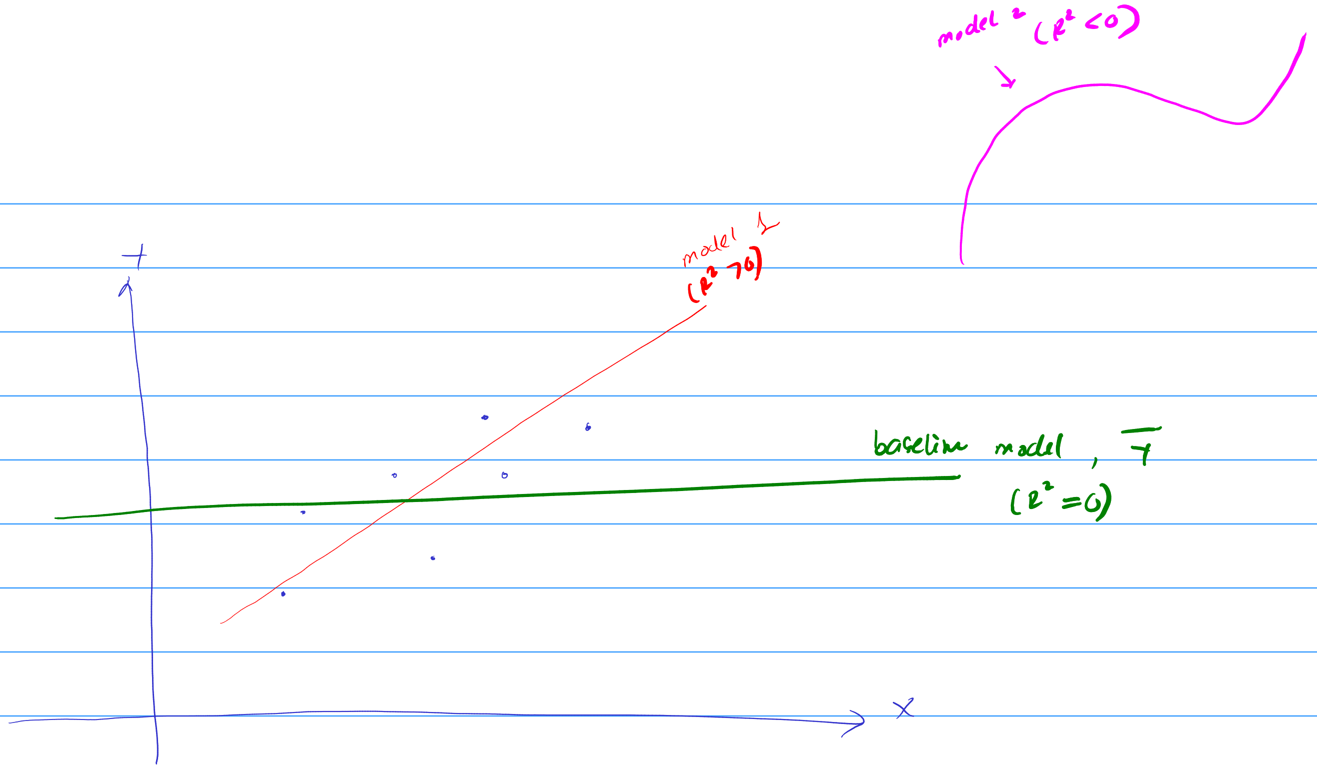
Example

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 $X_2 \geq 2$

(★) R^2



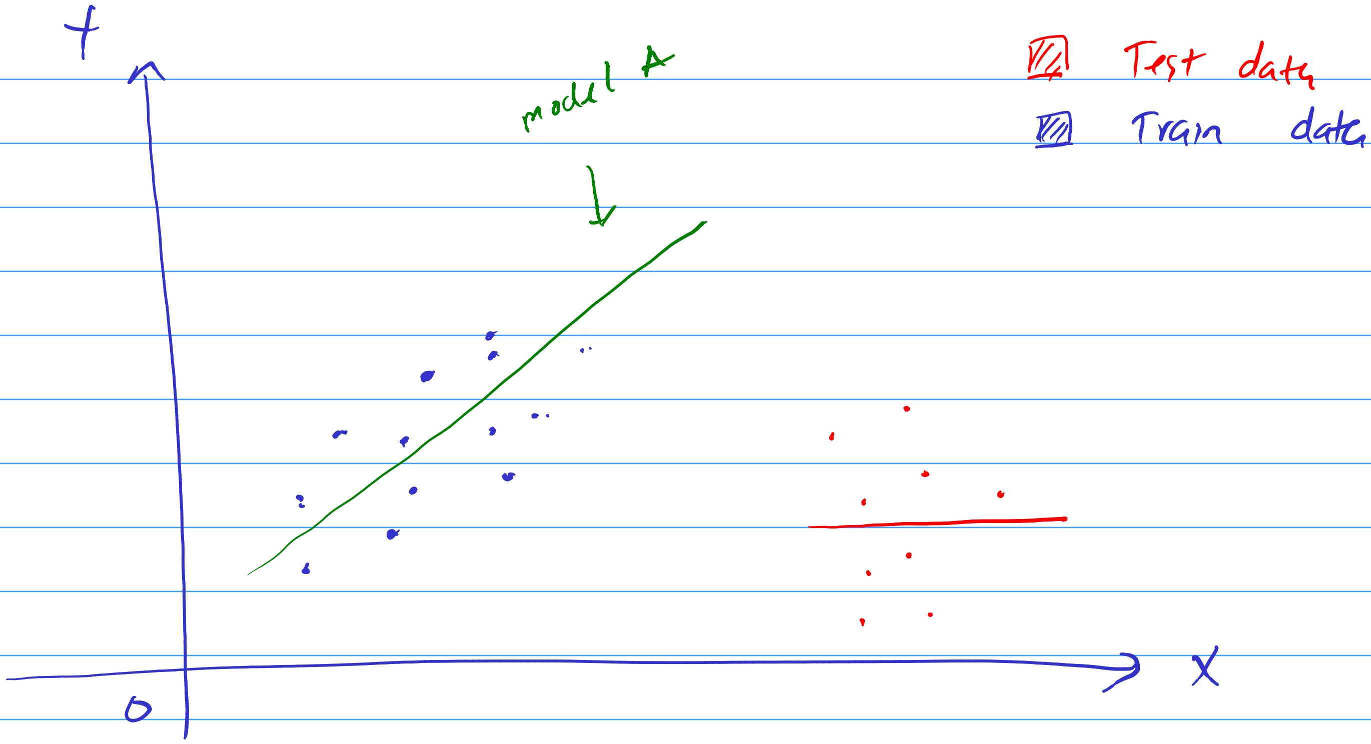
$$R^2 = 1 - \frac{\text{SSE of the model}}{\text{SSE of the baseline model}, \bar{y}}$$

(1) when the model A is the baseline model, R^2 of A $= 0$

(2) when the model A is "perfect", $R^2 = 1 - \frac{0}{\text{SSE of baseline}}$

$\Rightarrow$ $\boxed{R^2 = 1}$

(3) when the model is "worse" than the baseline model, $R^2 < 0$



R^2 of A on training is positive but on testing is negative

(Back to the example)

Split 1 : $\boxed{X_1 < 4}$

True

D $\qquad$ A, B, C, EF

$y = 3.0$ $\qquad$ $y = \{1.2, 2.1, 1.5, 2.0, 1.6\}$

$\hat{y} = 3.0$ $\qquad$ $\hat{y} = \frac{1.2 + 2.1 + 1.5 + 2.0 + 1.6}{5} = 1.68$

	X_1	X_2	Y
A	1	0	1.2
B	2	1	2.1
C	3	2	1.5
D	4	1	3.0
E	2	2	2.0
F	1	1	1.6

Let calculate the R^2 of this model | split

SSE of the baseline model $= \Sigma(y - \bar{y})^2$

$$\bar{y} = \frac{1.2 + 2.1 + 1.5 + 3.0 + 2.0 + 1.6}{6} = 1.9$$

$$\text{SSE of } \bar{y} = (1.2 - 1.9)^2 + (2.1 - 1.9)^2$$

$$+ (1.5 - 1.9)^2 + (3.0 - 1.9)^2$$

$$+ (2 - 1.9)^2 + (1.6 - 1.9)^2$$

$$= 2$$

$$R^2 = 1 - \frac{.548}{2} = .726.$$