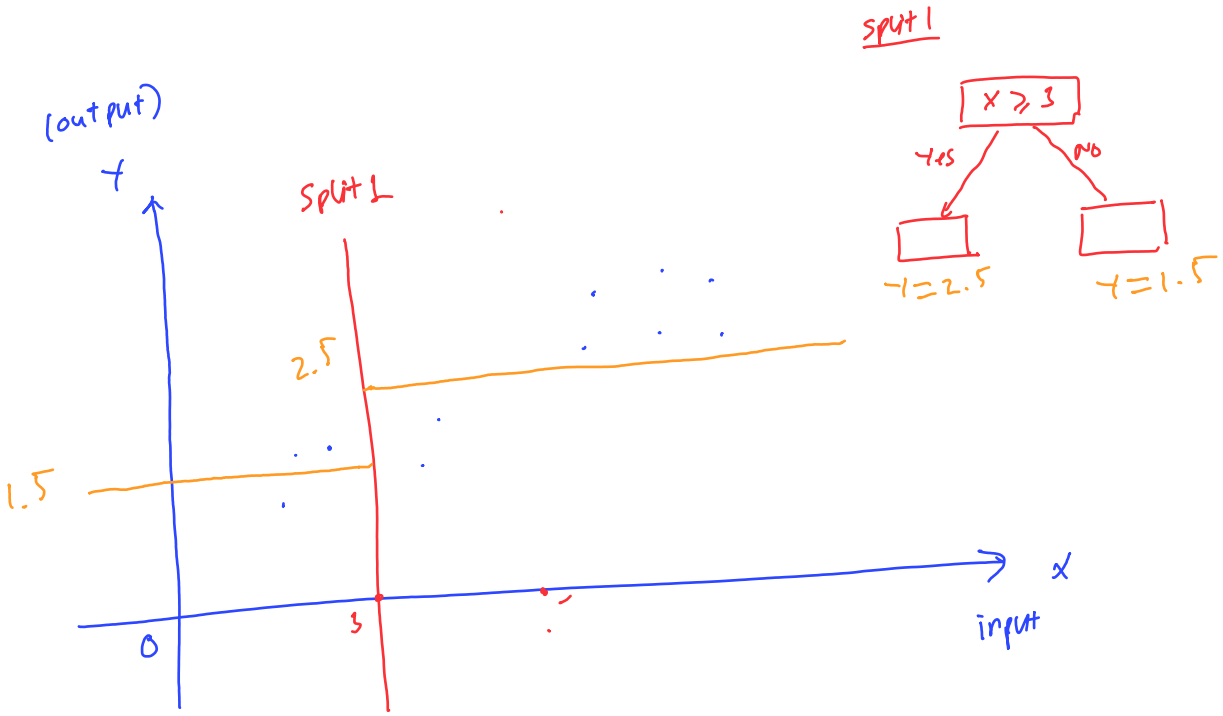
Regression Trees

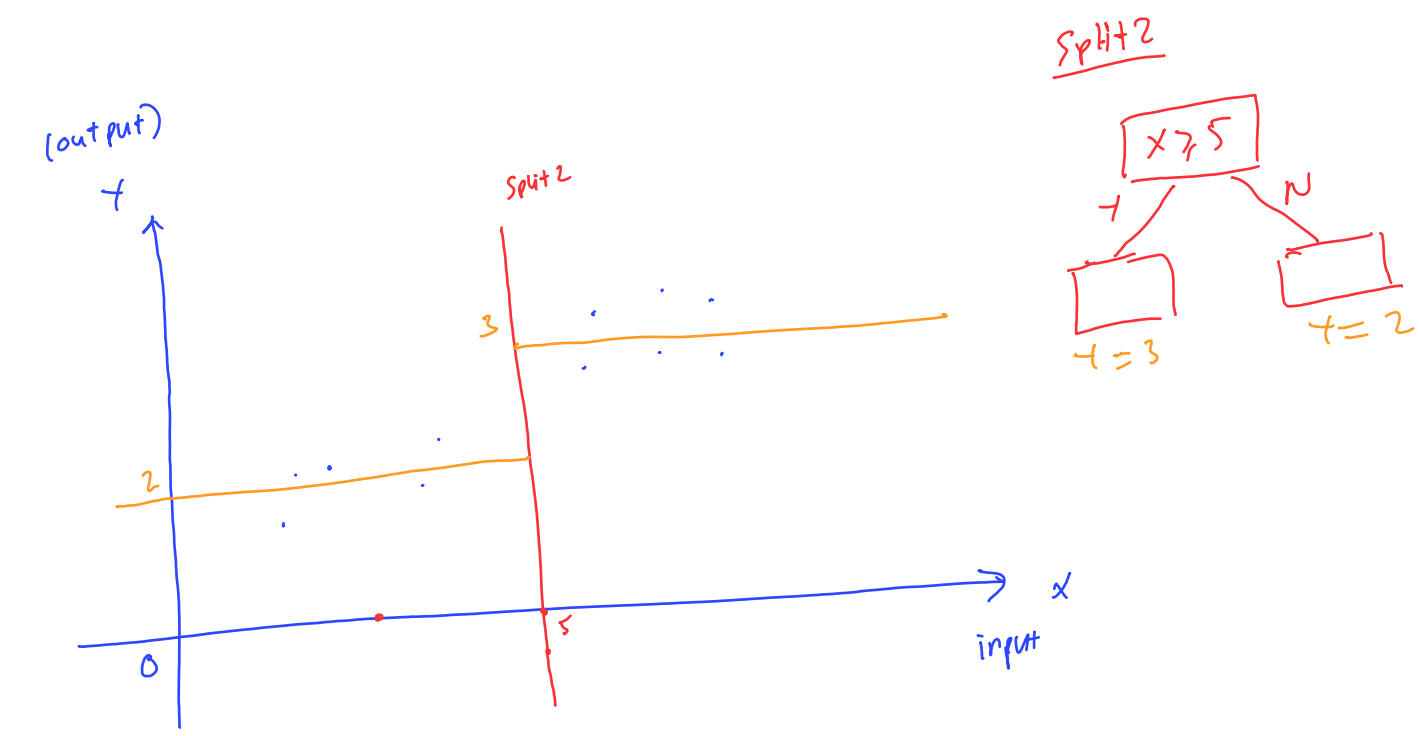
Regression Trees

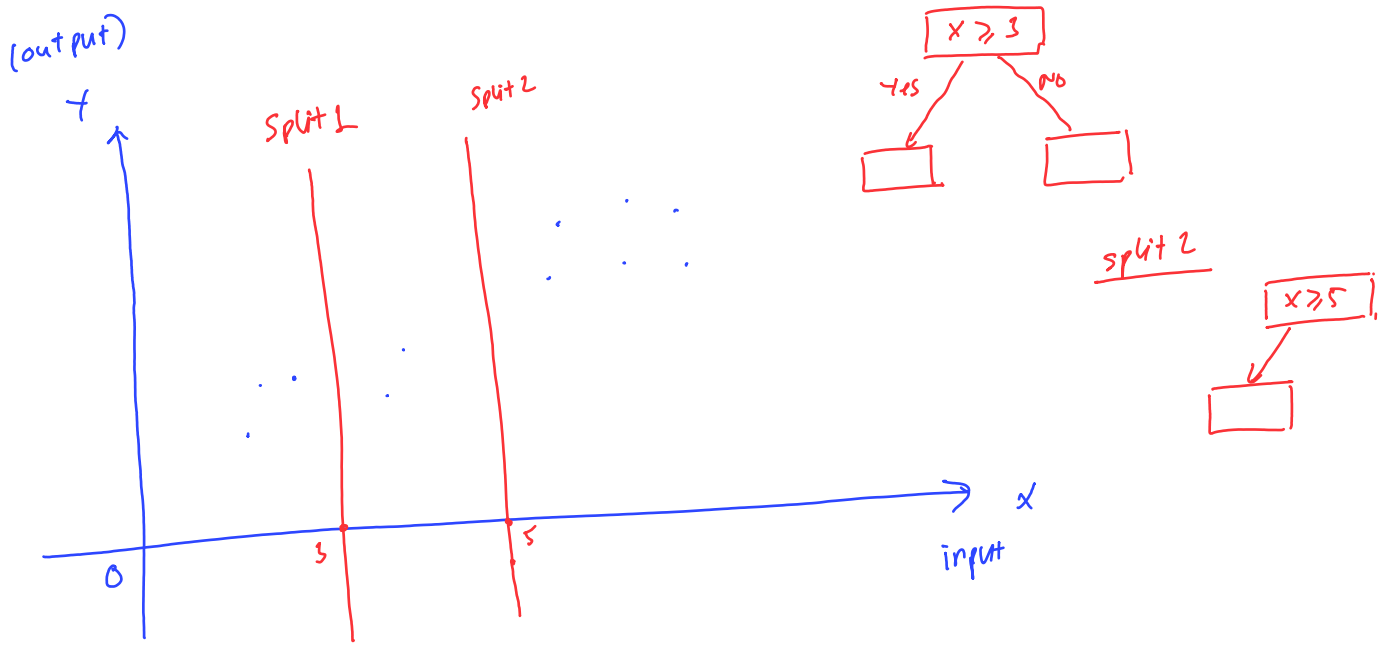
- The tree will search for all combination of predictors and cutoff value to decide the best split
- In Regression tree, the best split is the split that minimizes

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

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

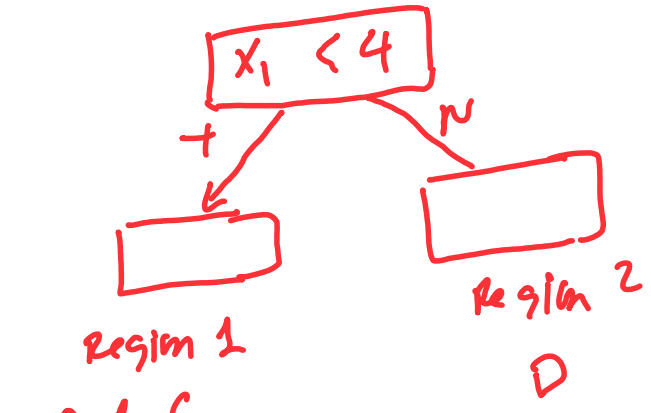

split 1
x ≥ 3
yes
no
y=2.5
y=1.5
(output)
y
Split 1
2.5
1.5
0
3
x
input


Split 2
x ≥ 5
y
N
y=3
y=2
(output)
y
Split 2
3
2
0
5
x
input


(output)
y
Split 1
Split 2
0
3
5
x
input
x ≥ 3
Yes
No
split 2
x ≥ 5

Example

	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

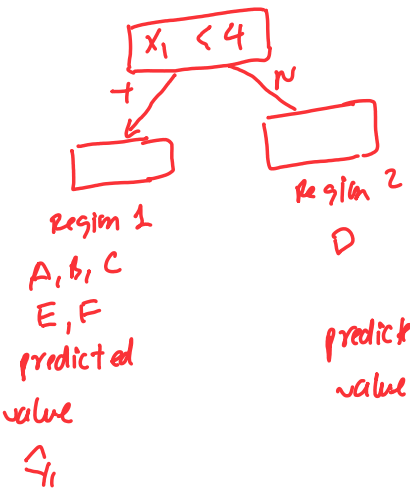


$$\hat{y}_1 = \frac{1.2 + 2.1 + 1.5 + 2.0 + 1.6}{5}$$

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 :



$\hat{y}_1 = 1.68$

$\hat{y}_2 = 3.0$

split 1

X_1	X_2	Y	$\hat{y}$
1	0	1.2	1.68
2	1	2.1	1.68
3	2	1.5	1.68
4	1	3.0	3.0
2	2	2.0	1.68
1	1	1.6	1.68

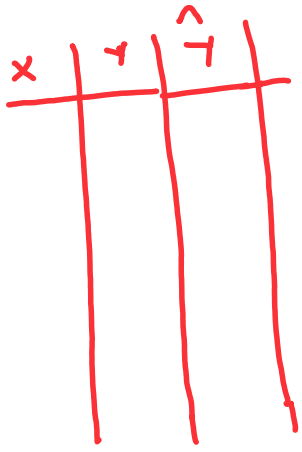
$$RSS = \sum (y - \hat{y})^2$$

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

$$RSS = .548$$

Similarly we can calculate RSS of Split 2.

⑧ R^2



$$RSS = \sum (y - \hat{y})^2$$

* The base line model :

Predict y without using information of x.

Predict y by $\bar{y}$: $\hat{y} = \bar{y}$

RSS of this model : $\sum (y - \bar{y})^2$

(*)

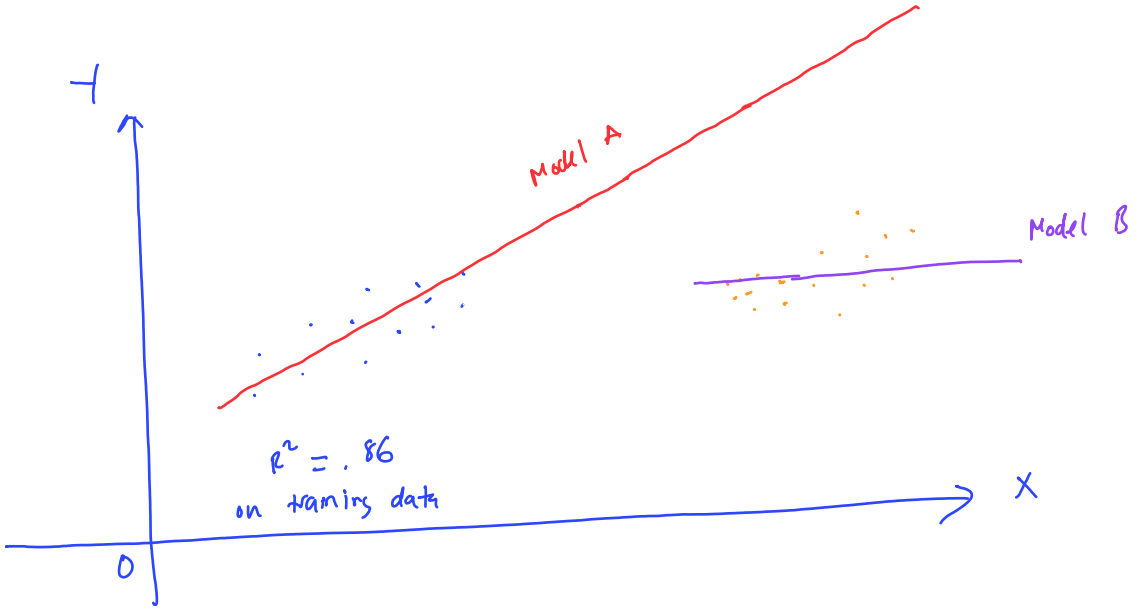
$$R^2 \text{ of Model A} = 1 - \frac{\text{RSS of the model A}}{\text{RSS of the baseline model}}$$

$$= 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

(1) If RSS of Model A $= 0$, $\Rightarrow R^2 = 1$

(2) If Model A is as good as the baseline model $\Rightarrow R^2 = 0$

(3) Can $R^2 < 0$?



on validation data: model A is worse than model B (baseline model)

$\Rightarrow$ R^2 of model A on validation data would be negative.

Suppose that your regression tree contain only one split which is the best split in the previous question. Calculate the R^2 of this regression tree on the training data.

RSS of the baseline model

X_1	X_2	Y	$\hat{Y} = \bar{Y}$
1	0	1.2	1.9
2	1	2.1	1.9
3	2	1.5	1.9
4	1	3.0	1.9
2	2	2.0	1.9
1	1	1.6	1.9

$$\Rightarrow \sum (Y - \bar{Y})^2$$

$$= (1.2 - 1.9)^2 + (2.1 - 1.9)^2 + (1.5 - 1.9)^2 + (3 - 1.9)^2 + (2 - 1.9)^2 + (1.6 - 1.9)^2$$

$$= 2$$

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

Use your regression tree to predict the y for the below testing data. Calculate the R^2 of the tree on the testing data.

x_1	x_2	y
3	1	3.0
1	5	3.6
5	1	4.0
5	2	3.9