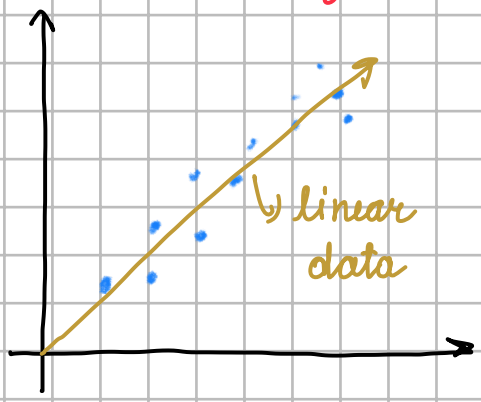
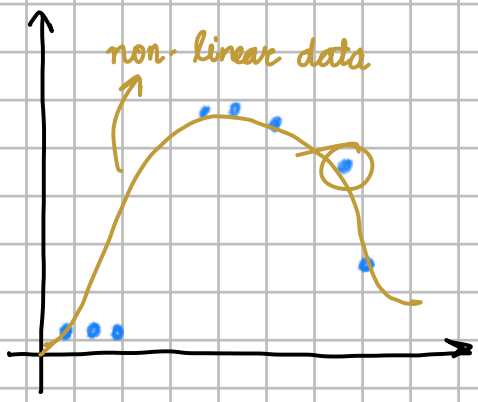
REGRESSION TREE (FROM SCRATCH)

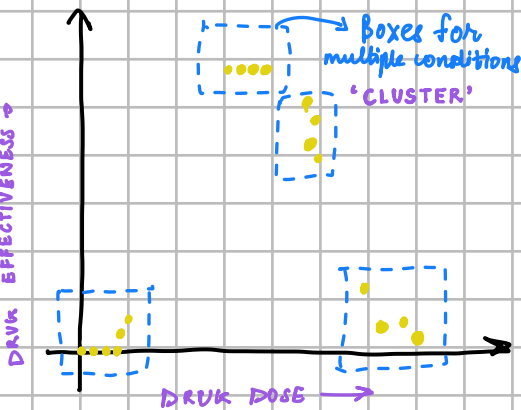
Can handle non-linear data much better than the other regression methods.





Regression tree is used here.

* Visualization

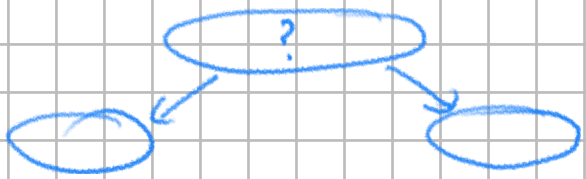


Why does this tree makes good prediction??

Problem: We have a training dataset which consists of drug effectiveness of different dosage.

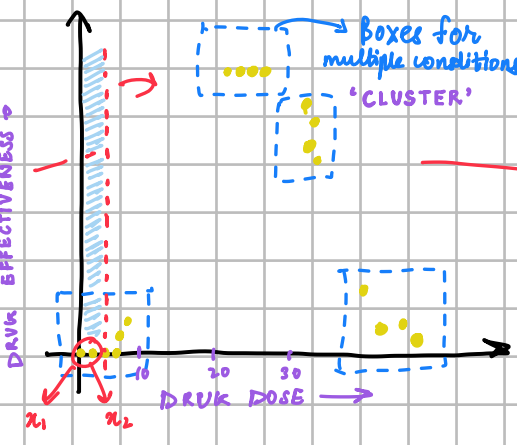
Given this training data, we want to build a regression tree that uses drug doses to predict drug effectiveness.

Step ①: We need to decide root node first!



Step ② Calculate the avg of first two doses (equals to 3)

$\frac{x_1 + x_2}{2} = 3$ (avg)



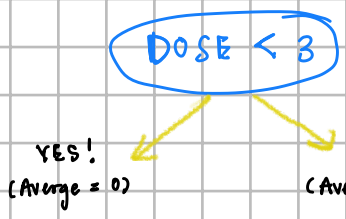
DOSE < 3

Drug effectiveness = 0
(avg is zero this side)

DOSE > 3

As we can see all points are there.
Drug effectiveness = Avg for all points
= 38.8 (Avg)

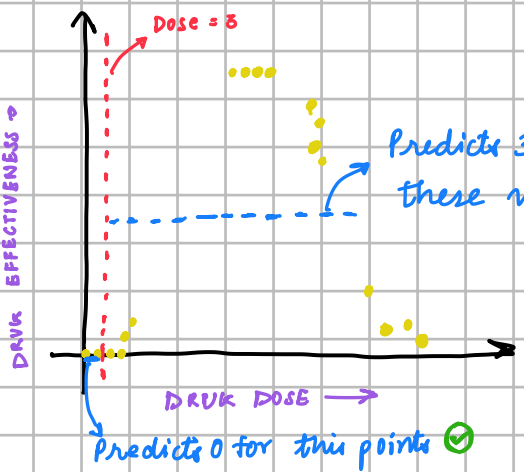
Tree will be formed like :-



← Just a trial!!

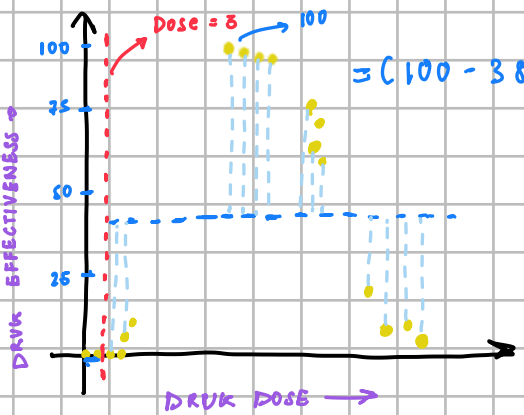
From visualization, we can say that this decision tree is not doing good job.

Let's see visual representation of this problem.



Step ③ : Sum of squared residuals (SSR)

Quantify how good or bad decision tree is!



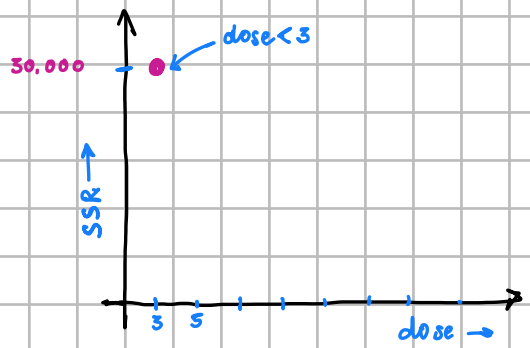
$= (100 - 38.8)^2$

Residual is a diff. b/w true value & predicted value.

$$SSR = (0-0)^2 + (0-38.8)^2 + (0-38.8)^2 + (5-38.8)^2 + (20-38.8)^2 + (100-38.8)^2 \dots (0-38.8)^2$$

$= 27468.5$

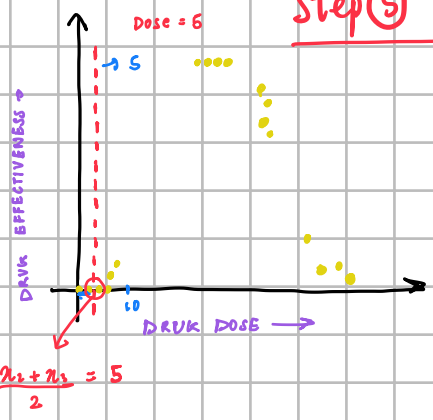
Step ④ : Plot SSR graph with dose on x axis & SSR on Y axis



But, we gotta know that dose < 3 & this is not our root node.

We will repeat this step of SSR until we get accurate value.

Step ⑤ : Repeat the above for different threshold.



for dose < 5

drug effectiveness = 0 (again both points is at zero value)

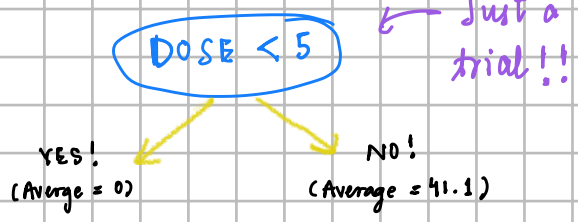
for dose > 5

drug effectiveness = Avg for all points on right side.

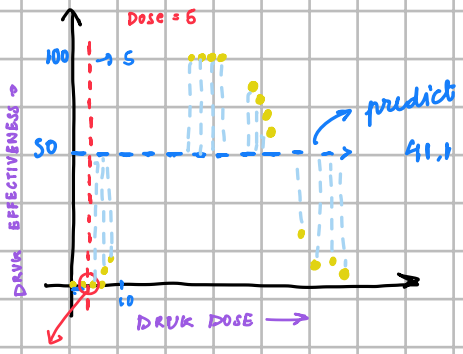
$= 41.4$

Visualization !

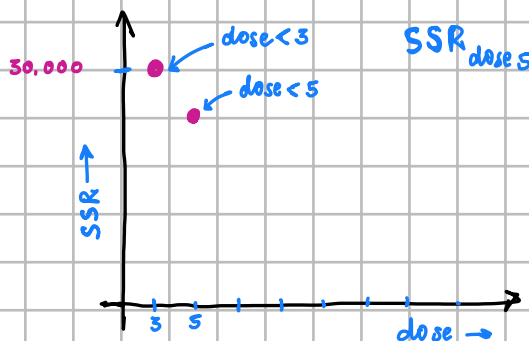
Tree will be formed like :-



← Just a trial !!



Let's plot SSR graph for 5 as well.

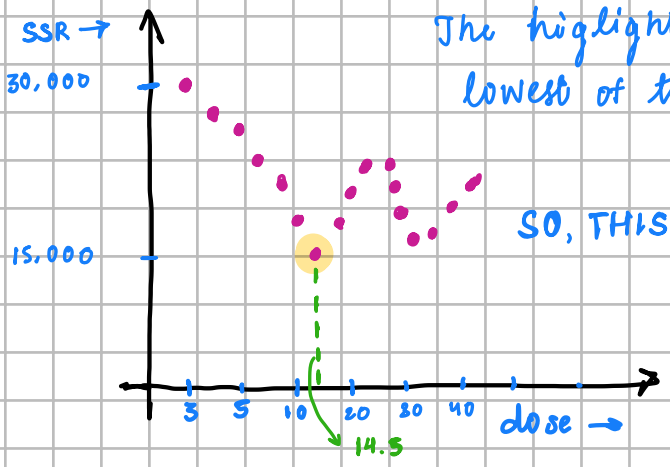


$SSR_{dose\,5} < SSR_{dose\,3}$ $\quad \frac{x_1 + x_2}{2} = 5$

Calculate for different doses & plot in this same graph to get value of root node.

⑥ Similarly, calculate for dose < 7, dose < 9, dose < 14.5, dose < 26 and dose < 36

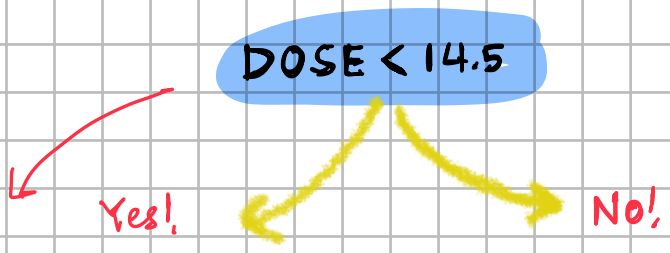
⑦ Plot them altogether in SSR graph :-



The highlighted point i.e. dosage < 14.5 gives the lowest of them all

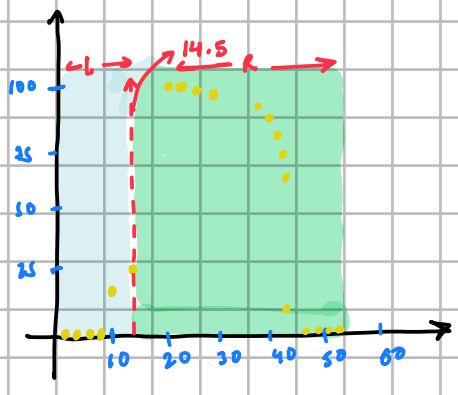
↓

SO, THIS WILL BE OUR ROOT NODE !!



Here, we got our root node

Step ⑧ : Our next goal is to find internal nodes



Dose < 14.5

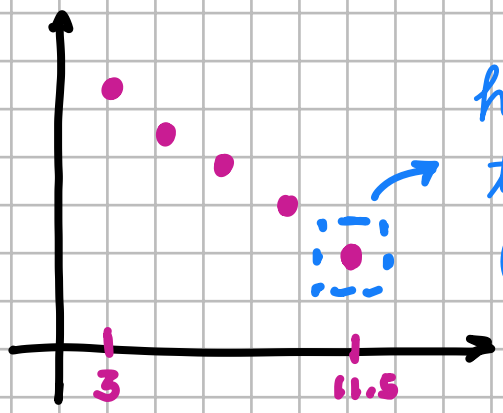
Yes! $\quad$ NO!

↓

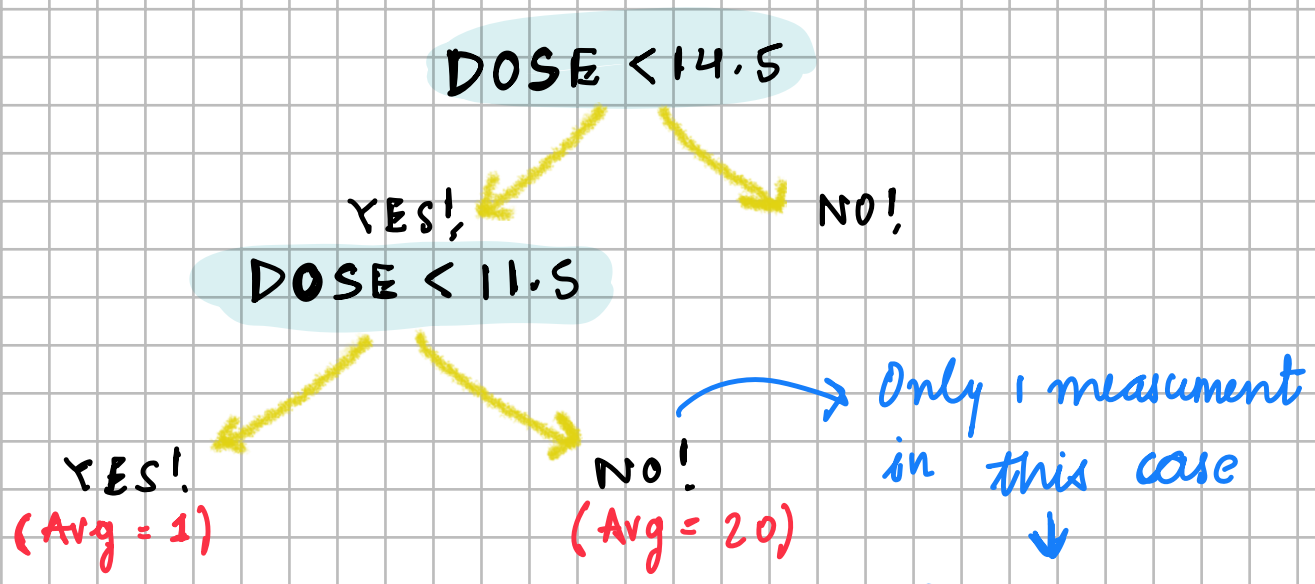
Left side : These 6 measurement should come here.

Step ⑨: Repeat SSR steps we did earlier for these 6 measurements.

Step ⑩: Plot the graph



here, dose < 11.5 gives lowest threshold (previously done in 14.5)



Step ⑪: Preventing overfitting

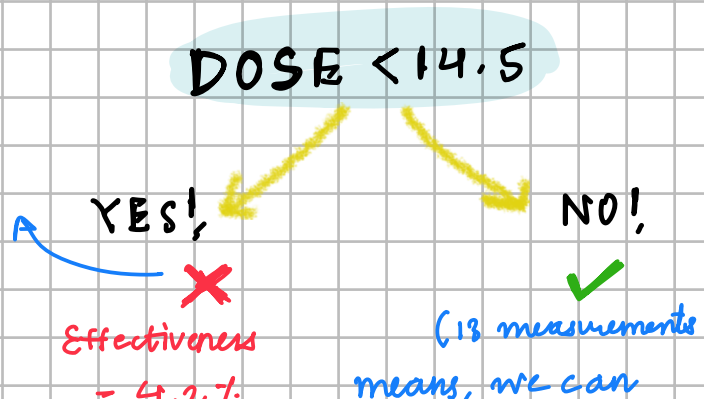
To prevent overfitting, we have to set the minimum no. of measurements for splitting tree.

Let's choose no. as 7

As, we know there are only 6 measurements (we cannot split it further)

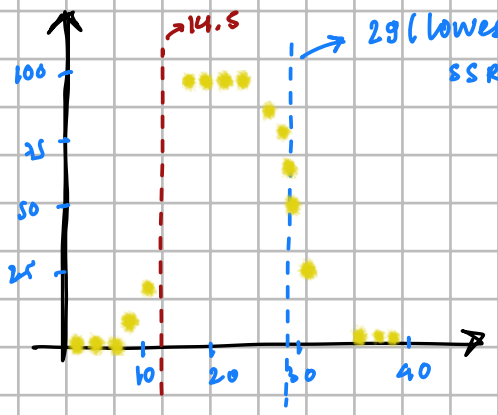
Effectiveness = avg of 6 measures.

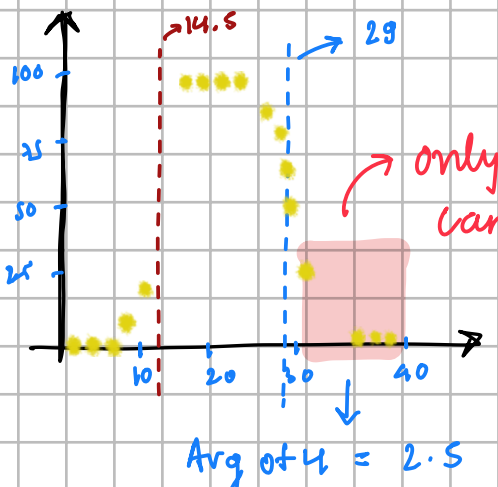
We will split the measurements from right side (into two groups & find lowest SSR)



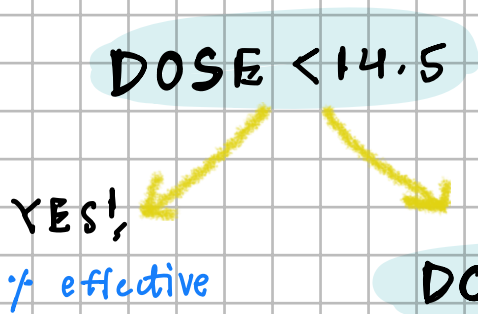
Again finding SSR values :

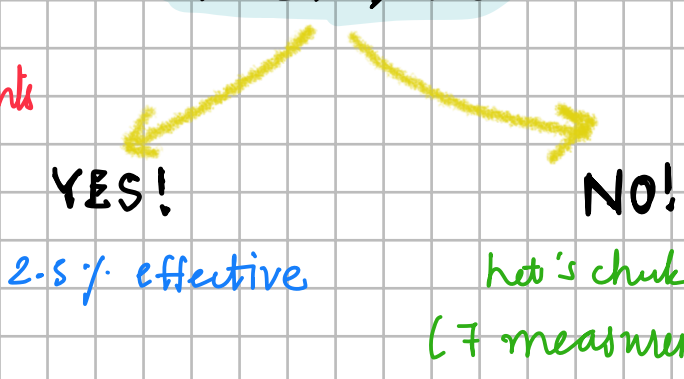
By this, we gotta know that our next internal node (right) would be dose > 29





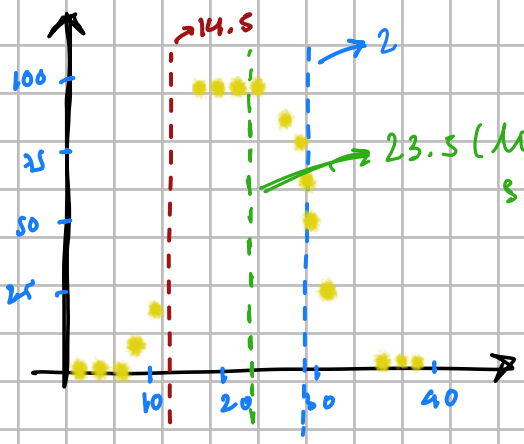
Avg of 4 = 2.5

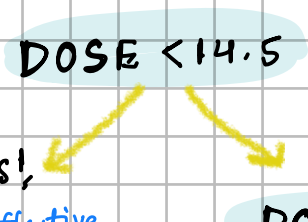


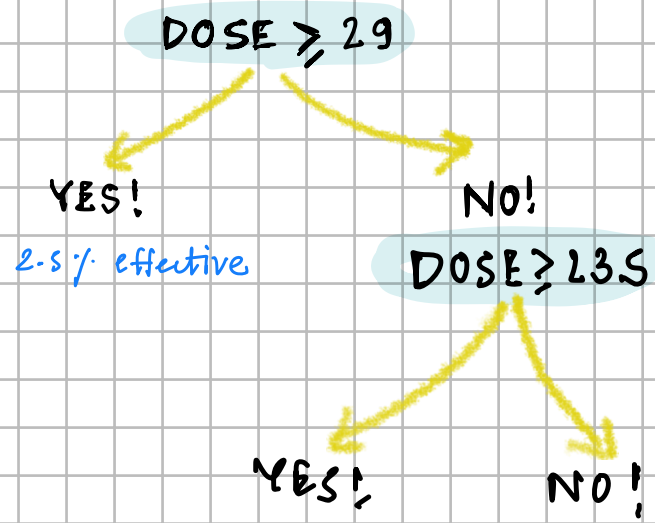


* Now, again we'll have to decide measurement to split b/w 14.5 and 29.

* Following same procedure, we will find lowest SSR.







* Here, both the values have lower value than 7. So we cannot split this further

Our final tree structure is ready !!

Multiple features in this tree

using dose, age and sex.

Dose	Age	Sex	Drug Effectiveness
10	25	F	98
20	73	M	0
35	54	F	6
5	12	M	44

* First, we'll completely ignore age and sex and use dose :-
* We select threshold with smallest sum of squared residuals (SSR) →

DOSE < 14.5

YES! NO!

* Then, we ignore dose and sex, and only age to predict drug effectiveness.
* By this, we individually find SSR and create preferences according to it

DOSE < 14.5

YES! NO!

(SSR = 19864)

Age > 50

YES! NO!

(SSR = 12017)

Sex = F

YES! NO!

(SSR = 20738)

lowest SSR will be root of regression tree. →

Age > 50

YES NO

ROOT NODE

* Then we grow the tree just like before the (only diff is that we'll have 3 candidates
  - Do this until we can no longer split further.

Age > 50

Yes! Age > 65

No! Sex = F

Yes! 0%

No! 5%

Yes! 75%

No! 39%

→ This final prediction is the drug effectiveness!