Statistics and Machine Learning

Decision tree regression

Remote lecture Week 11 03/29 — 04/02

Contents

Review of linear regression algorithm

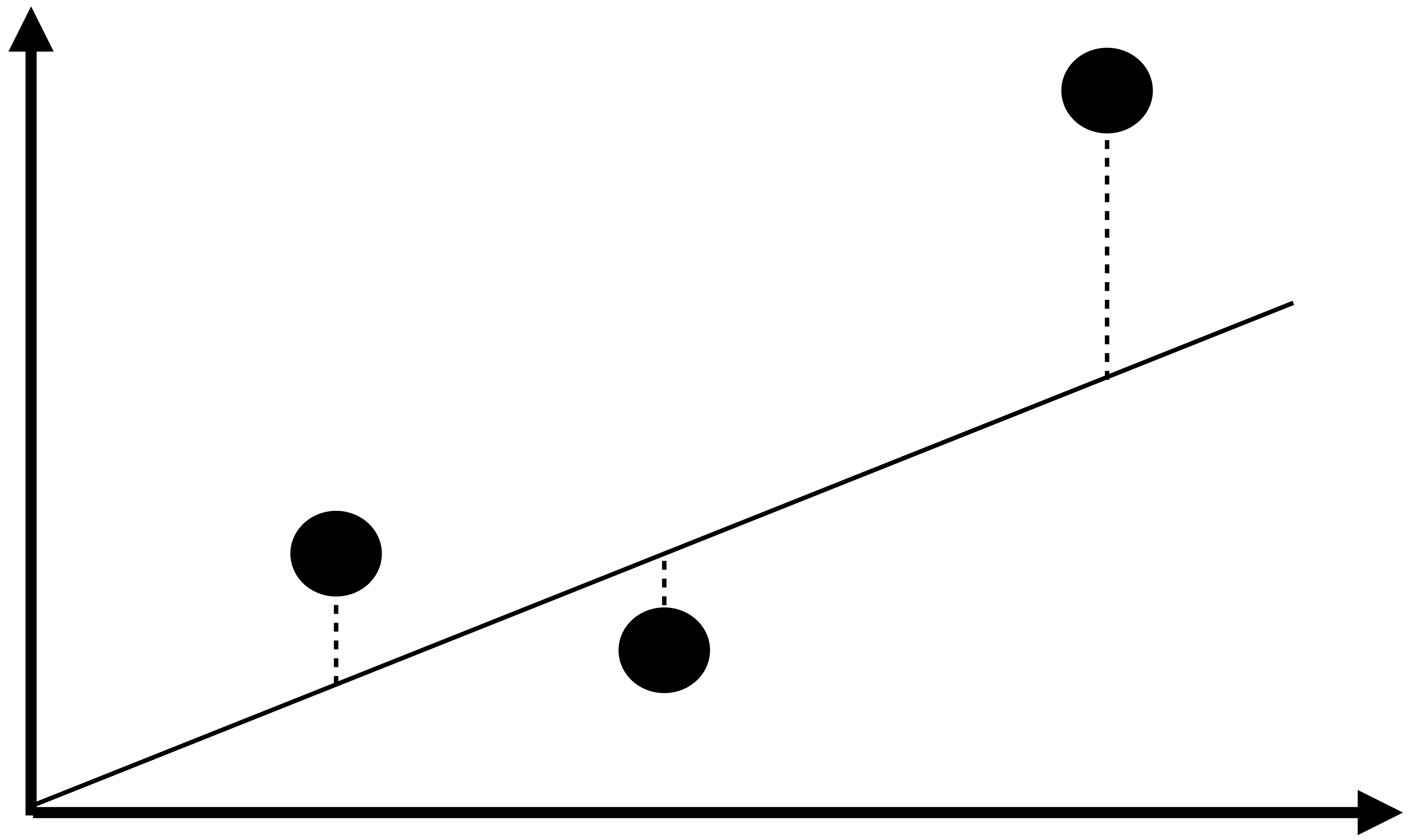
Why tree-based algorithm?

Tree-based regression algorithm (textbook 8.1.1)

Lab session: sk-learn library

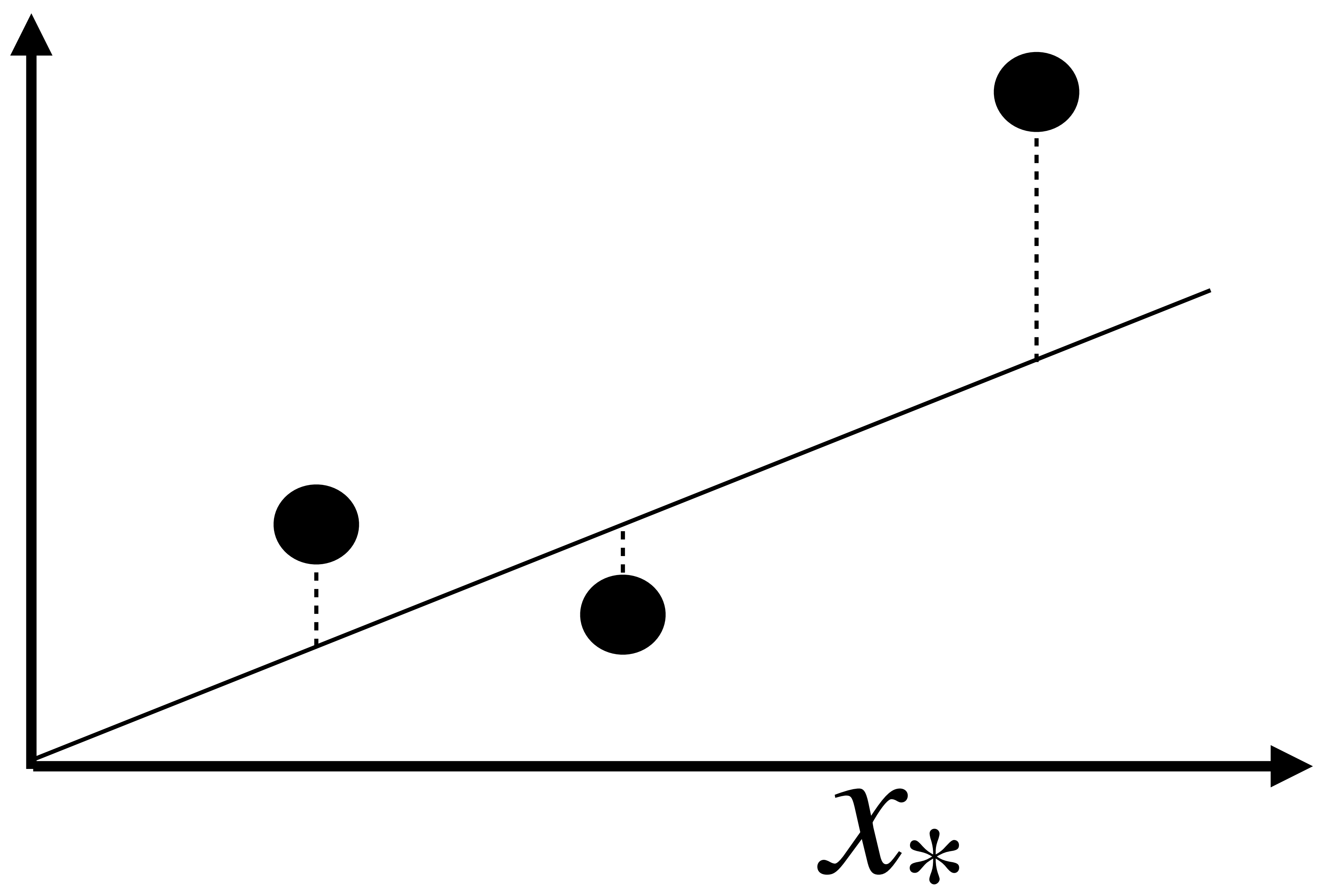
Quiz

Recall what linear regression does!



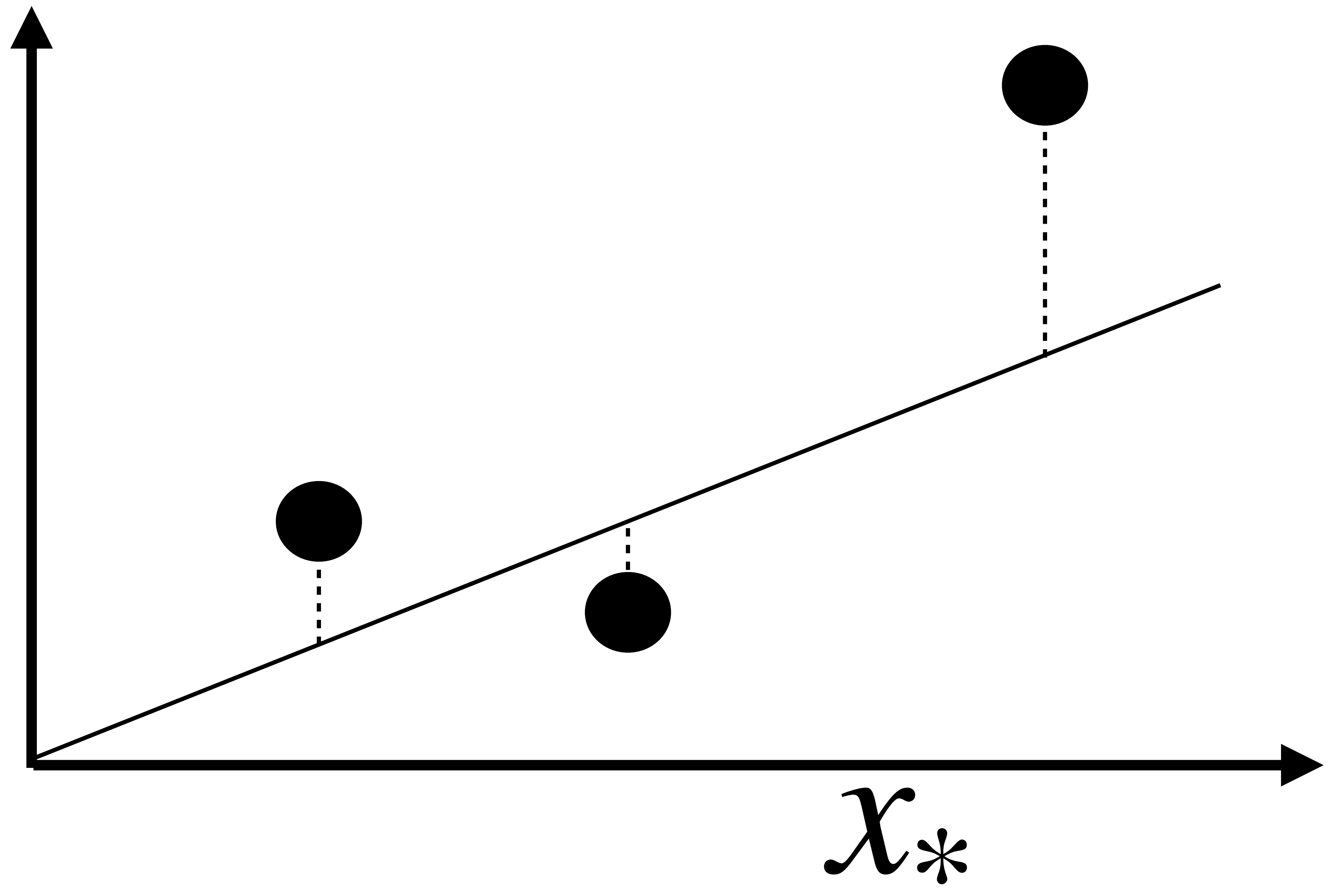
- Assume the straight line function is $y = ax + b$
- Calculate the Loss = sum of all squared error
- Tune 'a' and 'b' so that the Loss is minimized!
- Obtain the optimal 'a' and 'b', and we can use it in prediction

Recall what linear regression does!



- Assume the straight line function is y = ax + b
- Calculate the Loss = sum of all squared error
- Tune 'a' and 'b' so that the Loss is minimized!
- Obtain the optimal 'a' and 'b', and we can use it in prediction

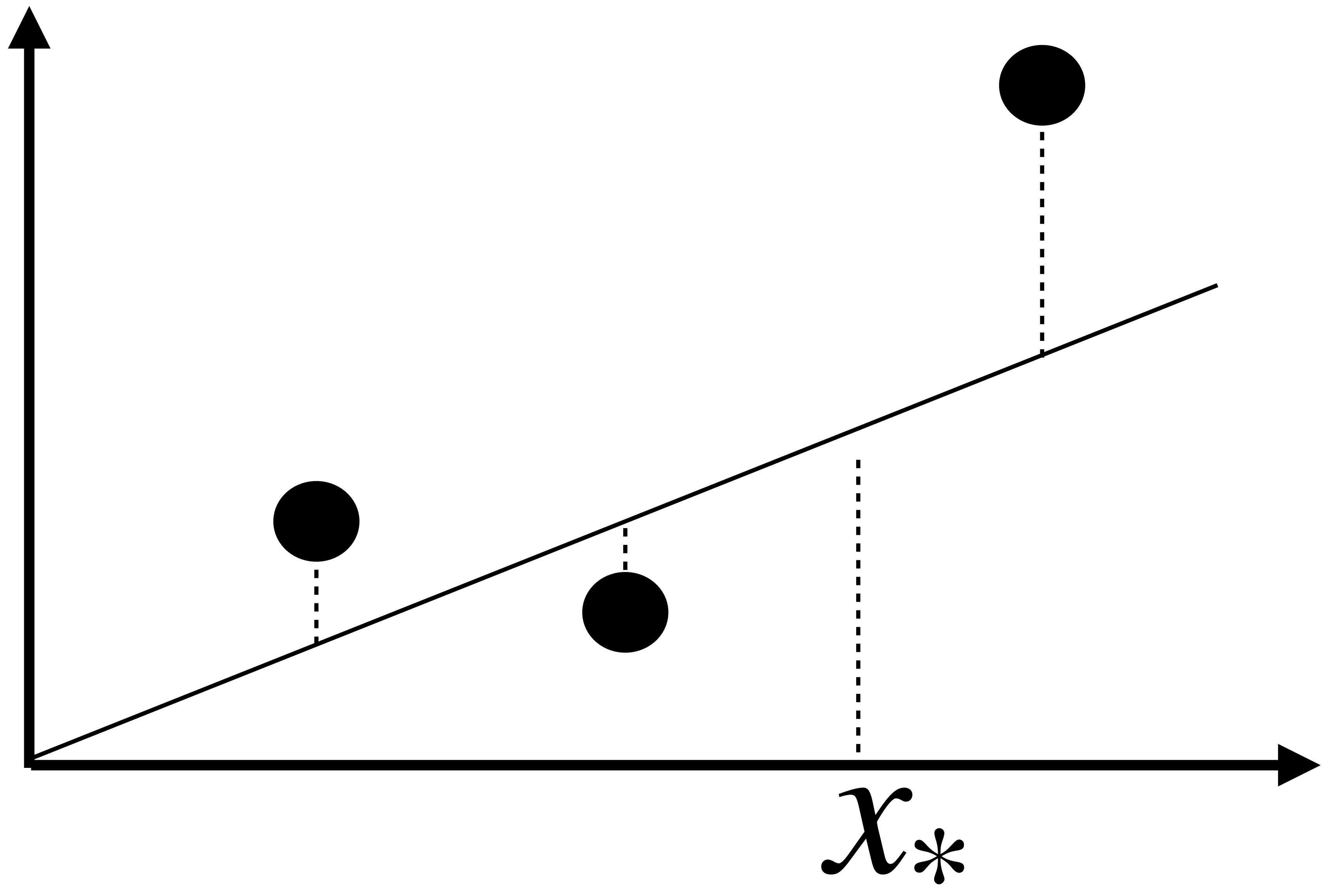
Recall what linear regression does!



$$y_{pred} = ax_* + b$$

- Assume the straight line function is y = ax + b
- Calculate the Loss = sum of all squared error
- Tune ‘a’ and ‘b’ so that the Loss is minimized!
- Obtain the optimal ‘a’ and ‘b’, and we can use it in prediction

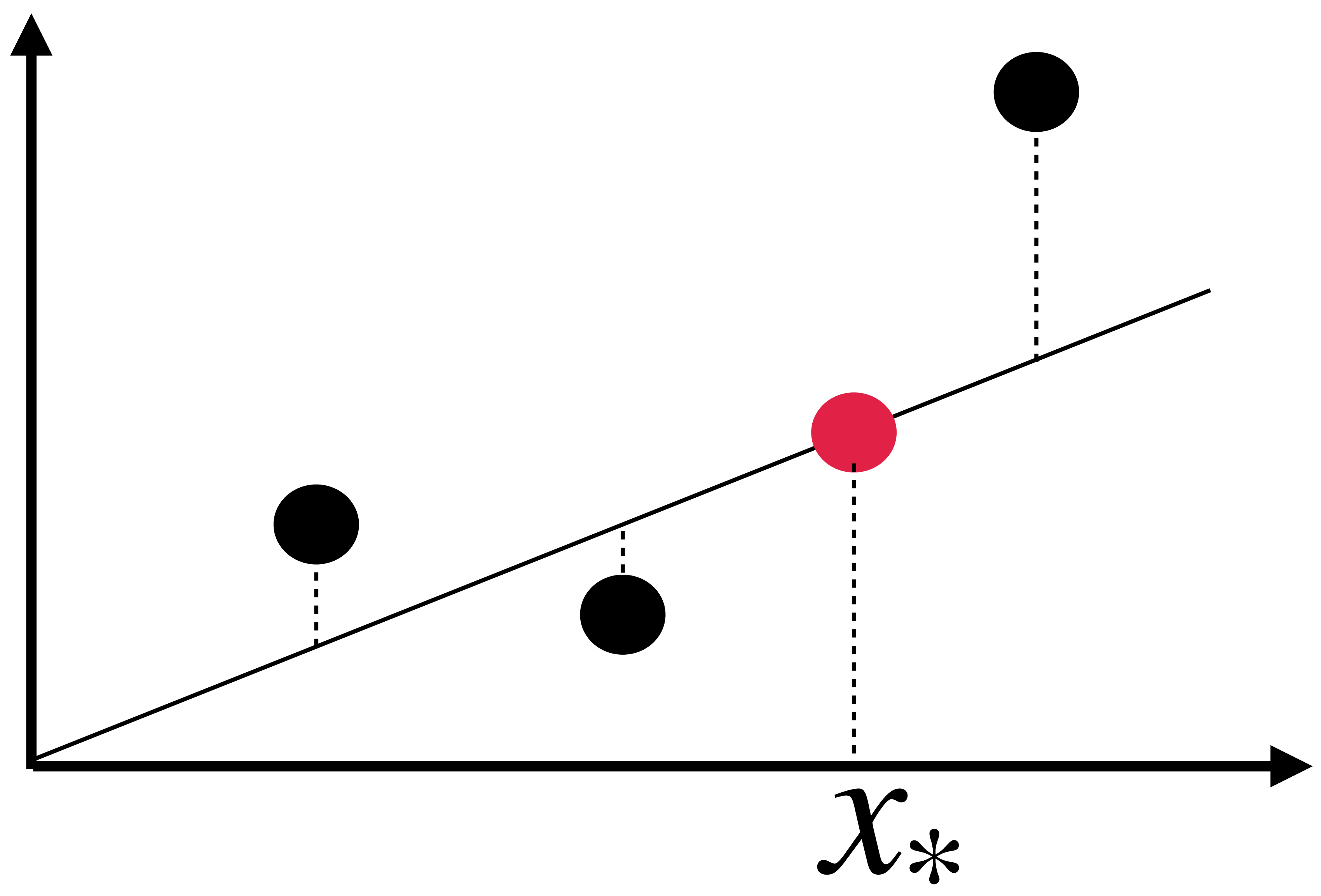
Recall what linear regression does!



$$y_{pred} = ax_* + b$$

- Assume the straight line function is y = ax + b
- Calculate the Loss = sum of all squared error
- Tune 'a' and 'b' so that the Loss is minimized!
- Obtain the optimal 'a' and 'b', and we can use it in prediction

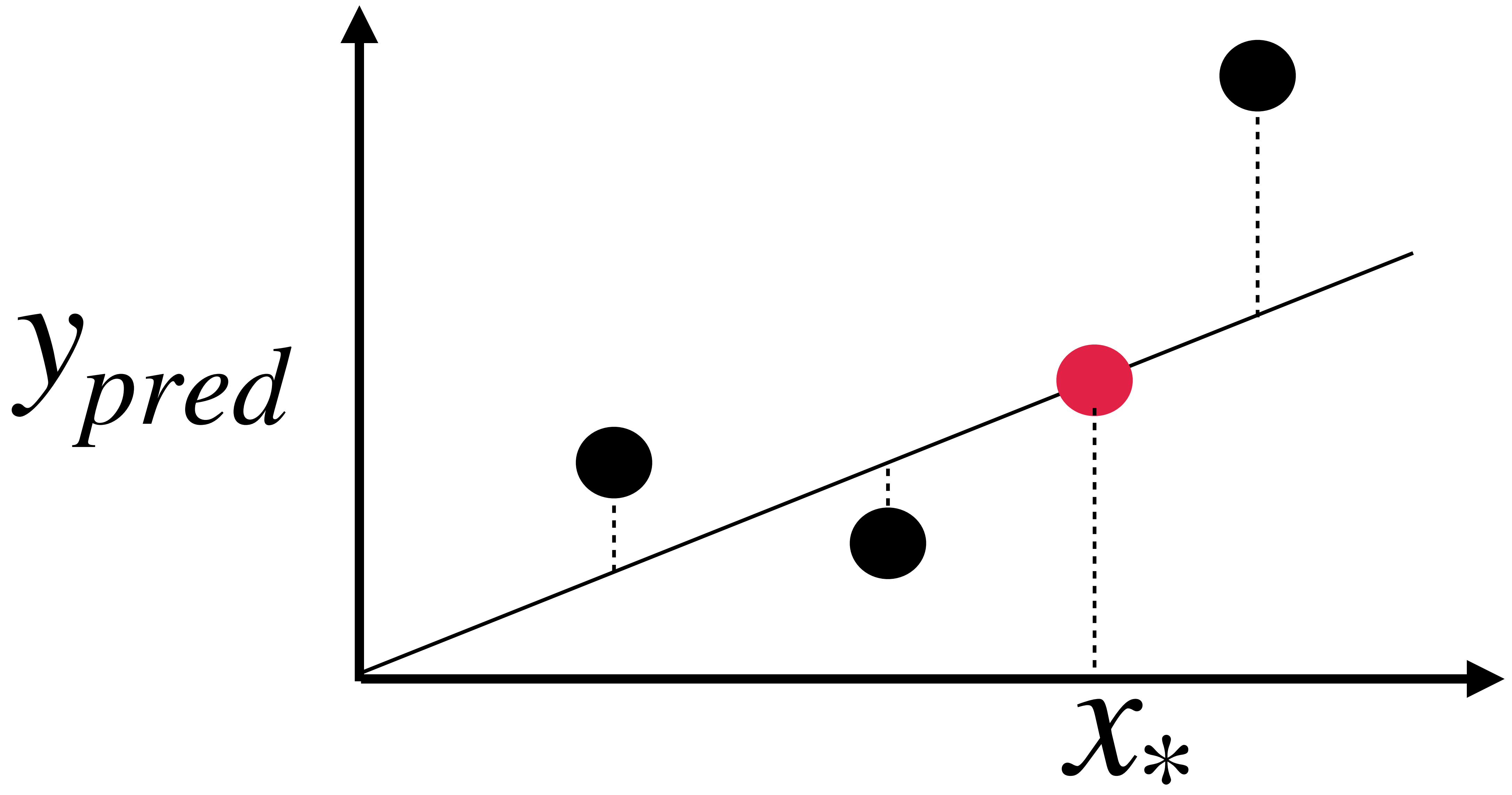
Recall what linear regression does!



$$y_{pred} = ax_* + b$$

- Assume the straight line function is y = ax + b
- Calculate the Loss = sum of all squared error
- Tune 'a' and 'b' so that the Loss is minimized!
- Obtain the optimal 'a' and 'b', and we can use it in prediction

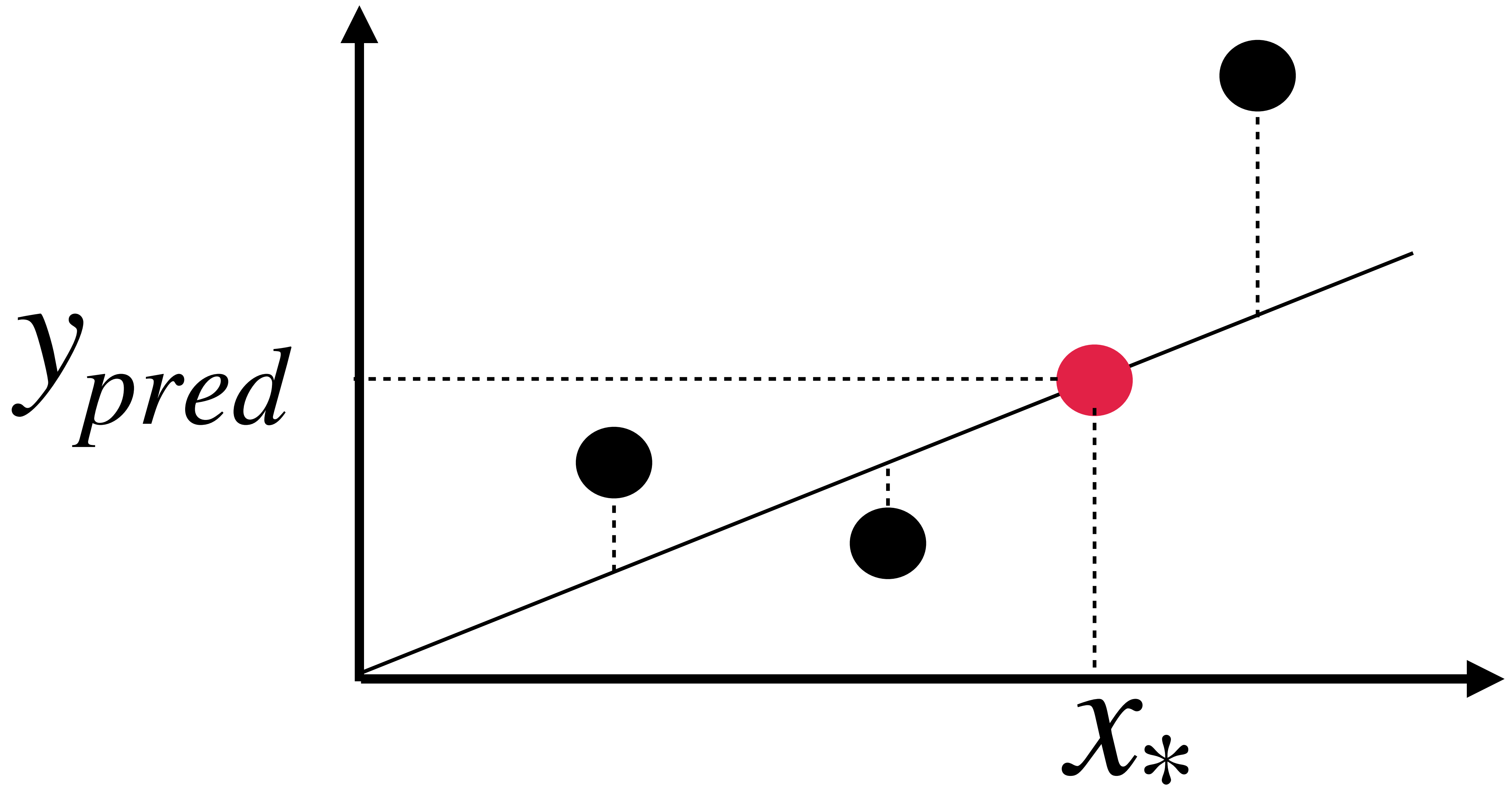
Recall what linear regression does!



$$y_{pred} = ax_* + b$$

- Assume the straight line function is y = ax + b
- Calculate the Loss = sum of all squared error
- Tune 'a' and 'b' so that the Loss is minimized!
- Obtain the optimal 'a' and 'b', and we can use it in prediction

Recall what linear regression does!



$$y_{pred} = ax_* + b$$

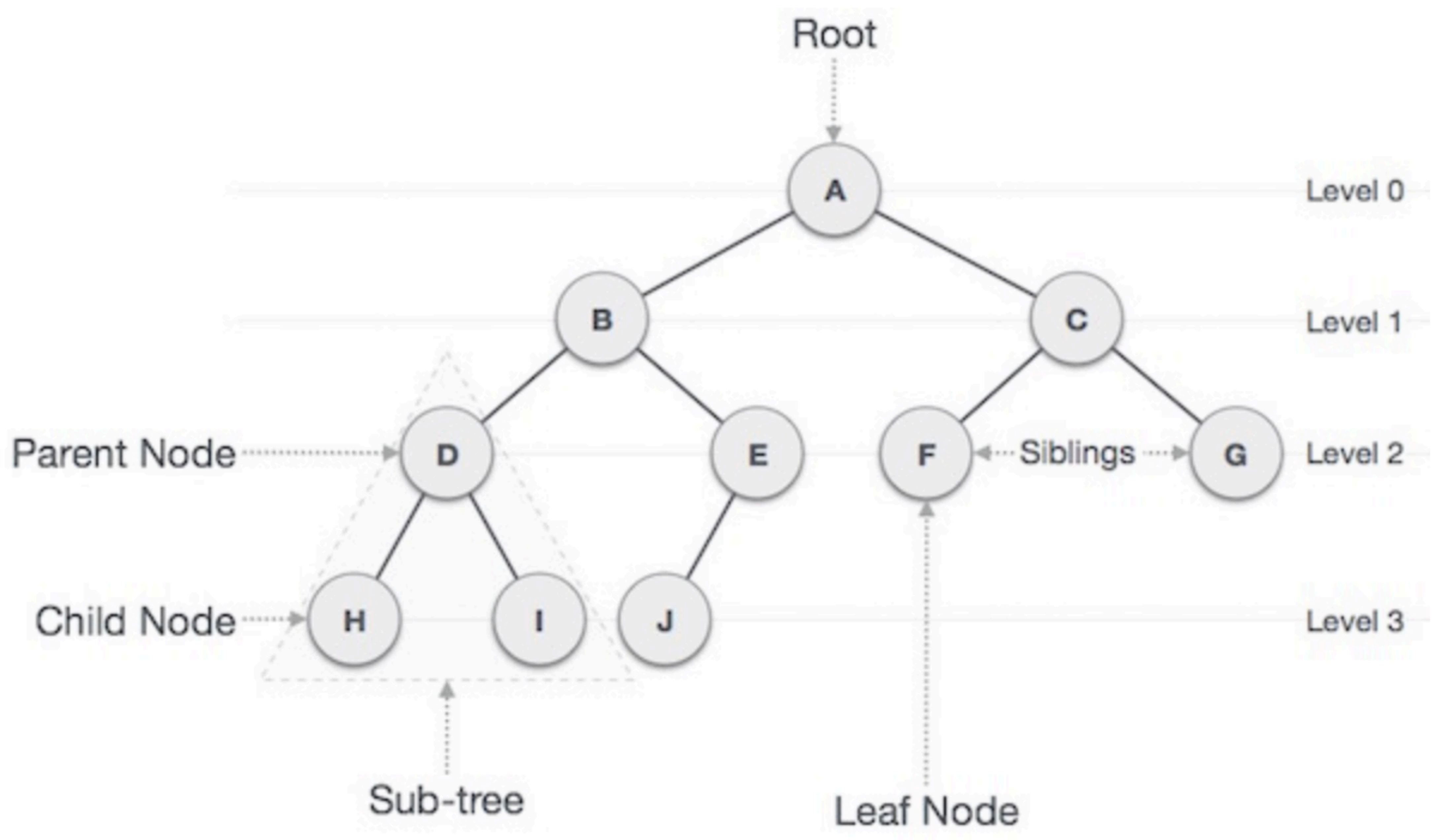
- Assume the straight line function is y = ax + b
- Calculate the Loss = sum of all squared error
- Tune ‘a’ and ‘b’ so that the Loss is minimized!
- Obtain the optimal ‘a’ and ‘b’, and we can use it in prediction

Why tree-based model?

Because it is very computer-sciencely!!

As a data structure, binary tree search, sort, etc...

An accompanying key phrase: divide and conquer



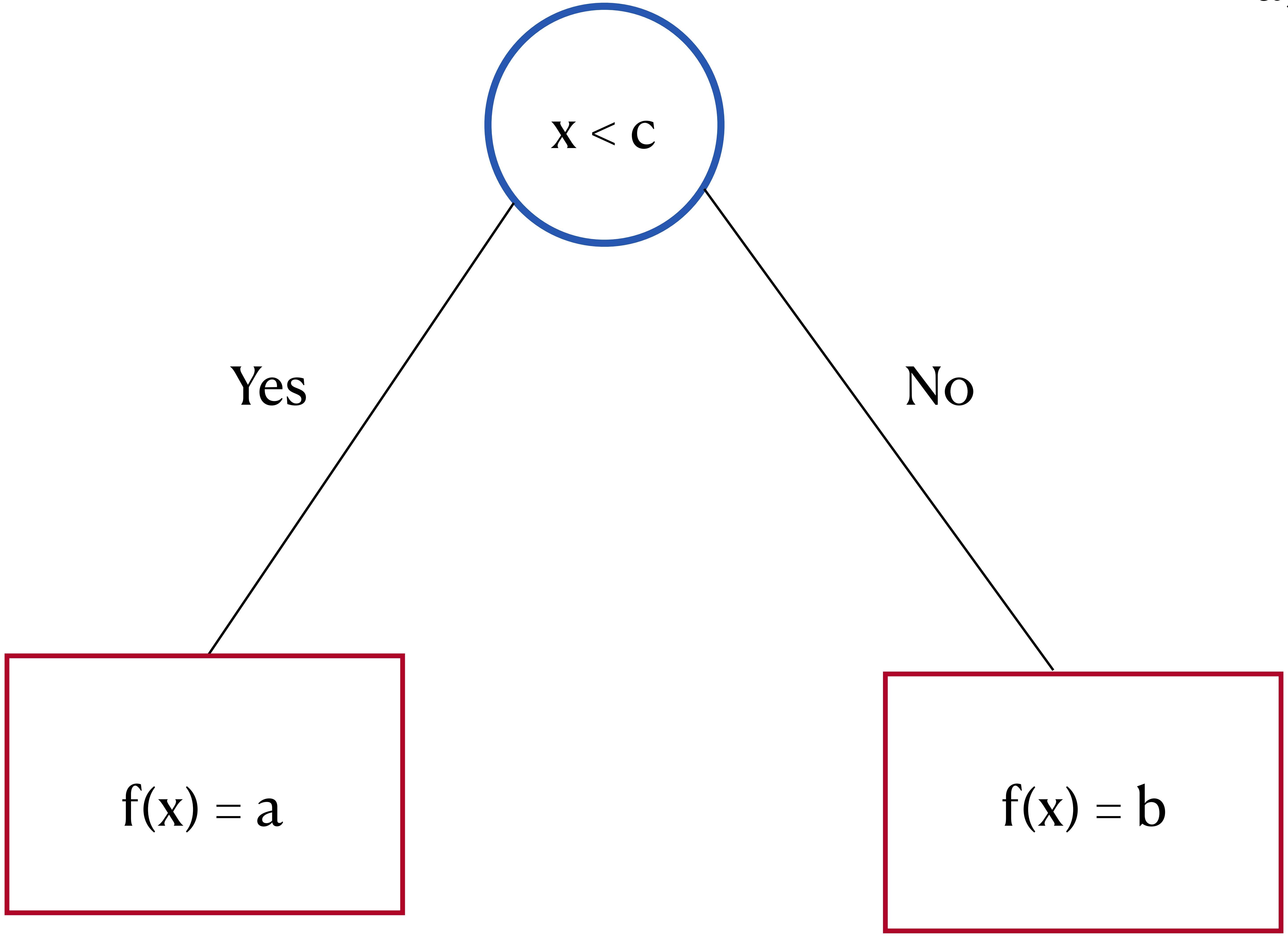
A simple numeric function of tree

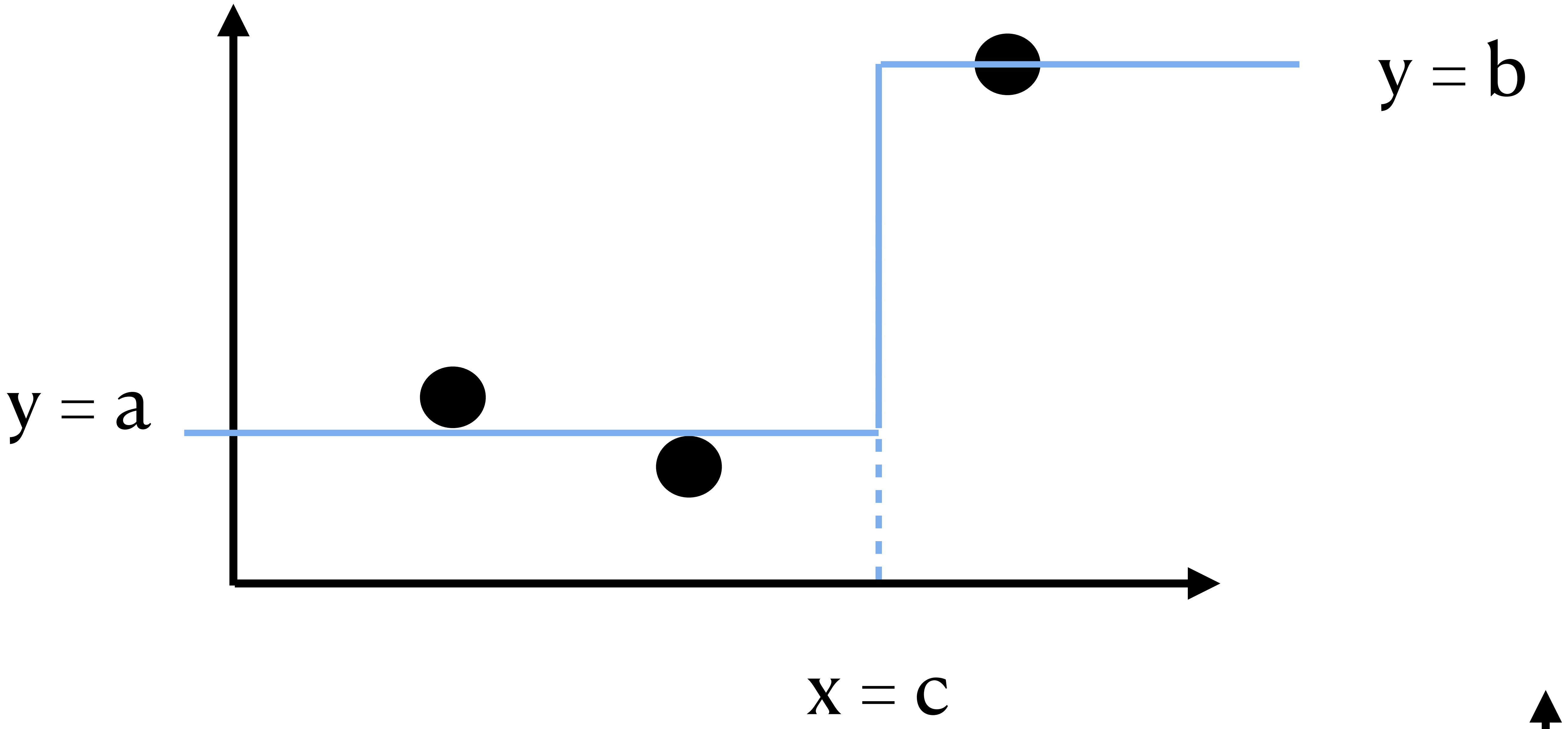
f(x):

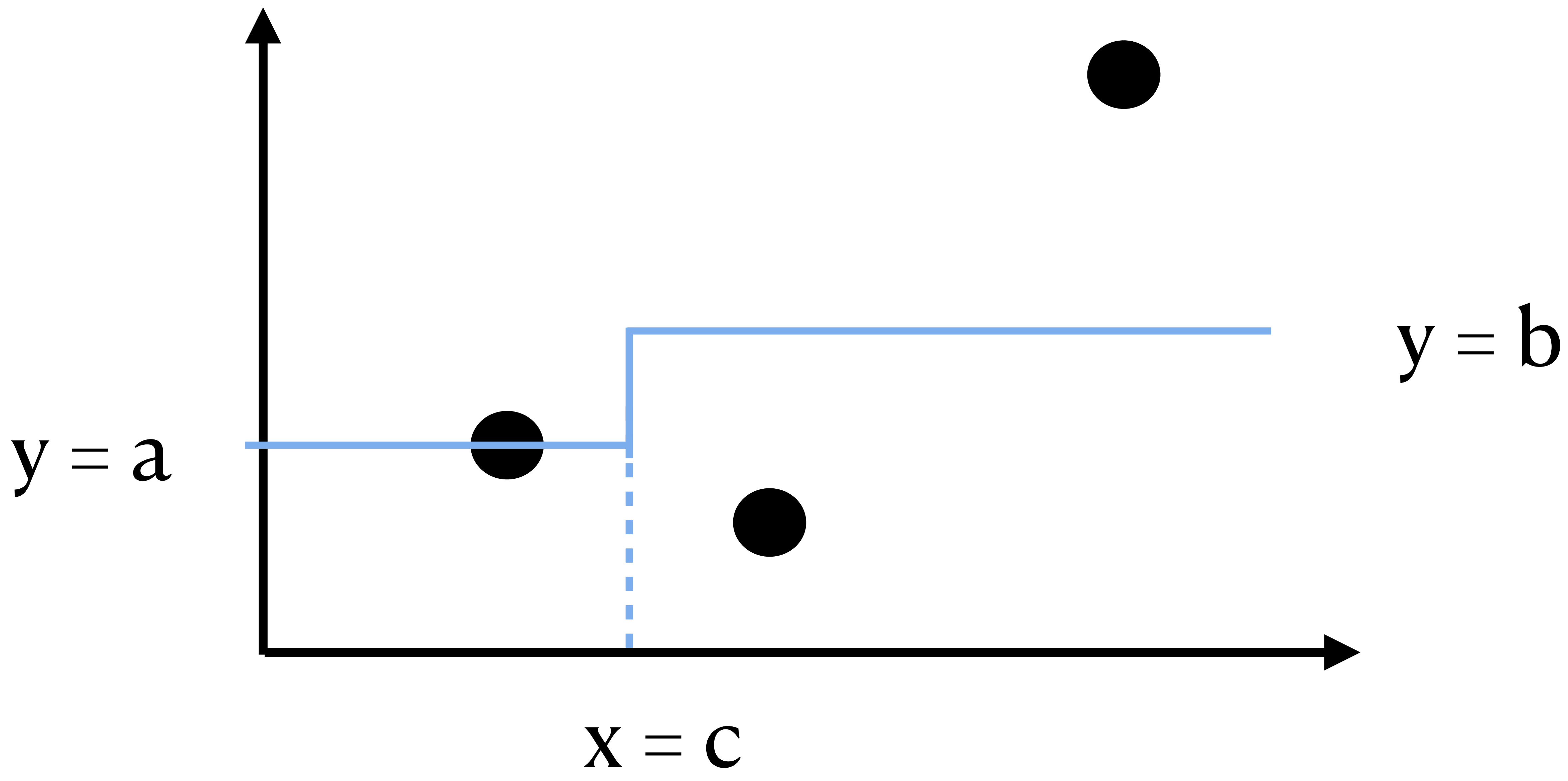
The model has

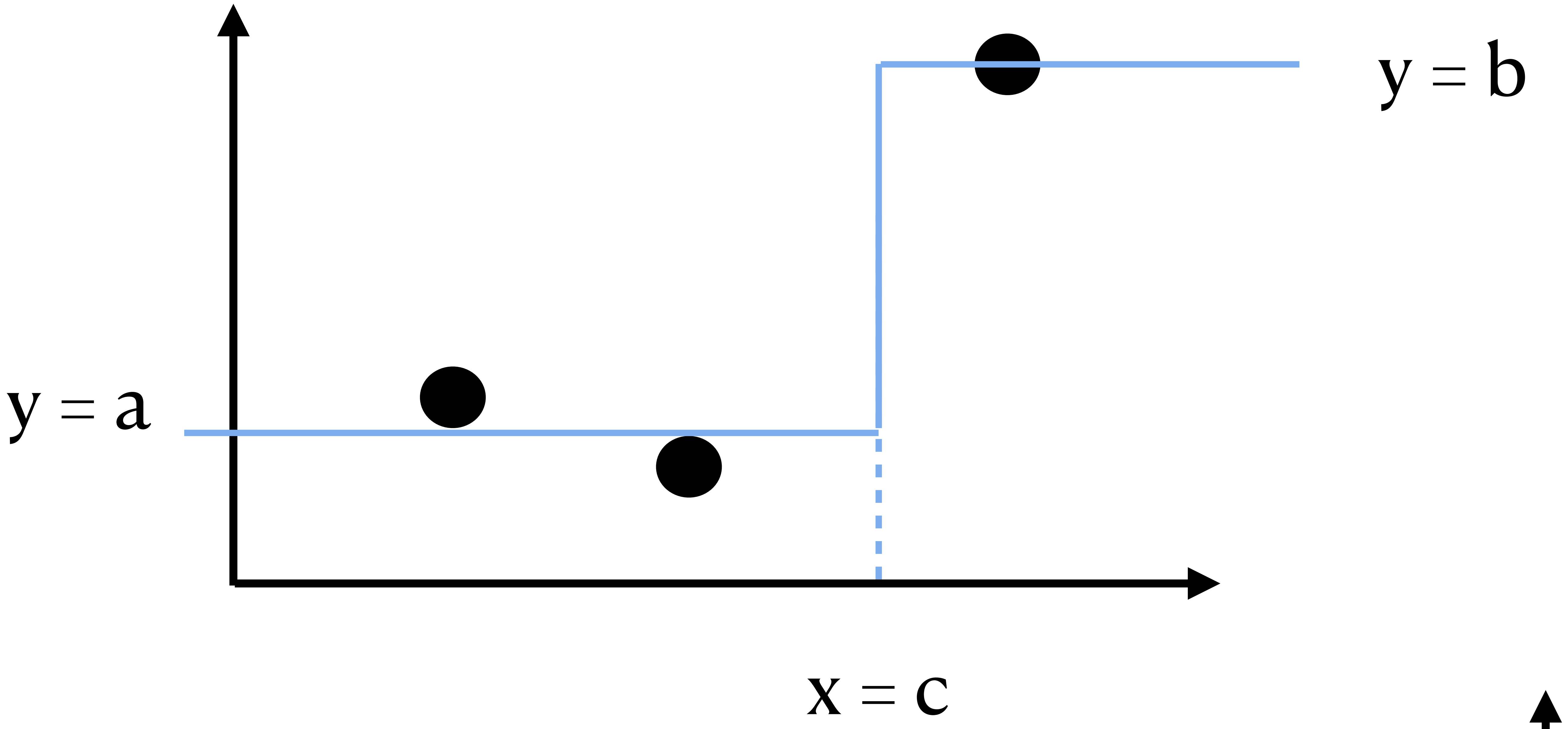
'3' parameters:

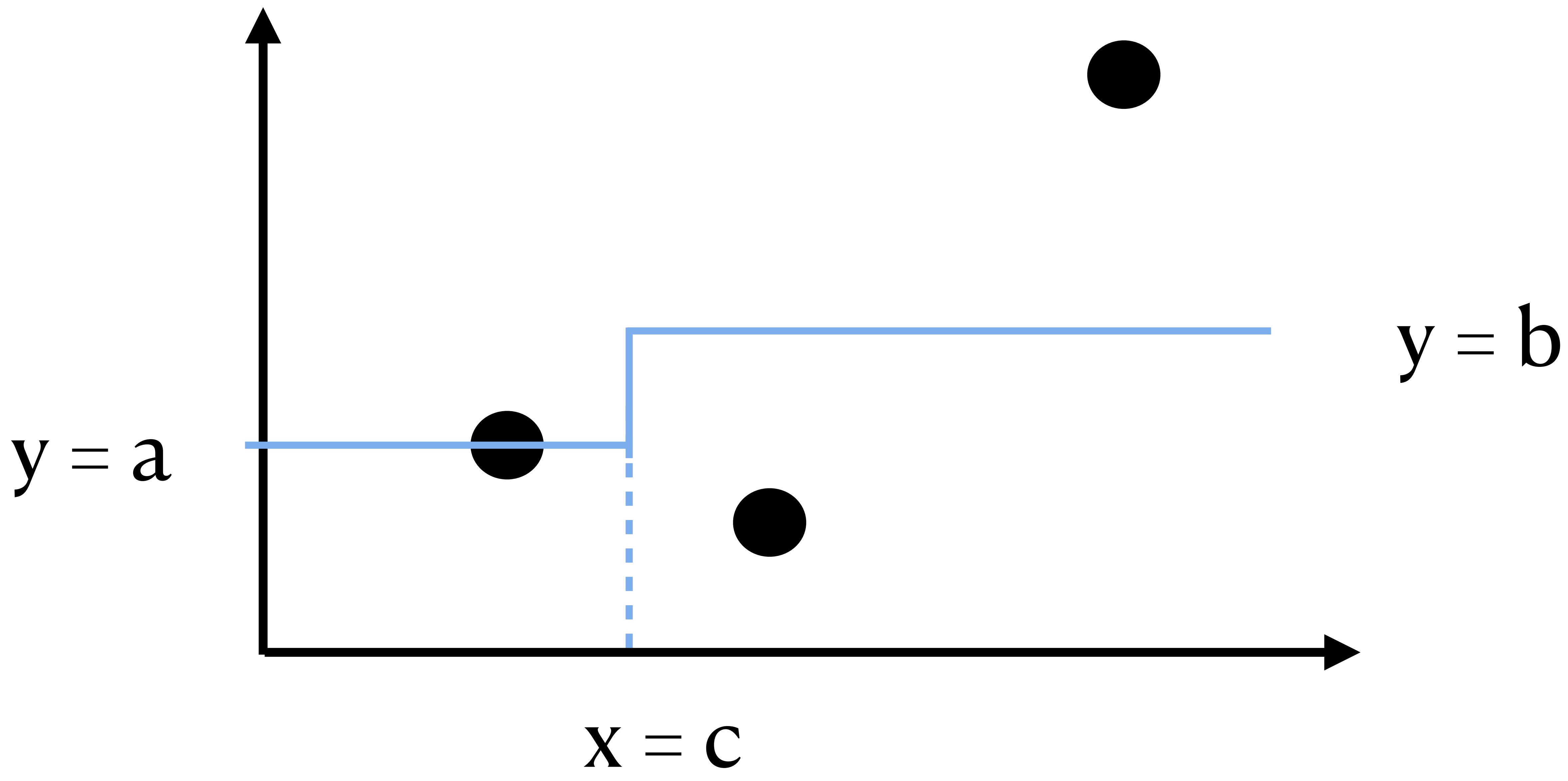
a, b, c




y = b
y = a
x = c


y = b
y = a
x = c





Which one has lower loss?

Baseball player salary data set

How player's salary is determined?

	Unnamed: 0	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
1	-Alan Ashby	315	81	7	24	38	39	14	3449	835	...	321	414	375	N	W	632	43	10	475.0
2	-Alvin Davis	479	130	18	66	72	76	3	1624	457	...	224	266	263	A	W	880	82	14	480.0
3	-Andre Dawson	496	141	20	65	78	37	11	5628	1575	...	828	838	354	N	E	200	11	3	500.0
4	-Andres Galarraga	321	87	10	39	42	30	2	396	101	...	48	46	33	N	E	805	40	4	91.5
5	-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	...	501	336	194	A	W	282	421	25	750.0

Baseball player salary data set

How player's salary is determined?

	Unnamed: 0	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
1	-Alan Ashby	315	81	7	24	38	39	14	3449	835	...	321	414	375	N	W	632	43	10	475.0
2	-Alvin Davis	479	130	18	66	72	76	3	1624	457	...	224	266	263	A	W	880	82	14	480.0
3	-Andre Dawson	496	141	20	65	78	37	11	5628	1575	...	828	838	354	N	E	200	11	3	500.0
4	-Andres Galarraga	321	87	10	39	42	30	2	396	101	...	48	46	33	N	E	805	40	4	91.5
5	-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	...	501	336	194	A	W	282	421	25	750.0

Baseball player salary data set

How player's salary is determined?

Y

	Unnamed: 0	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
1	-Alan Ashby	315	81	7	24	38	39	14	3449	835	...	321	414	375	N	W	632	43	10	475.0
2	-Alvin Davis	479	130	18	66	72	76	3	1624	457	...	224	266	263	A	W	880	82	14	480.0
3	-Andre Dawson	496	141	20	65	78	37	11	5628	1575	...	828	838	354	N	E	200	11	3	500.0
4	-Andres Galarraga	321	87	10	39	42	30	2	396	101	...	48	46	33	N	E	805	40	4	91.5
5	-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	...	501	336	194	A	W	282	421	25	750.0

Baseball player salary data set

How player's salary is determined?

Y

	Unnamed: 0	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
1	-Alan Ashby	315	81	7	24	38	39	14	3449	835	...	321	414	375	N	W	632	43	10	475.0
2	-Alvin Davis	479	130	18	66	72	76	3	1624	457	...	224	266	263	A	W	880	82	14	480.0
3	-Andre Dawson	496	141	20	65	78	37	11	5628	1575	...	828	838	354	N	E	200	11	3	500.0
4	-Andres Galarraga	321	87	10	39	42	30	2	396	101	...	48	46	33	N	E	805	40	4	91.5
5	-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	...	501	336	194	A	W	282	421	25	750.0

Baseball player salary data set

How player's salary is determined?

X_1

Y

	Unnamed: 0	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
1	-Alan Ashby	315	81	7	24	38	39	14	3449	835	...	321	414	375	N	W	632	43	10	475.0
2	-Alvin Davis	479	130	18	66	72	76	3	1624	457	...	224	266	263	A	W	880	82	14	480.0
3	-Andre Dawson	496	141	20	65	78	37	11	5628	1575	...	828	838	354	N	E	200	11	3	500.0
4	-Andres Galarraga	321	87	10	39	42	30	2	396	101	...	48	46	33	N	E	805	40	4	91.5
5	-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	...	501	336	194	A	W	282	421	25	750.0

Baseball player salary data set

How player's salary is determined?

Y

X_1

Unnamed: 0		AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
1	-Alan Ashby	315	81	7	24	38	39	14	3449	835	...	321	414	375	N	W	632	43	10	475.0
2	-Alvin Davis	479	130	18	66	72	76	3	1624	457	...	224	266	263	A	W	880	82	14	480.0
3	-Andre Dawson	496	141	20	65	78	37	11	5628	1575	...	828	838	354	N	E	200	11	3	500.0
4	-Andres Galarraga	321	87	10	39	42	30	2	396	101	...	48	46	33	N	E	805	40	4	91.5
5	-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	...	501	336	194	A	W	282	421	25	750.0

Baseball player salary data set

How player’s salary is determined?

Y

X_2

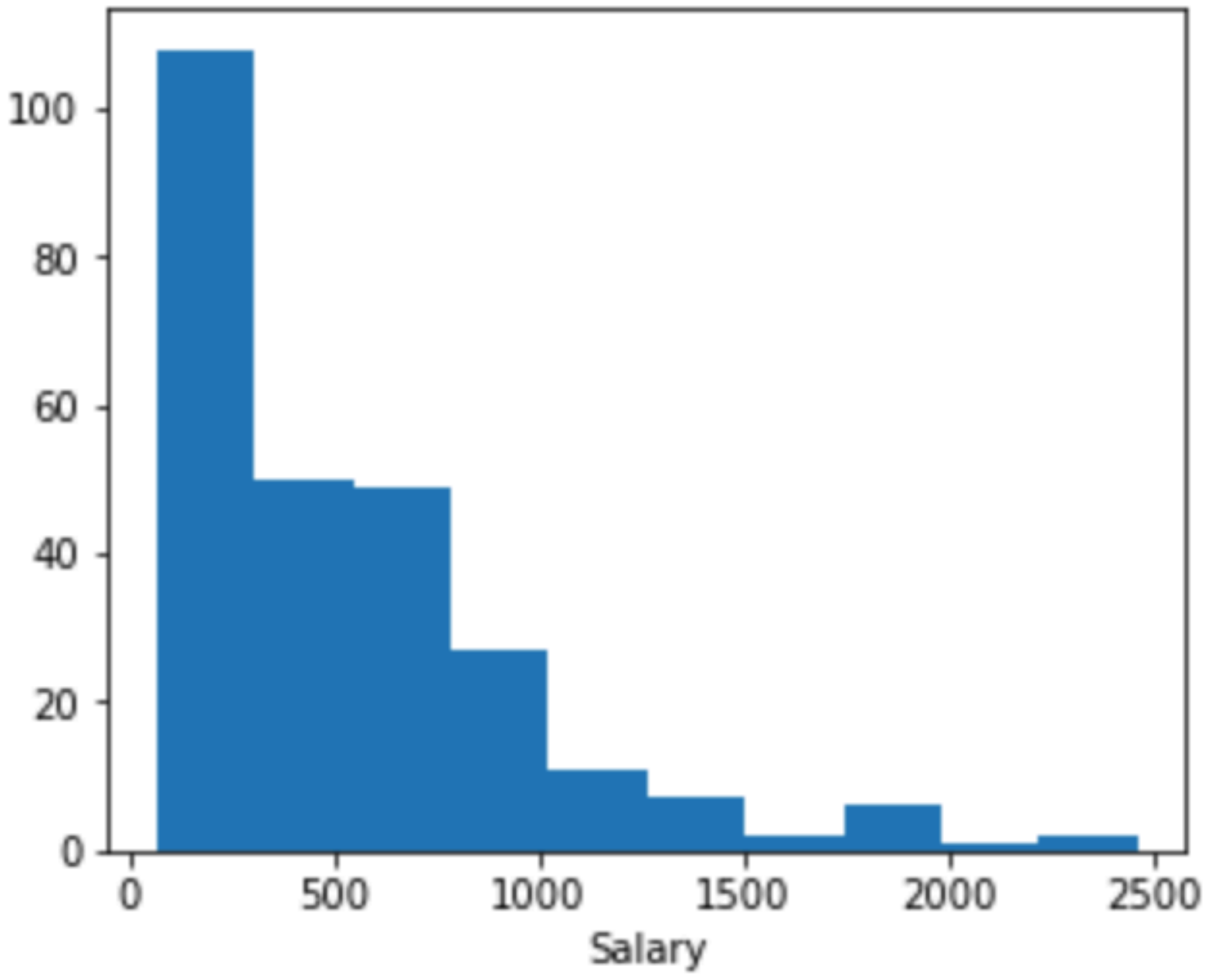
X_1

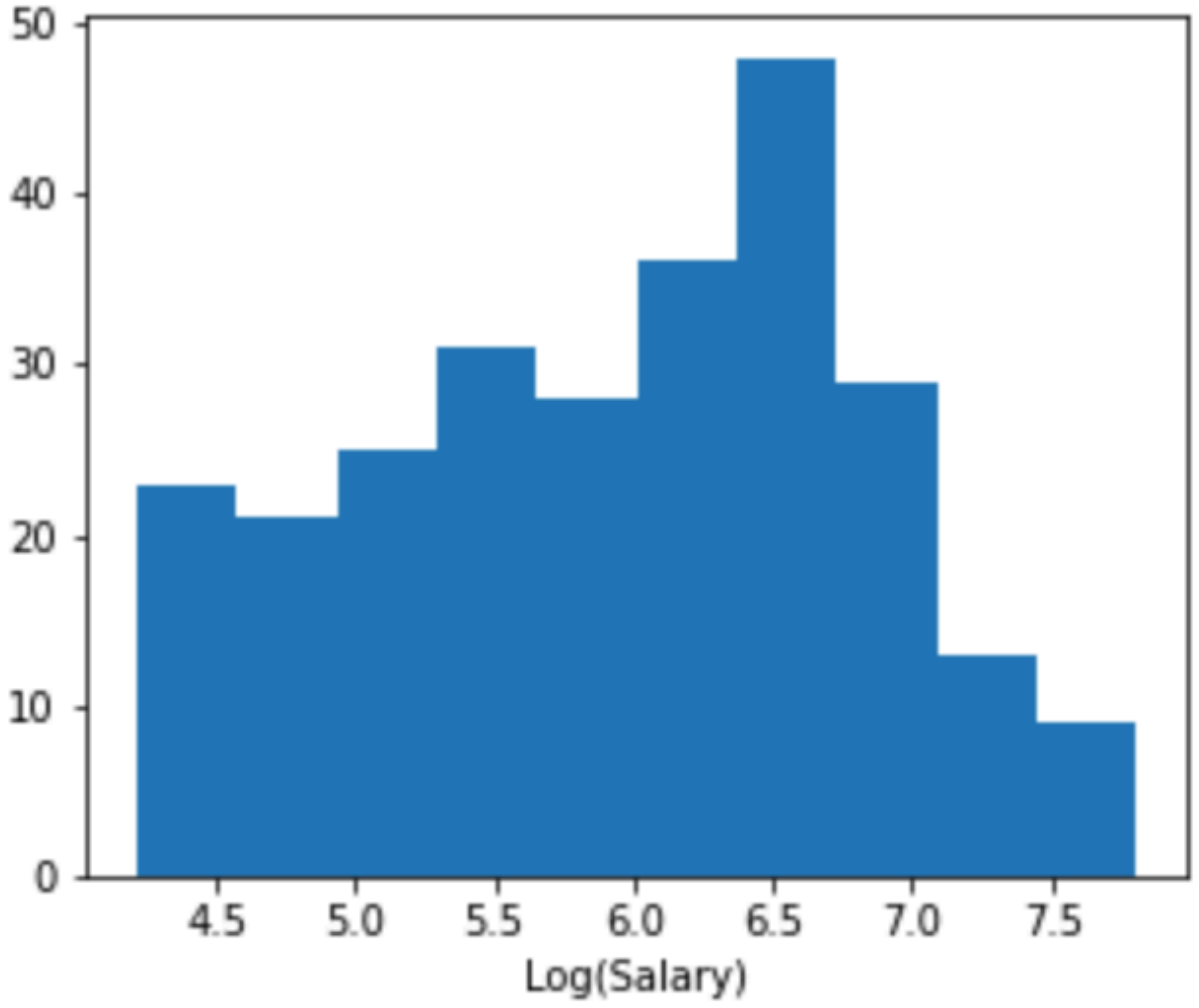
	Unnamed: 0	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	...	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary
1	-Alan Ashby	315	81	7	24	38	39	14	3449	835	...	321	414	375	N	W	632	43	10	475.0
2	-Alvin Davis	479	130	18	66	72	76	3	1624	457	...	224	266	263	A	W	880	82	14	480.0
3	-Andre Dawson	496	141	20	65	78	37	11	5628	1575	...	828	838	354	N	E	200	11	3	500.0
4	-Andres Galarraga	321	87	10	39	42	30	2	396	101	...	48	46	33	N	E	805	40	4	91.5
5	-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	...	501	336	194	A	W	282	421	25	750.0

Pre-processing the data

Taking log of salary

$$salary \rightarrow \log salary$$





Tree-based model for Hitters data

Tree-based model for Hitters data

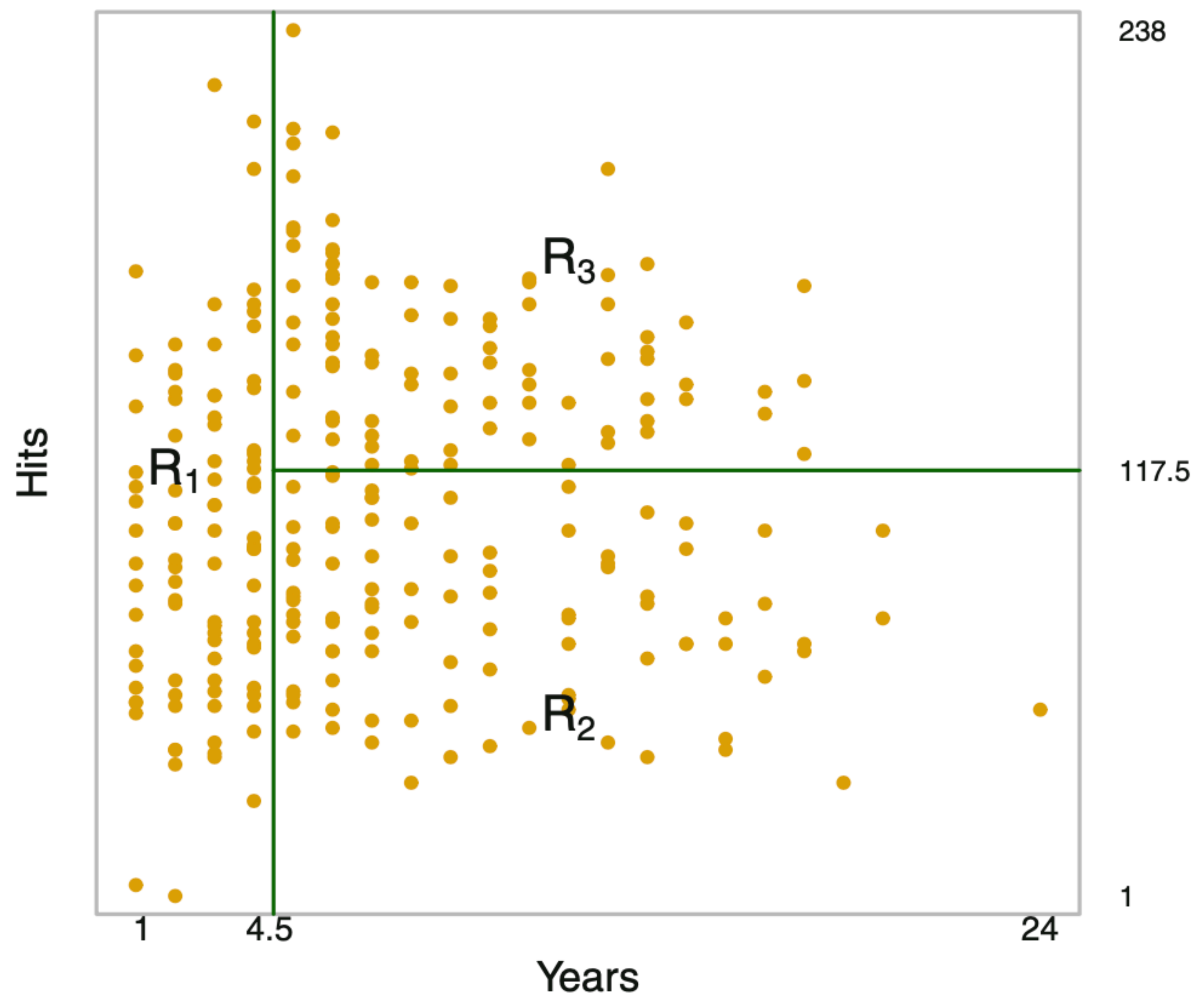


FIGURE 8.2. *The three-region partition for the* Hitters *data set from the regression tree illustrated in Figure 8.1.*

Tree-based model for Hitters data

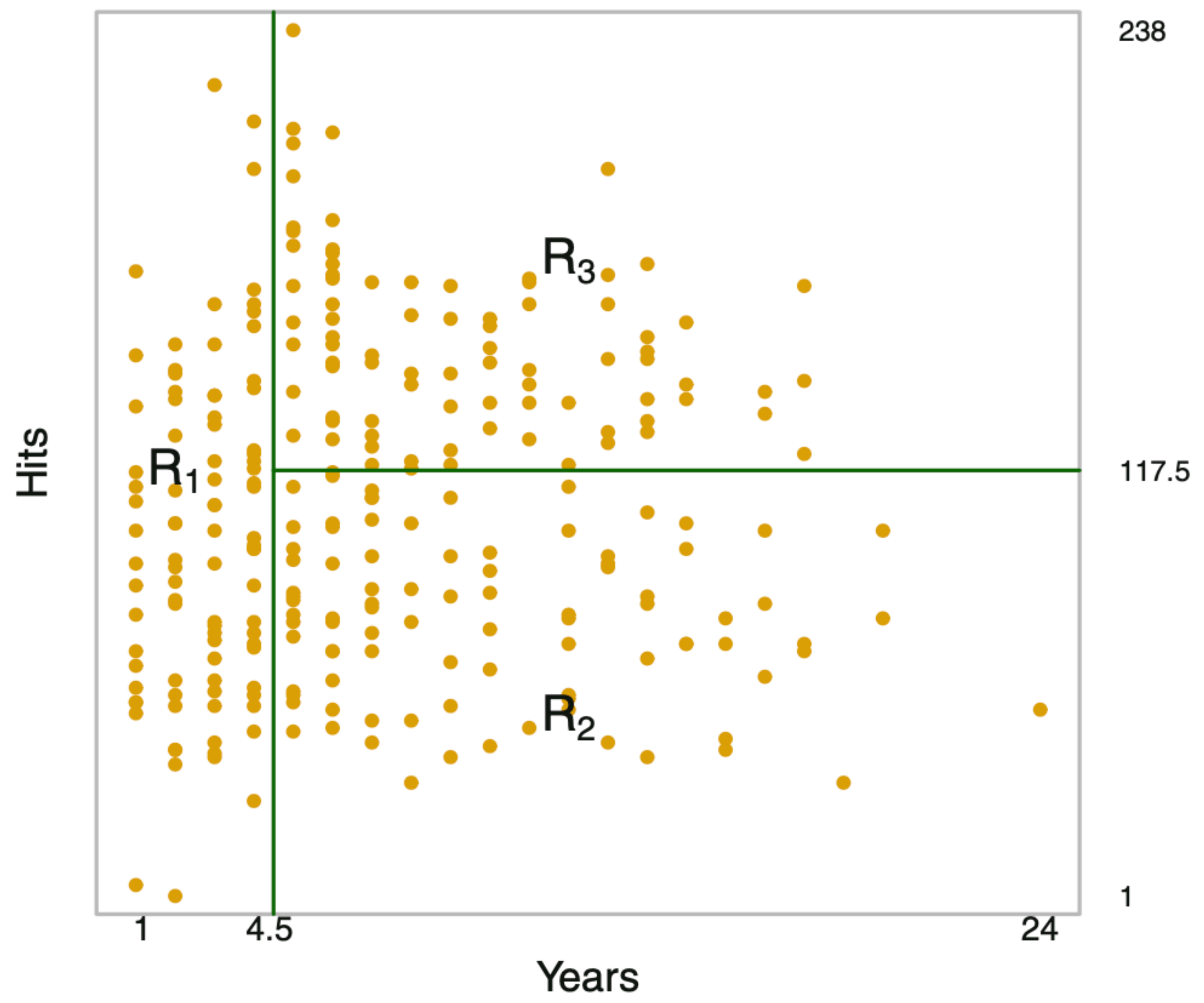
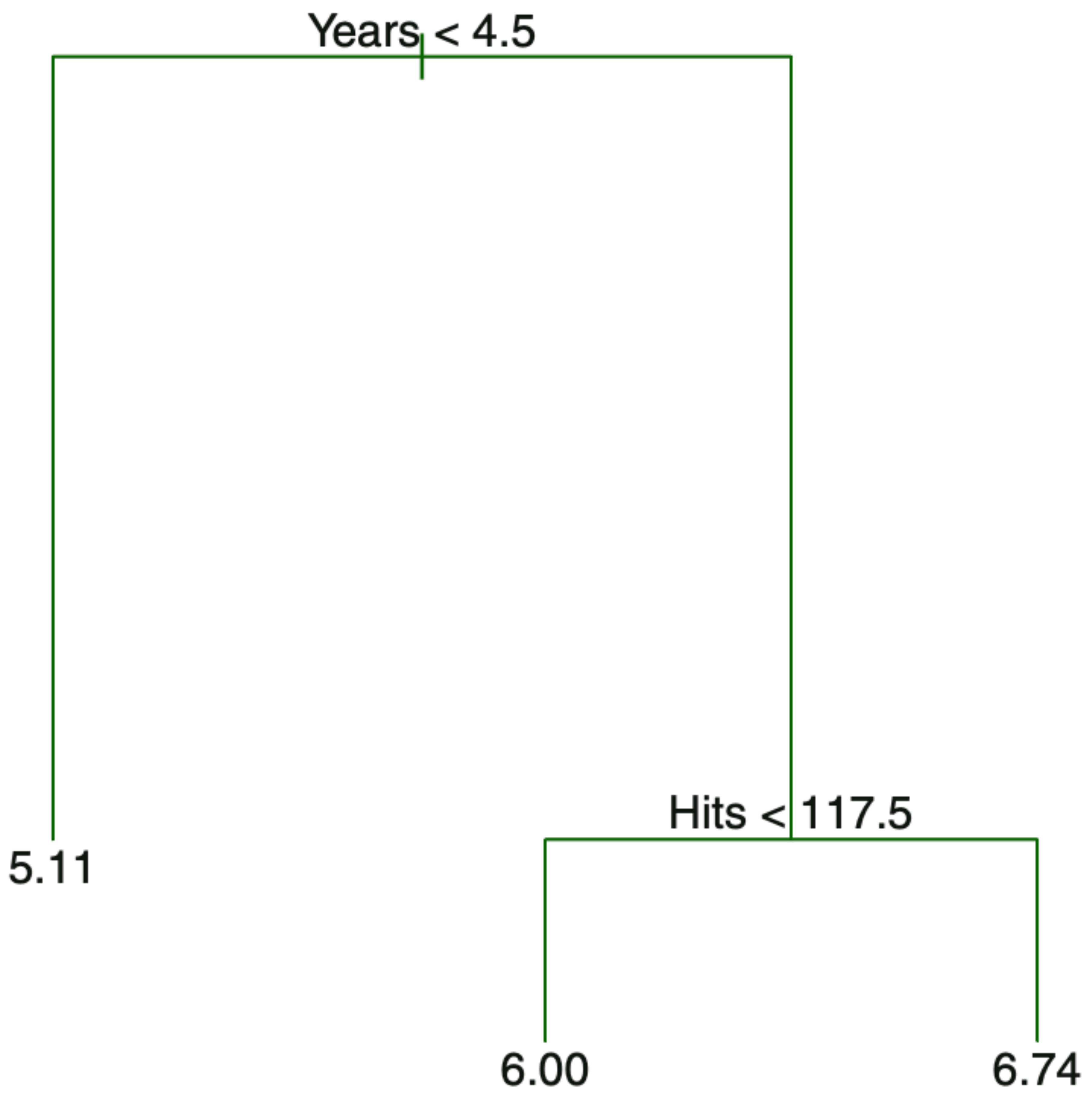


FIGURE 8.2. *The three-region partition for the* Hitters *data set from the regression tree illustrated in Figure 8.1.*

$$R_1 = \{X \,|\, Years < 4.5\}$$

Tree-based model for Hitters data



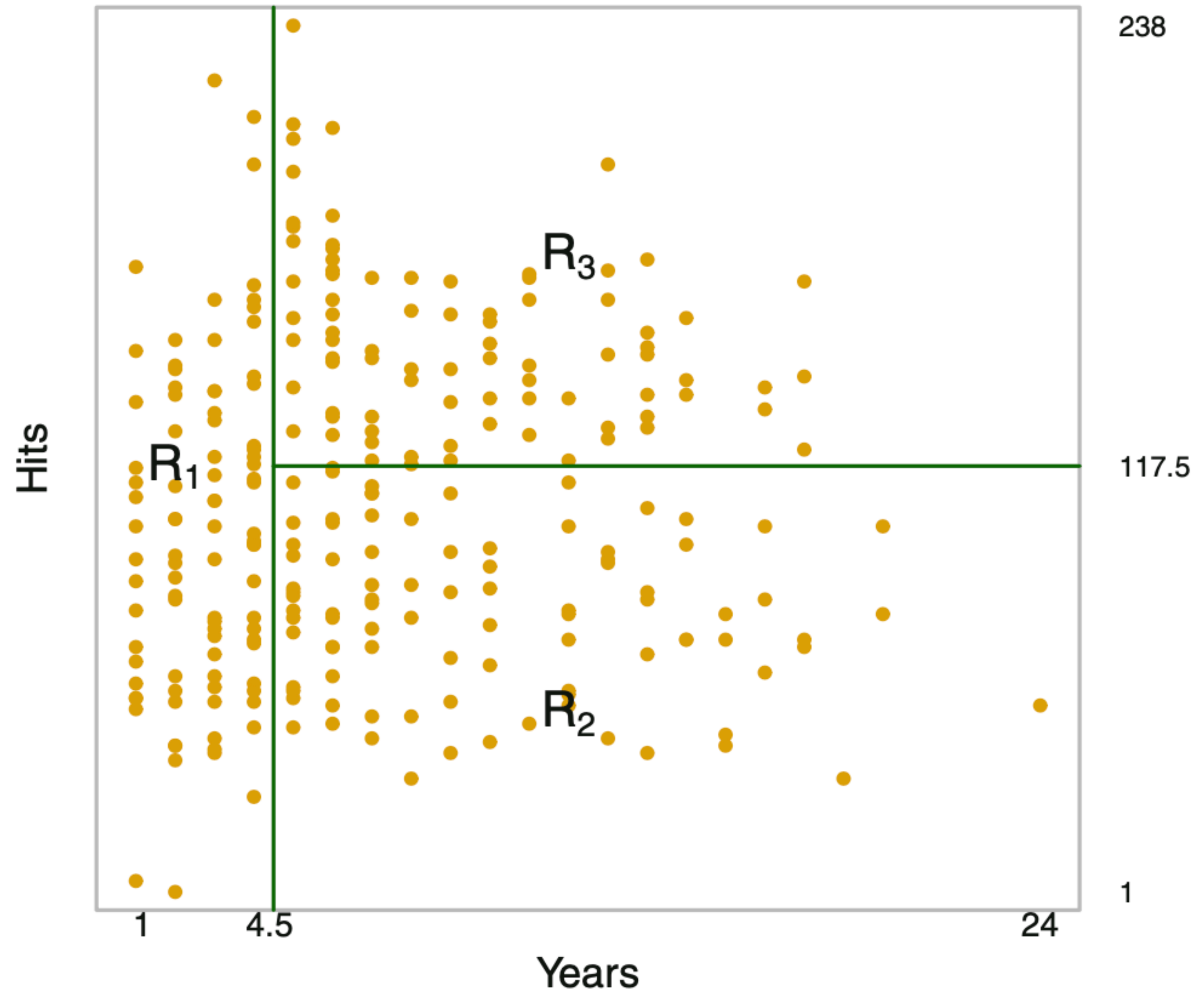


FIGURE 8.2. *The three-region partition for the* `Hitters` *data set from the regression tree illustrated in Figure 8.1.*

$$R_1 = \{X \mid Years < 4.5\}$$

$$R_2 = \{X \mid Years > 4.5,\ Hitts < 117.5\}$$

Tree-based model for Hitters data

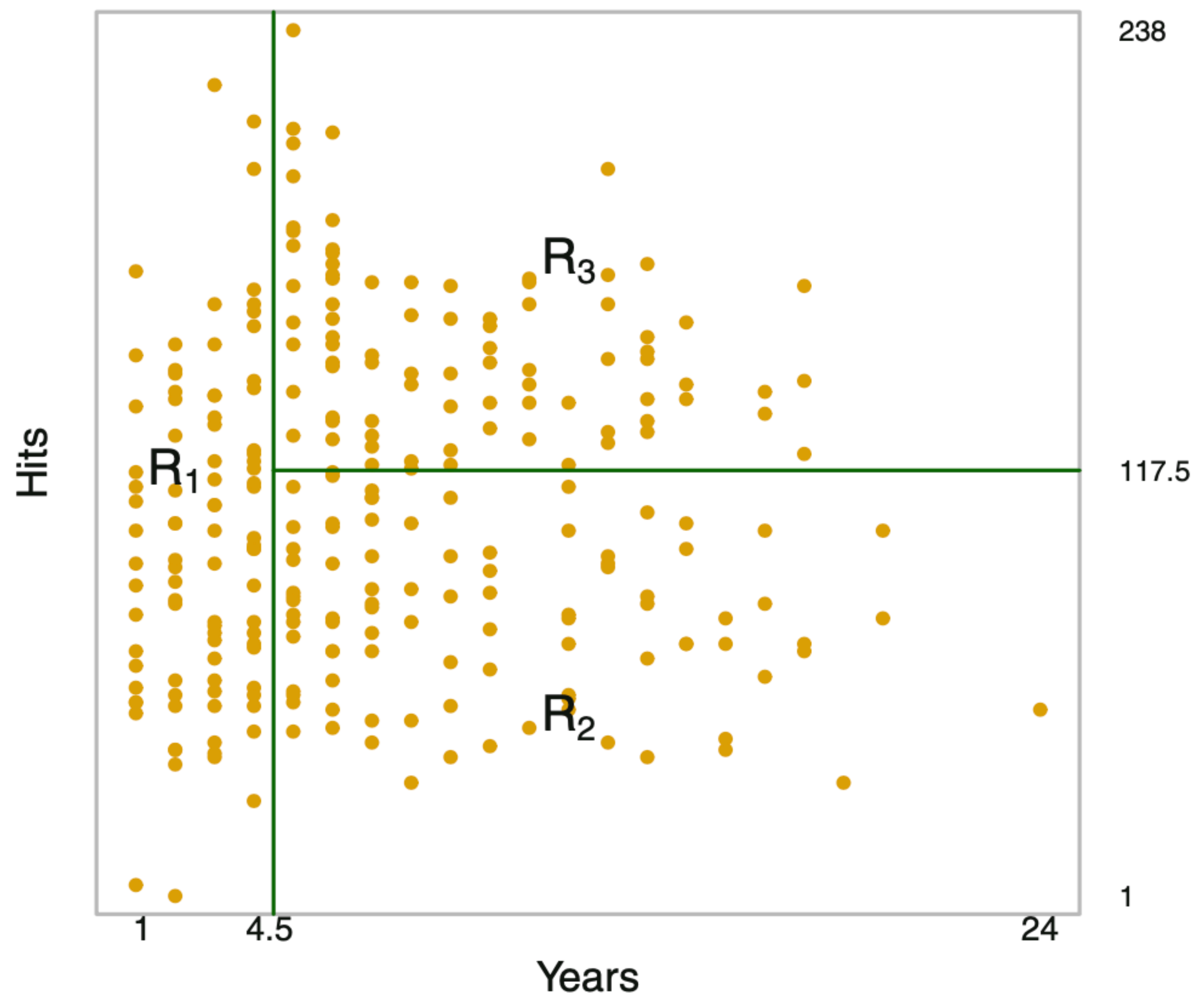


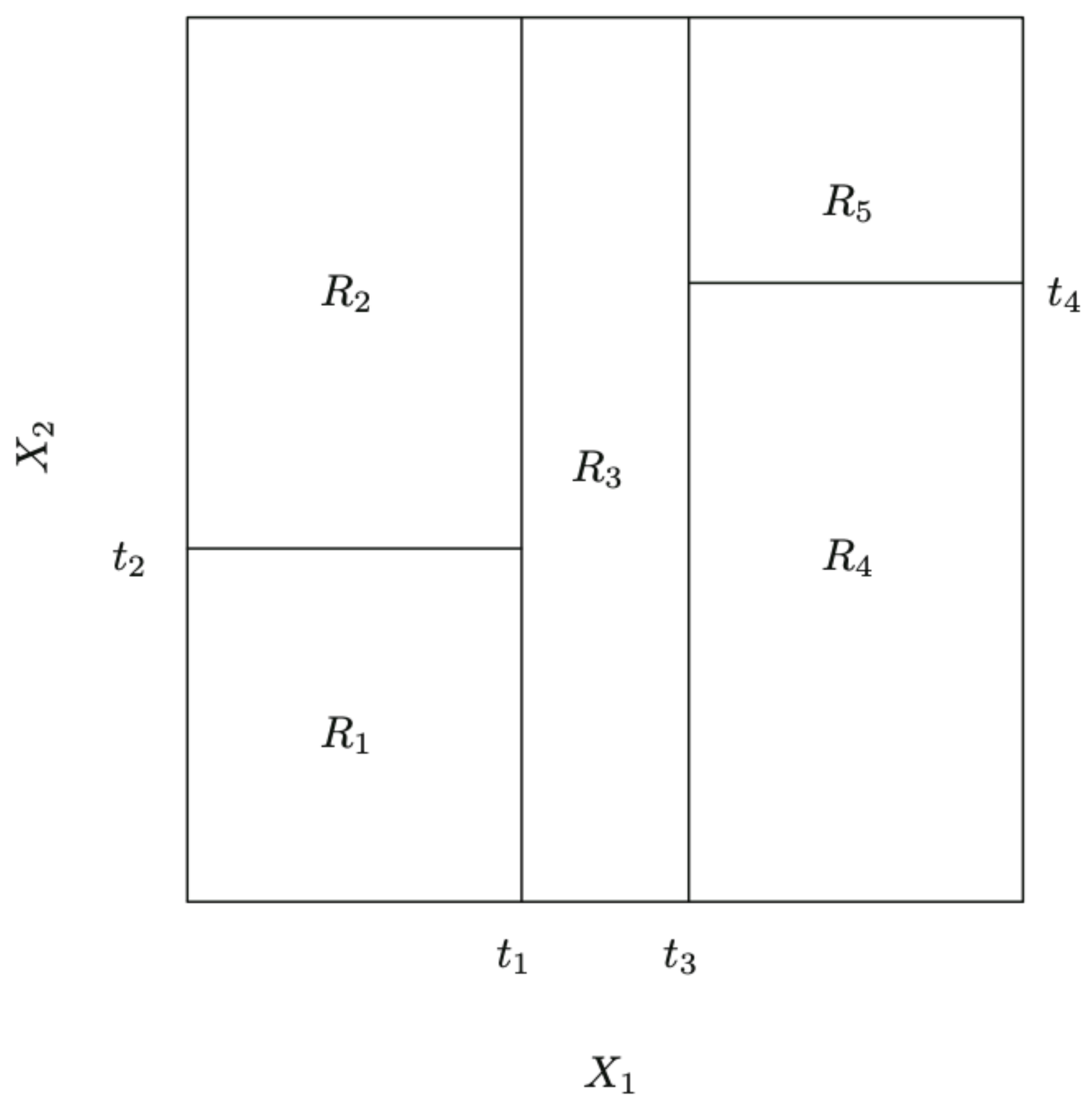
FIGURE 8.2. *The three-region partition for the* Hitters *data set from the regression tree illustrated in Figure 8.1.*

$$R_1 = \{X \mid Years < 4.5\}$$

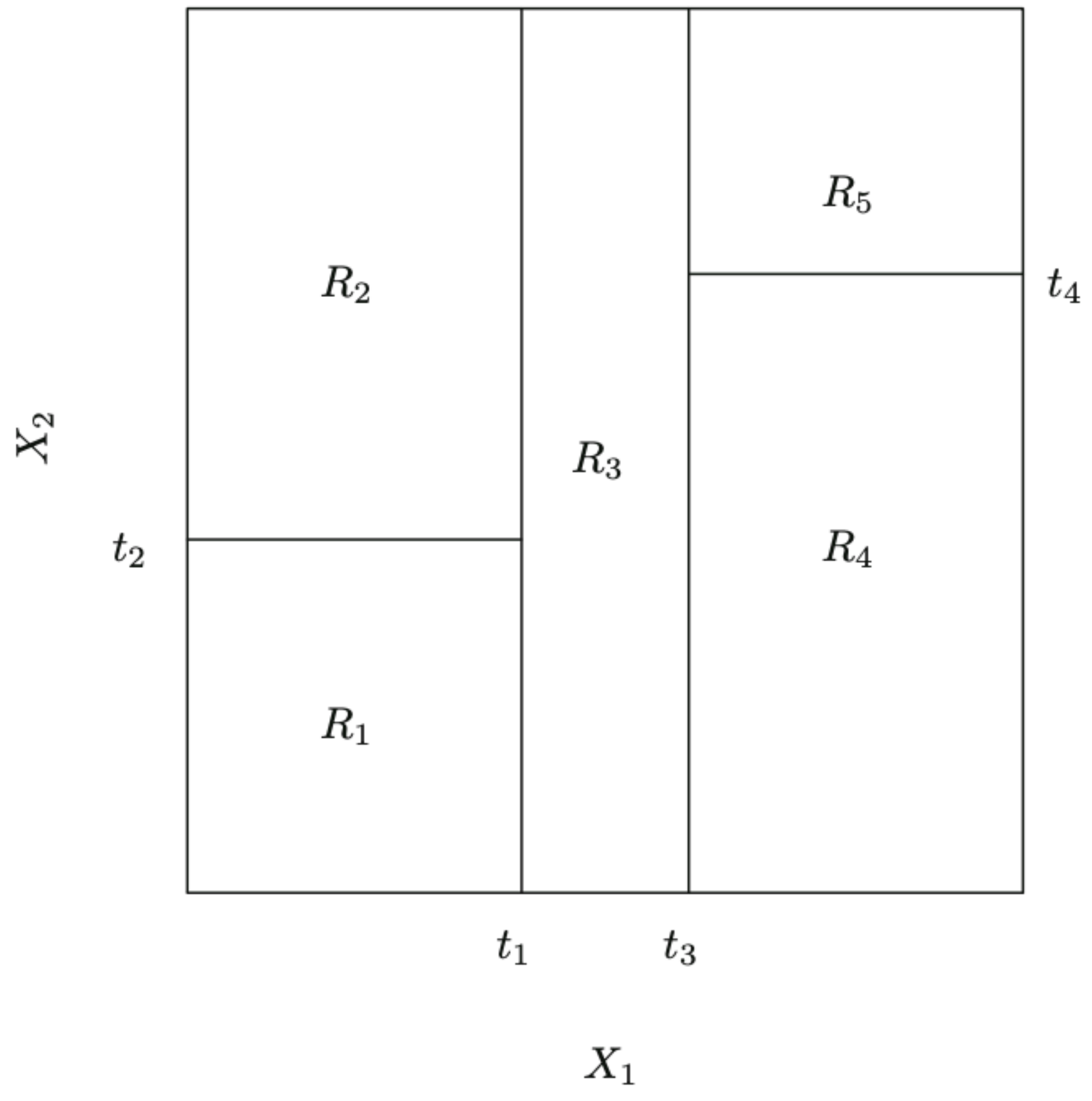
$$R_2 = \{X \mid Years > 4.5,\ Hitts < 117.5\}$$

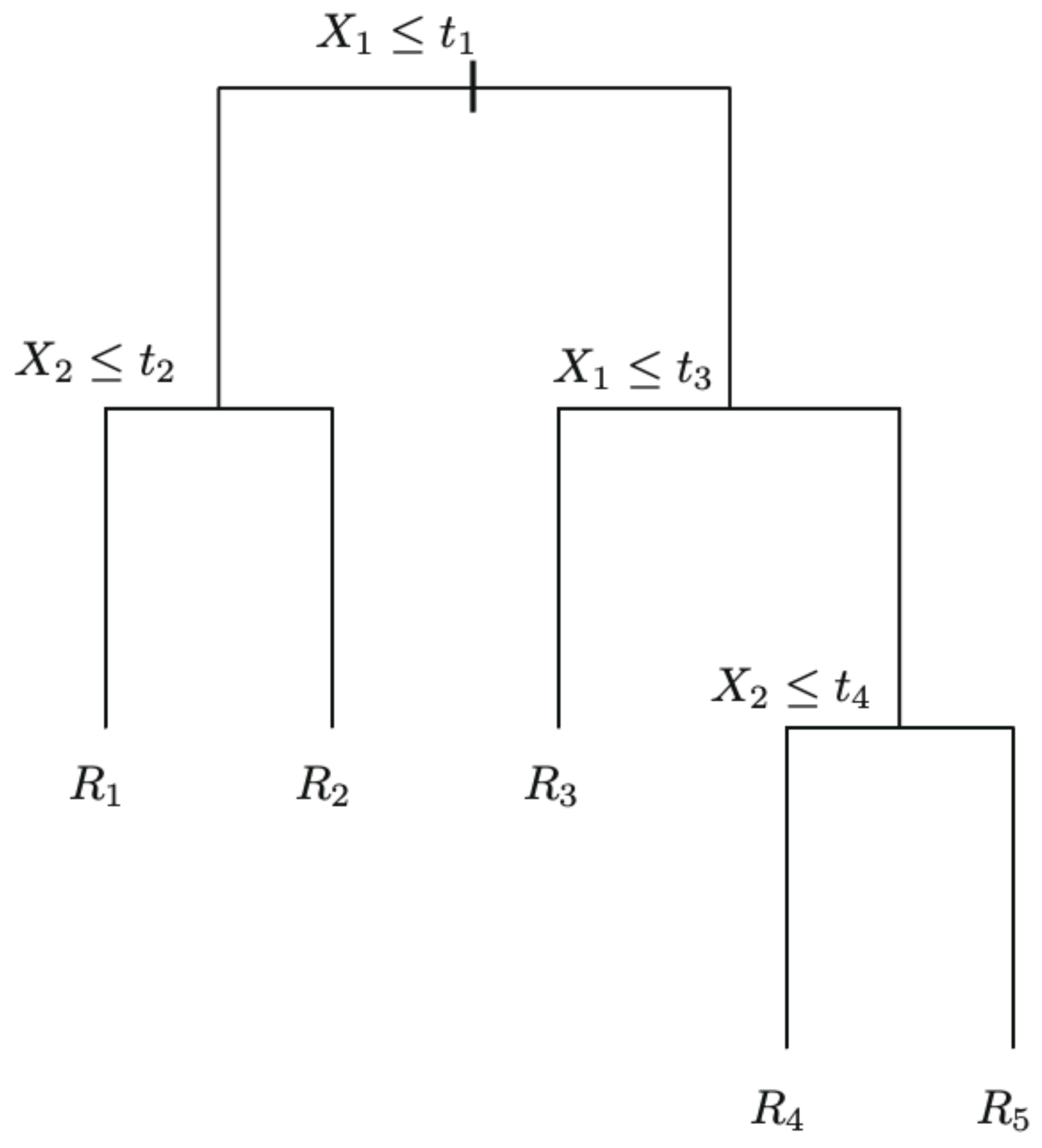
$$R_3 = \{X \mid Years > 4.5,\ Hitts > 117.5\}$$

Deeper tree models

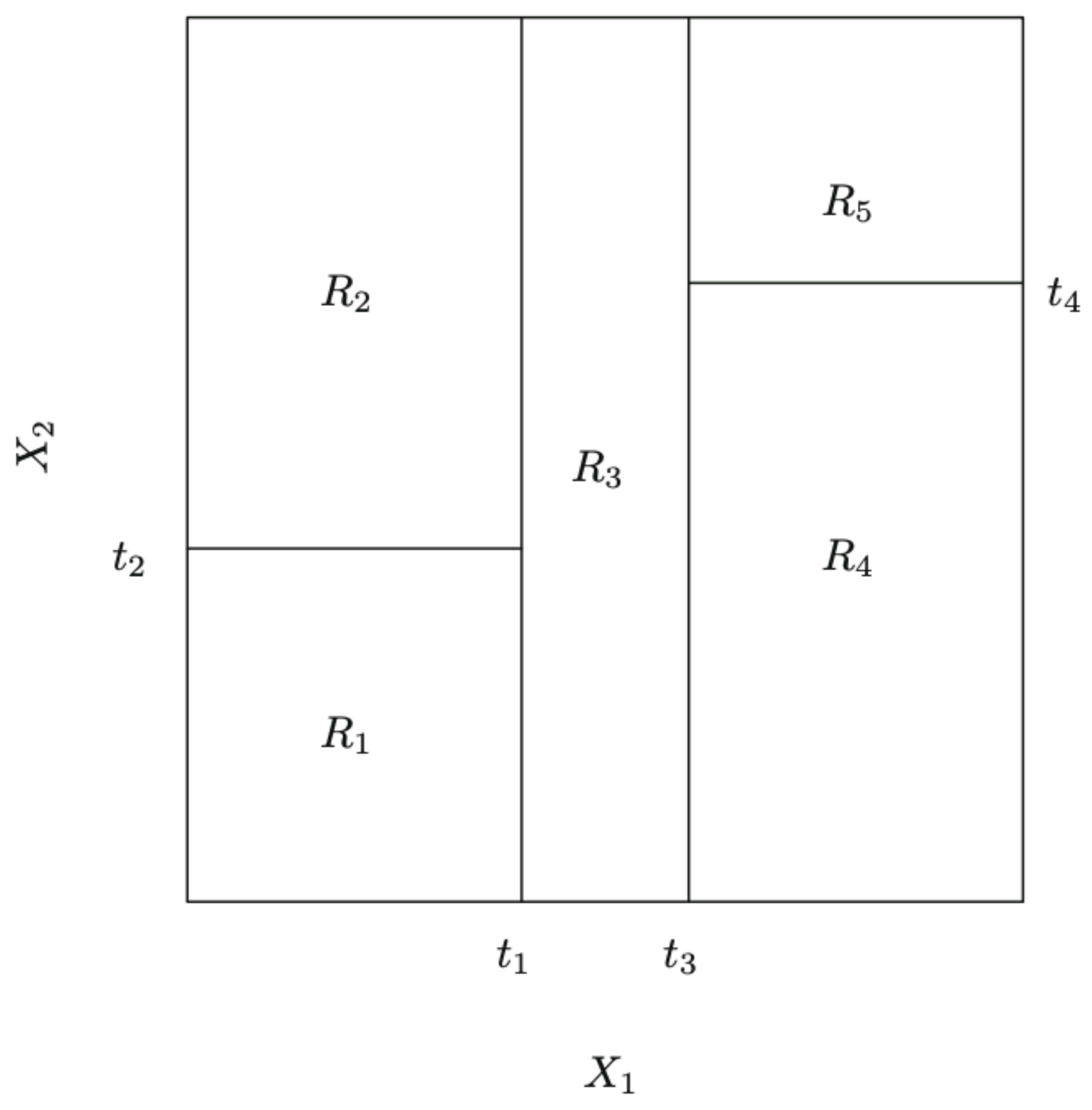


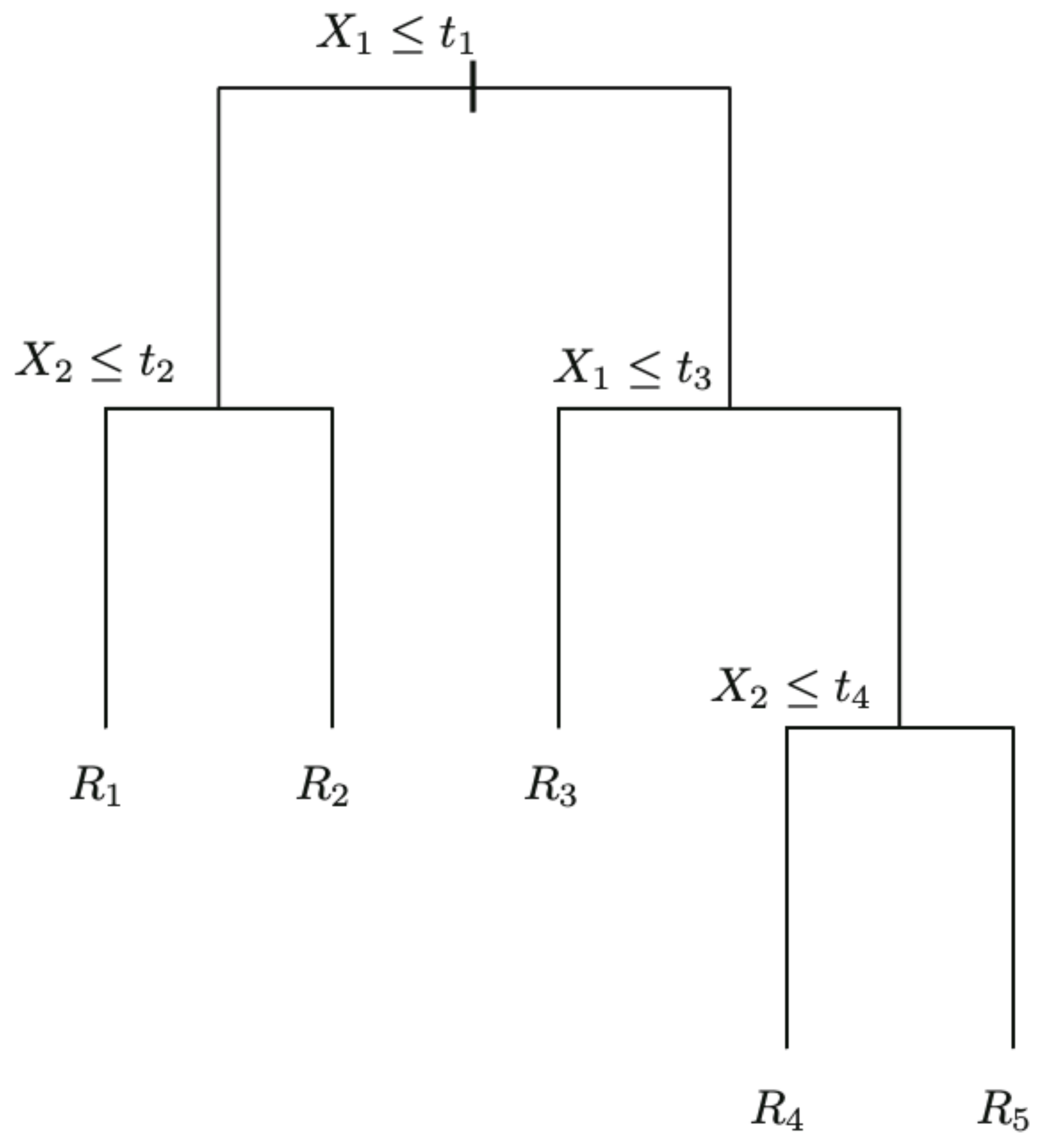
Deeper tree models





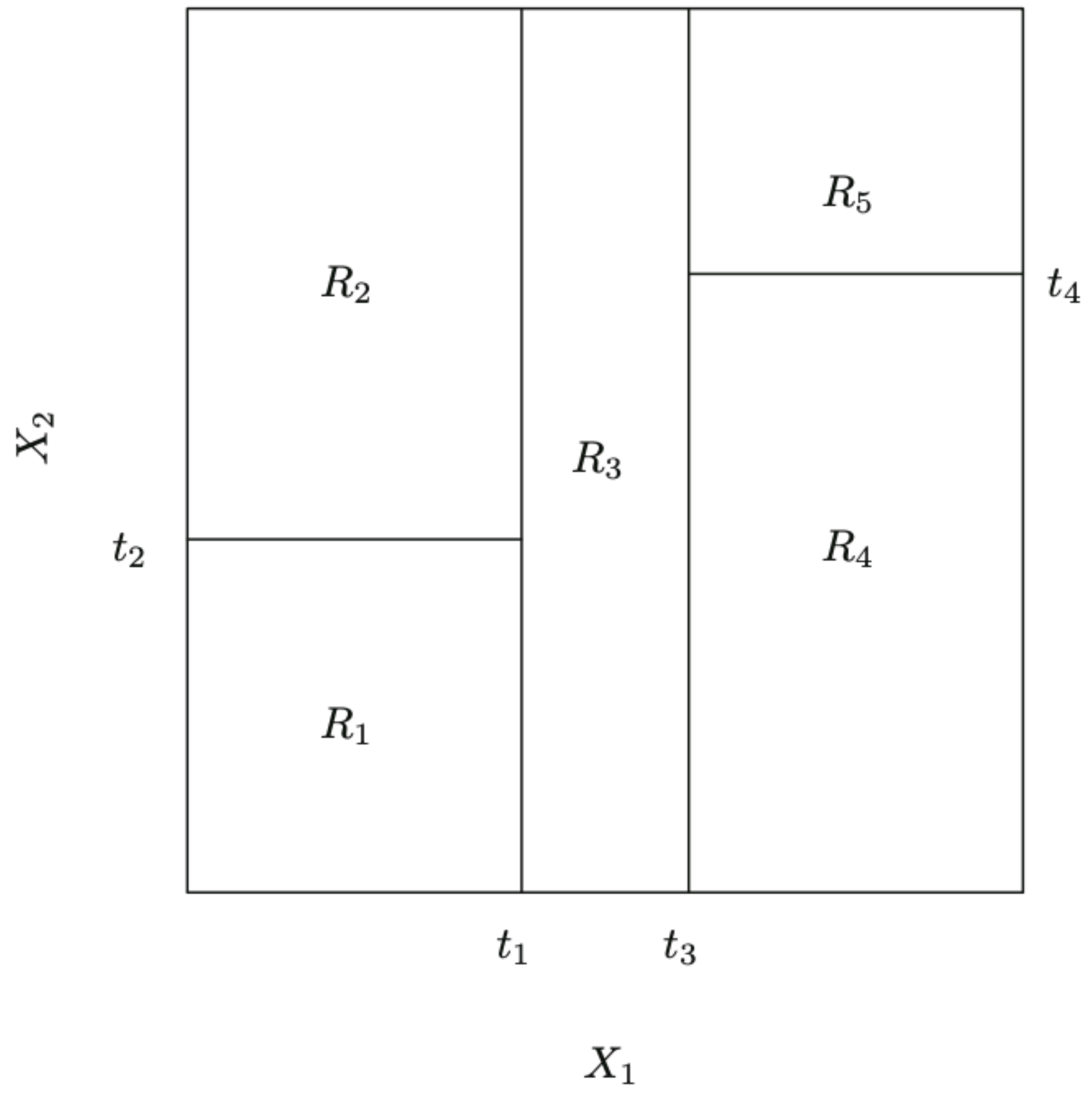
Deeper tree models

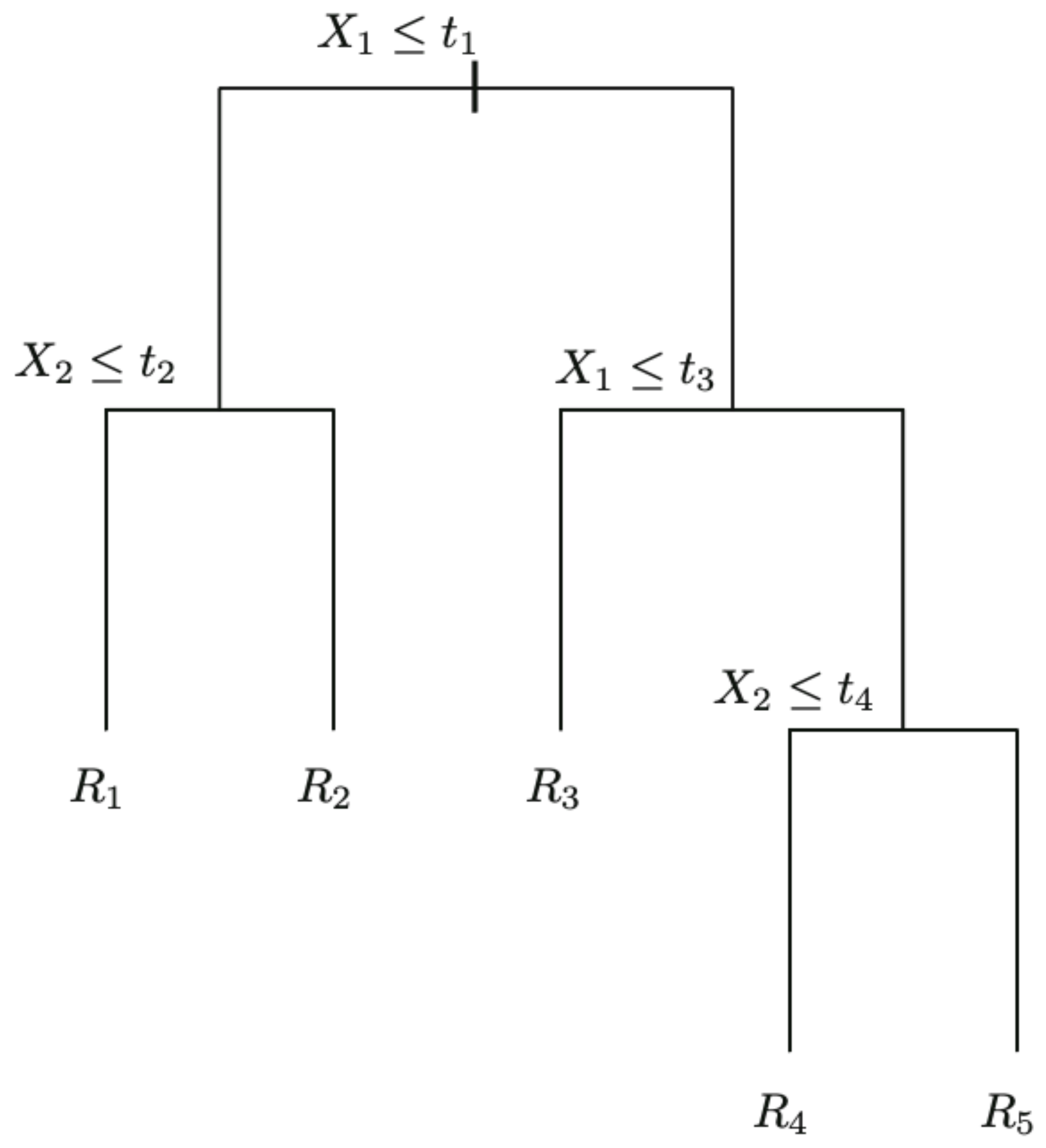


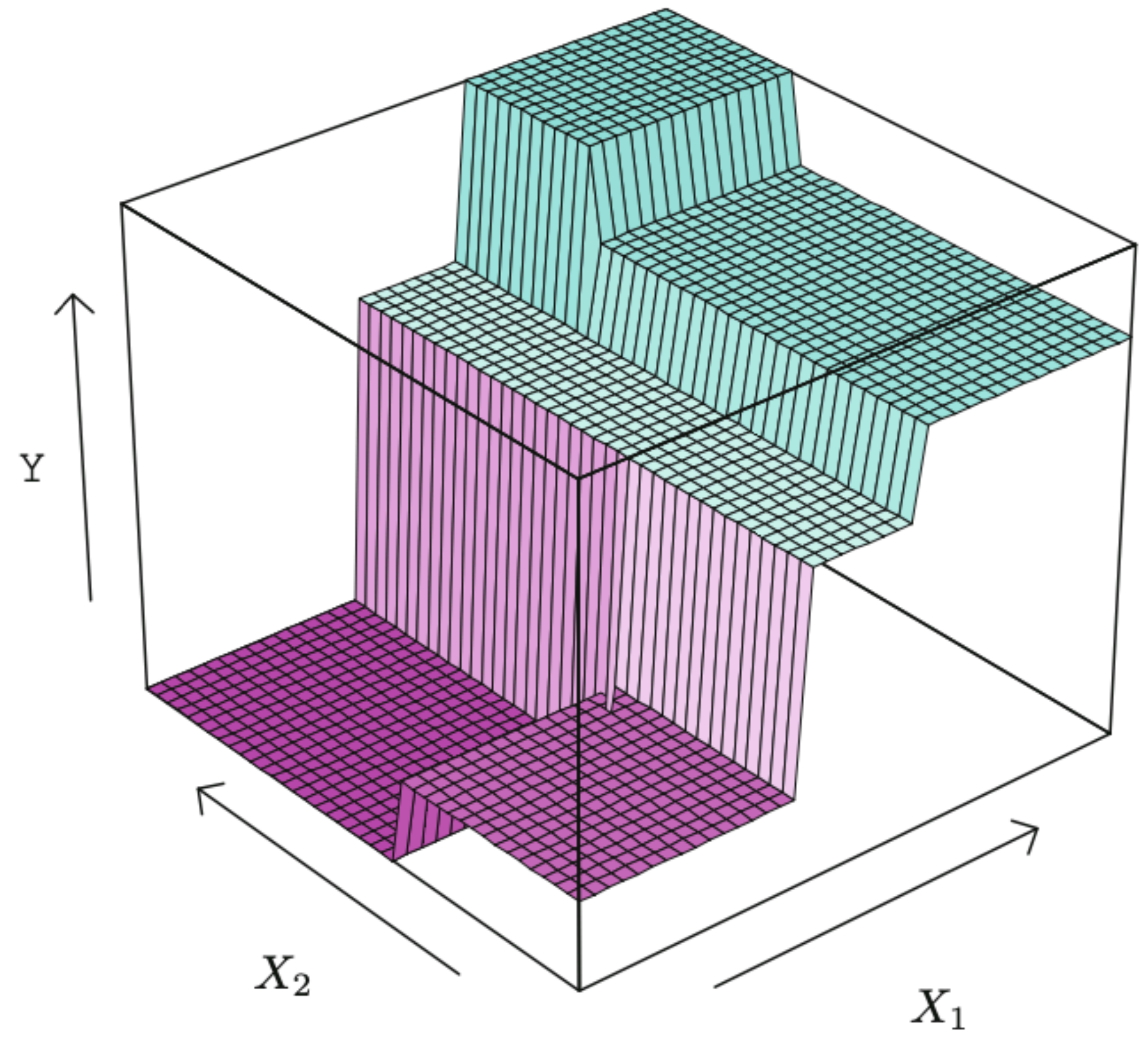


Convince yourself that the tree leads to the partition of X into the left diagram!!

Deeper tree models



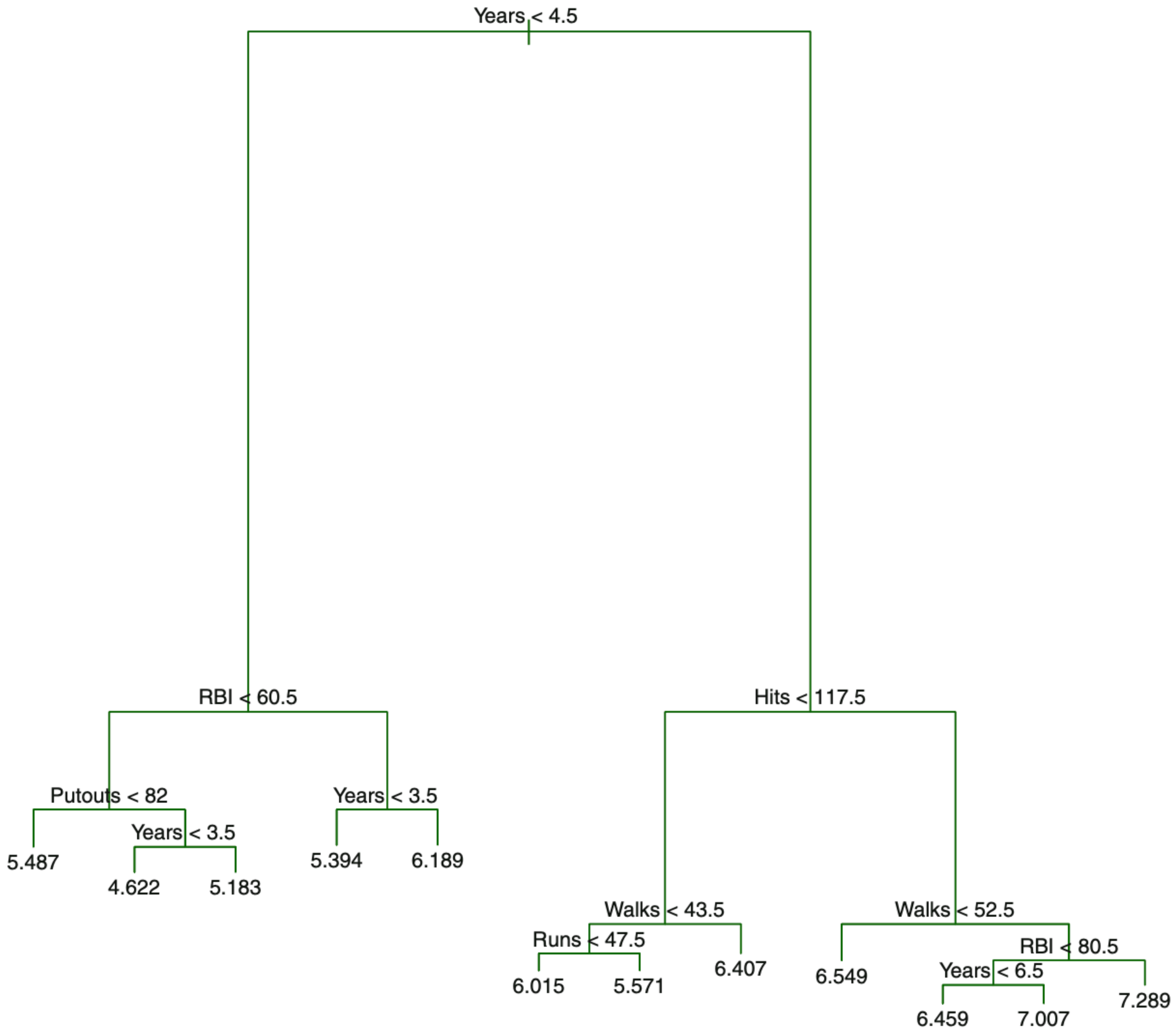




Convince yourself that the tree leads to the partition of X into the left diagram!!

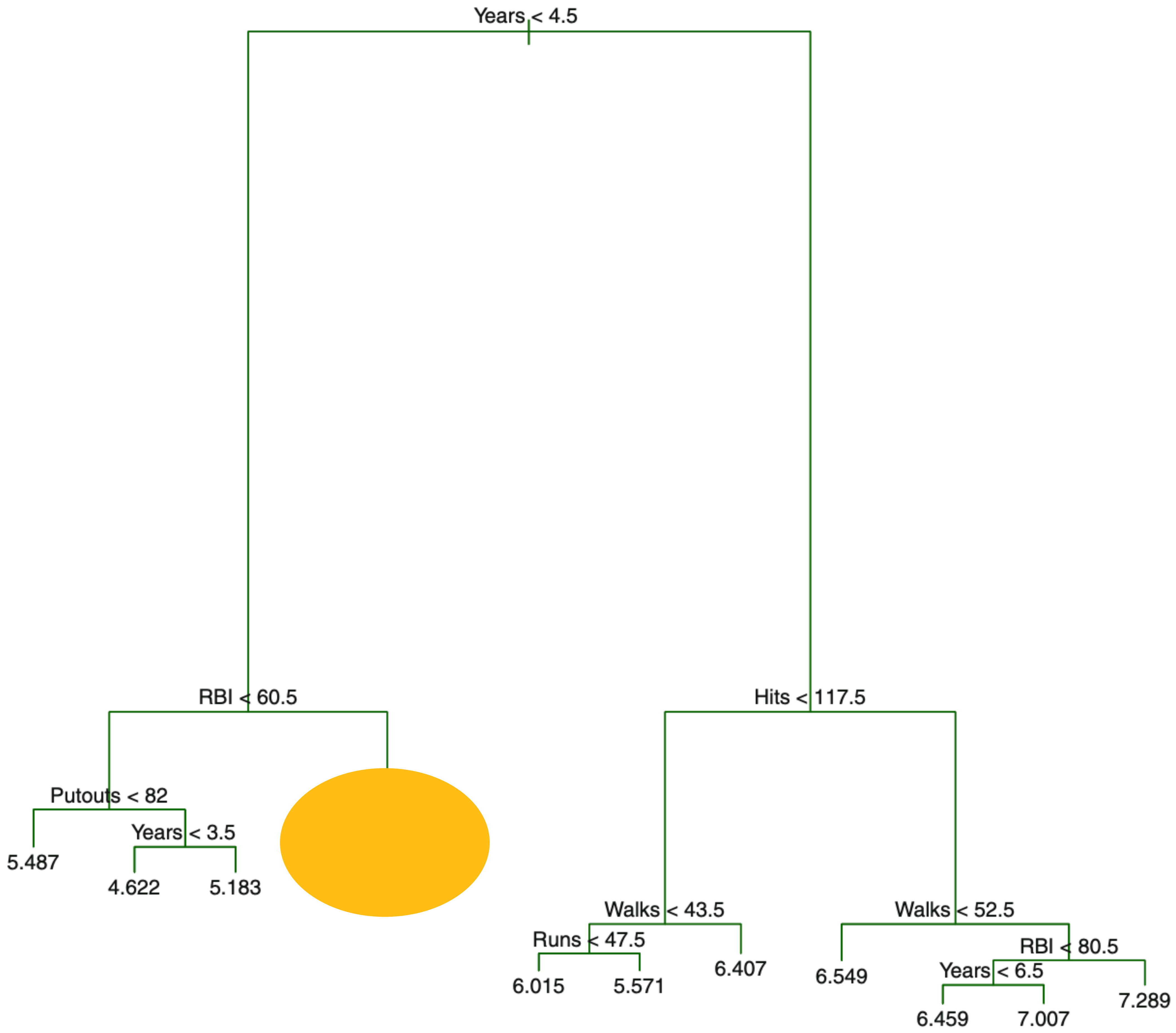
Deeper trees are overfitting

Deeper not always better



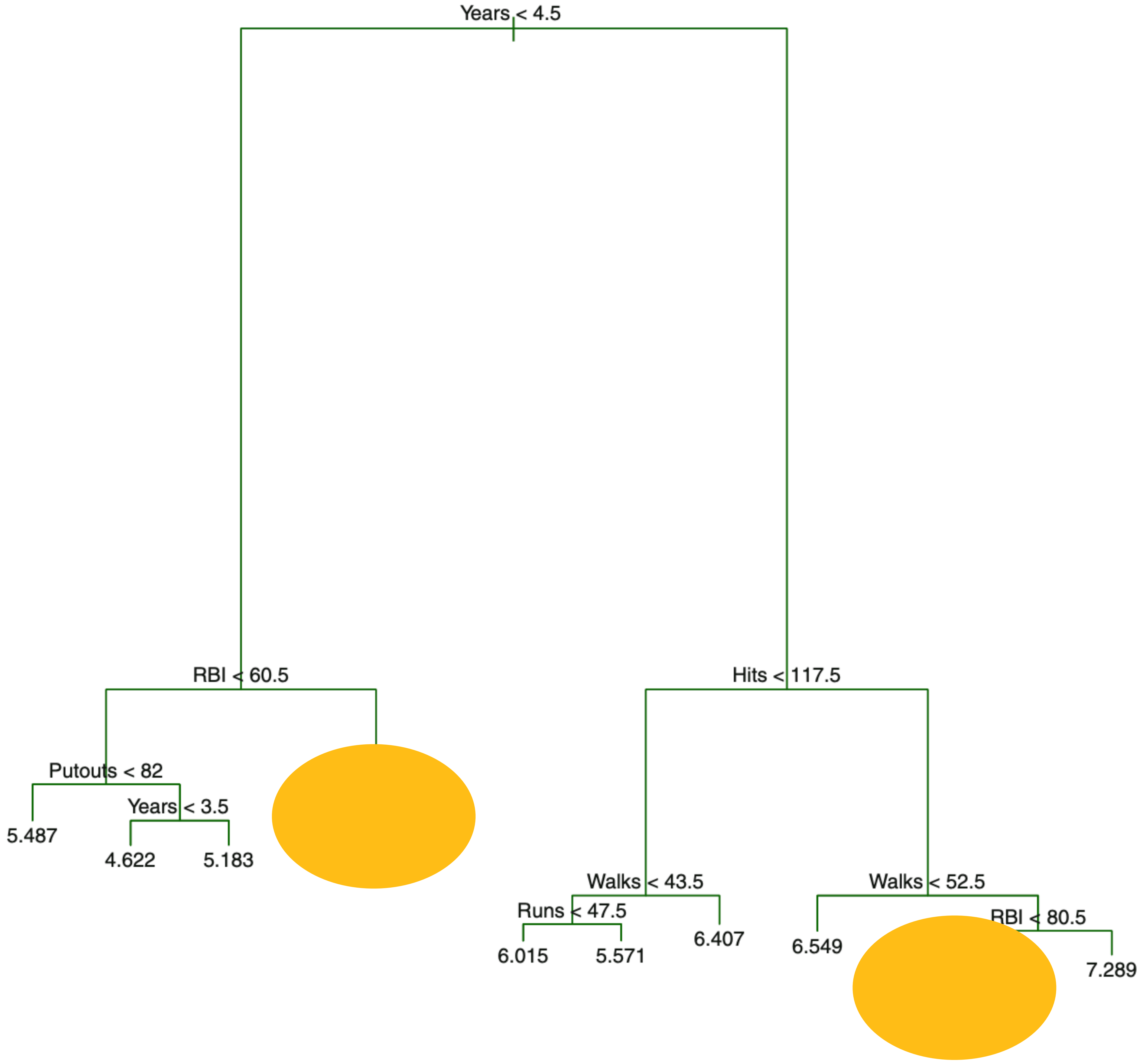
Deeper trees are overfitting

Deeper not always better



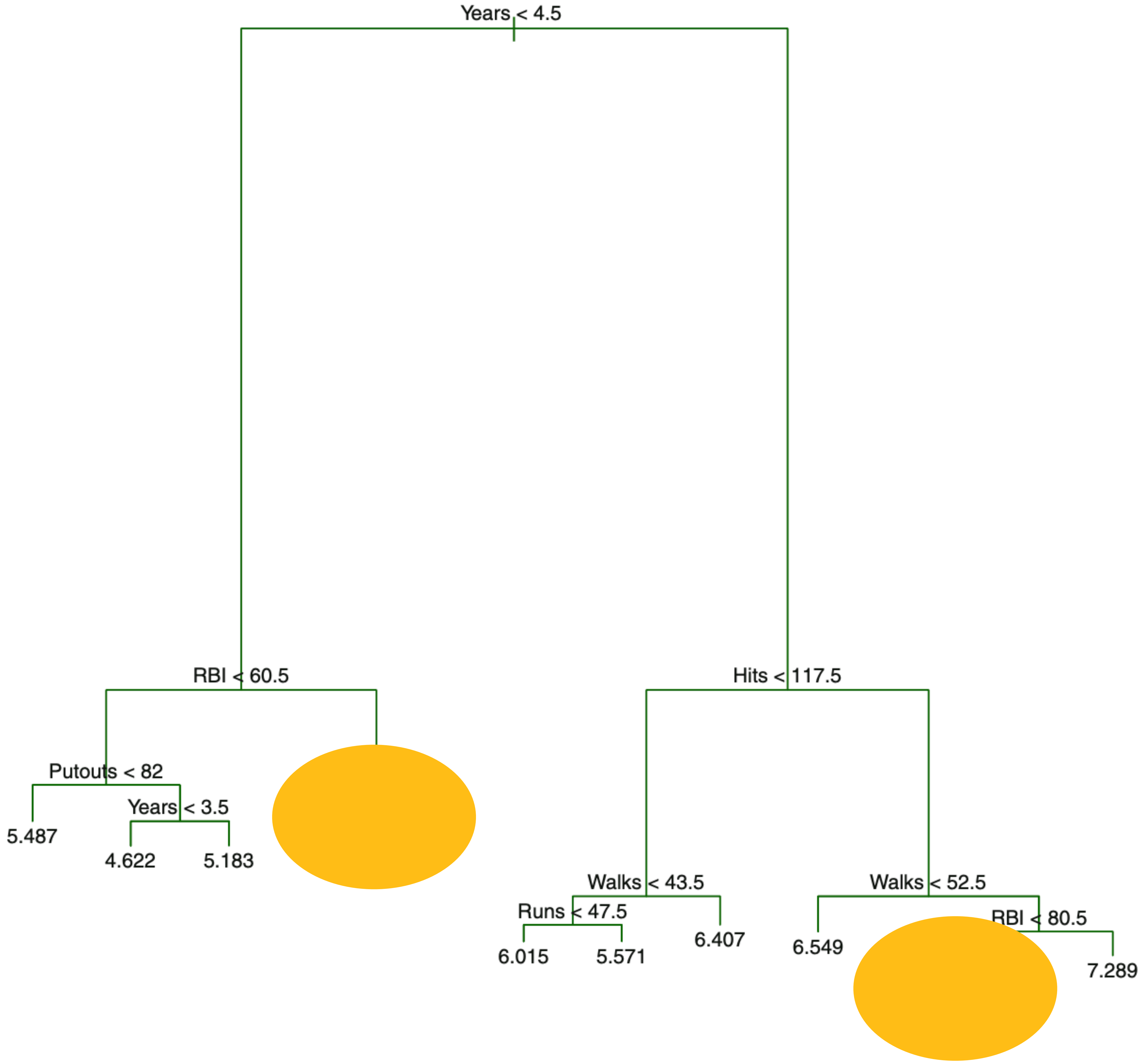
Deeper trees are overfitting

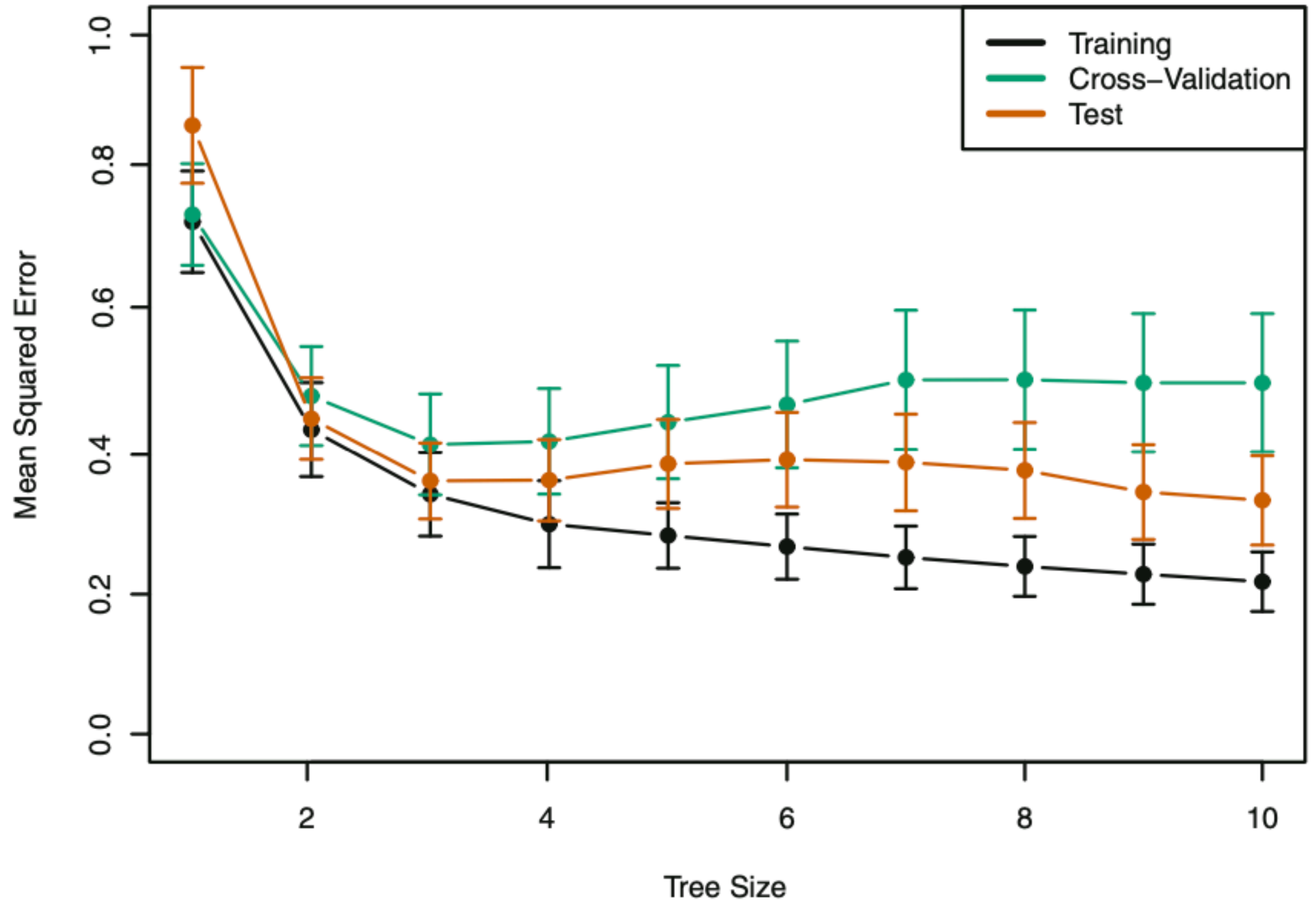
Deeper not always better



Deeper trees are overfitting

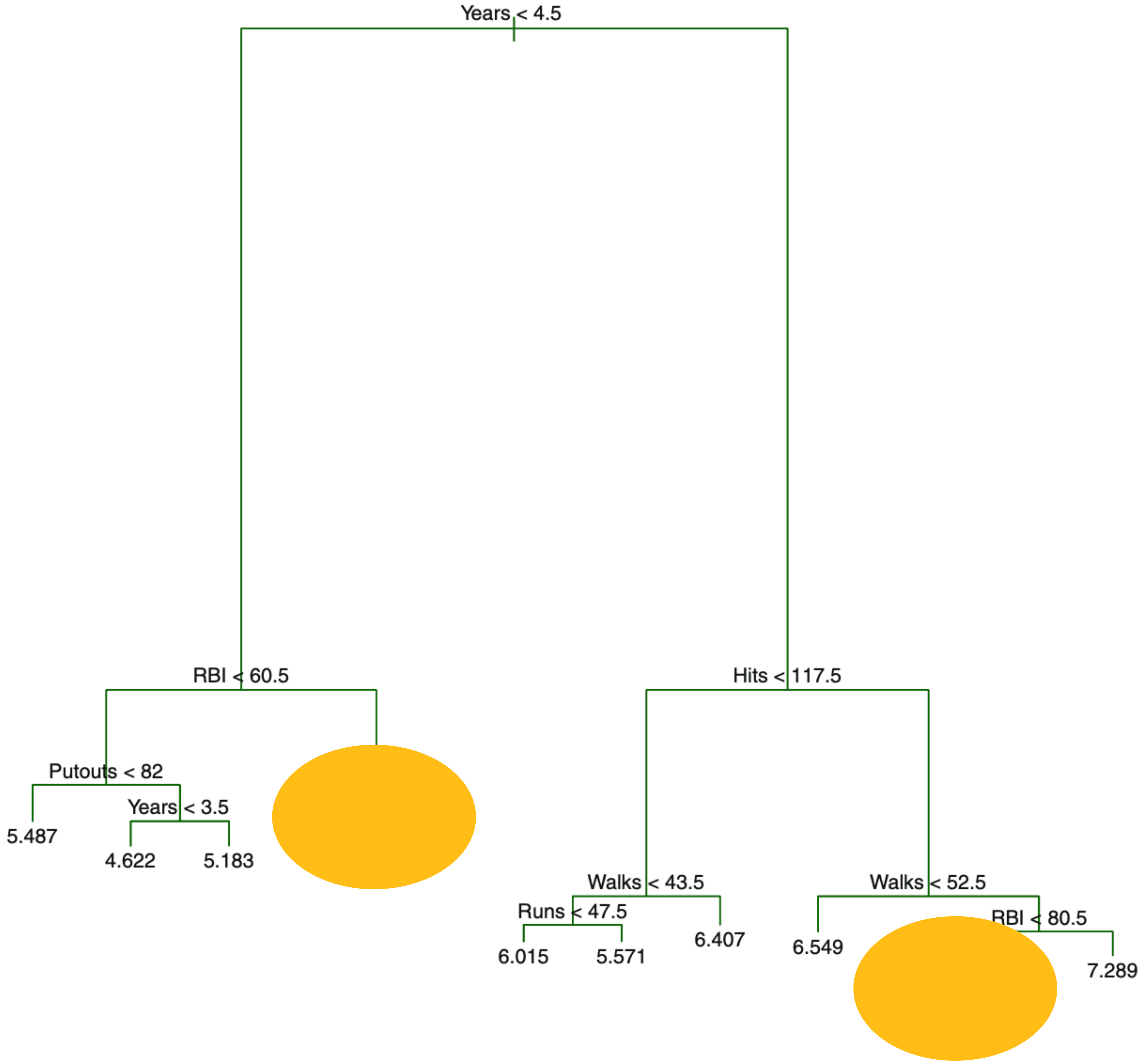
Deeper not always better

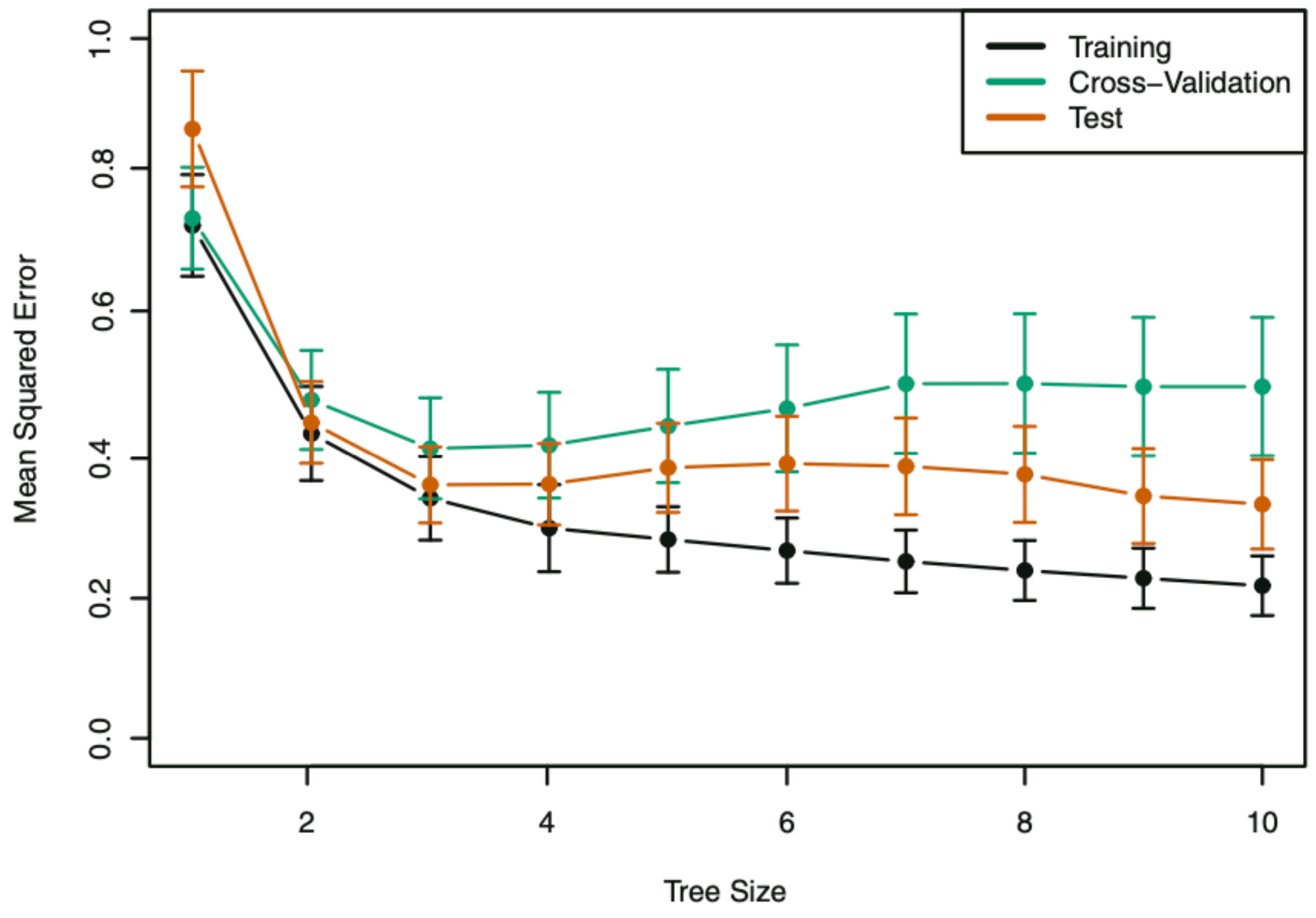




Deeper trees are overfitting

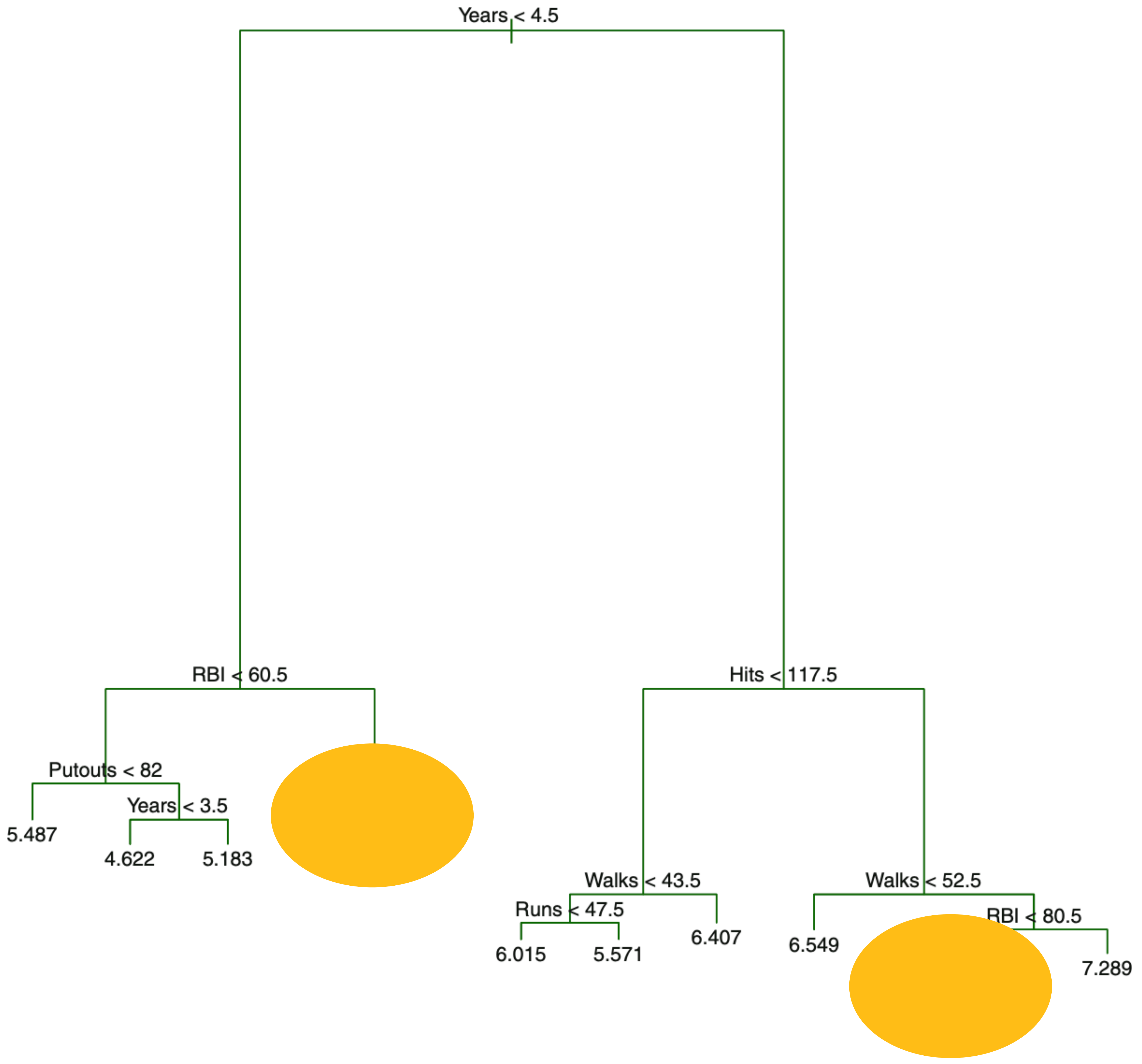
Deeper not always better

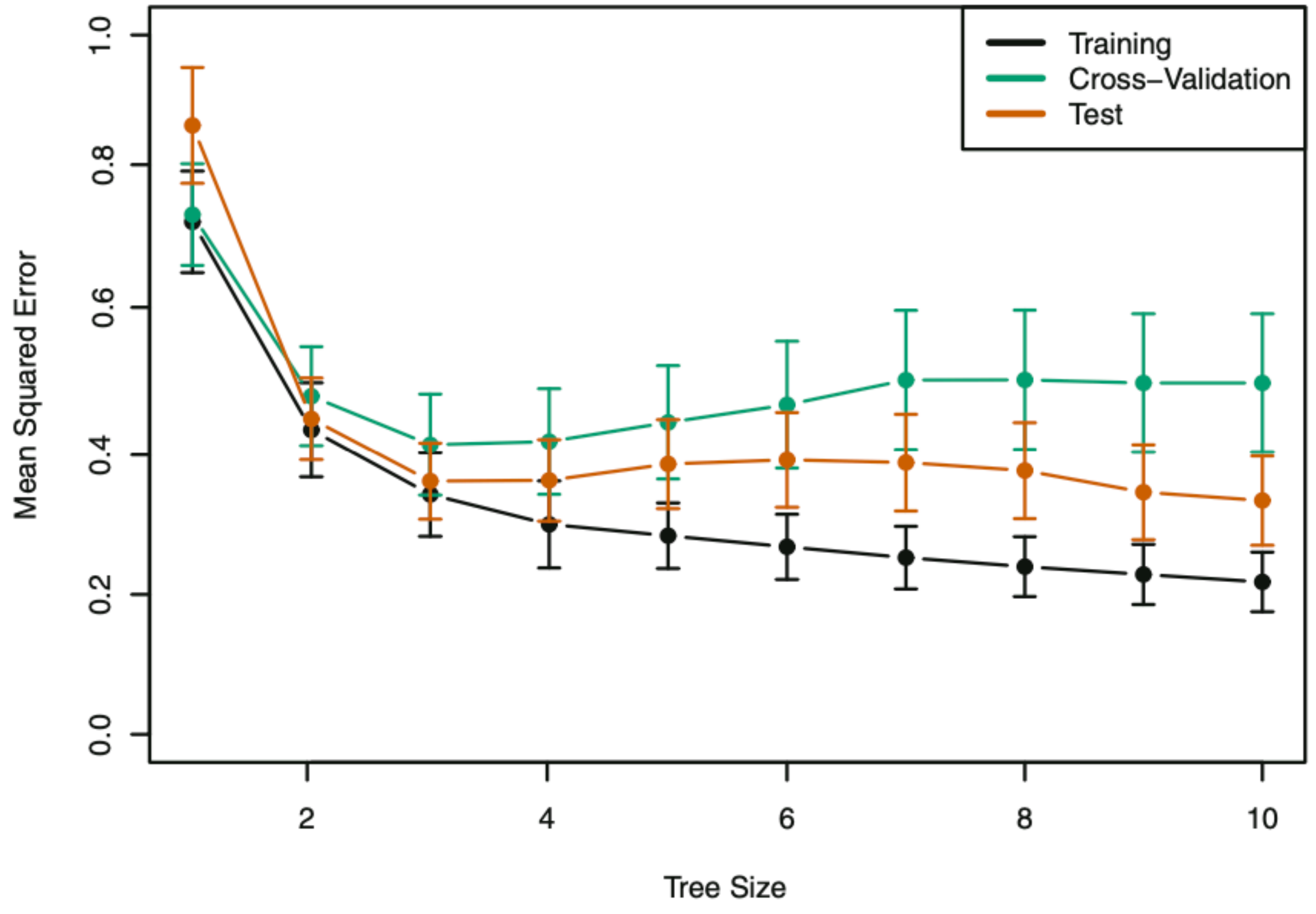




Deeper trees are overfitting

Deeper not always better

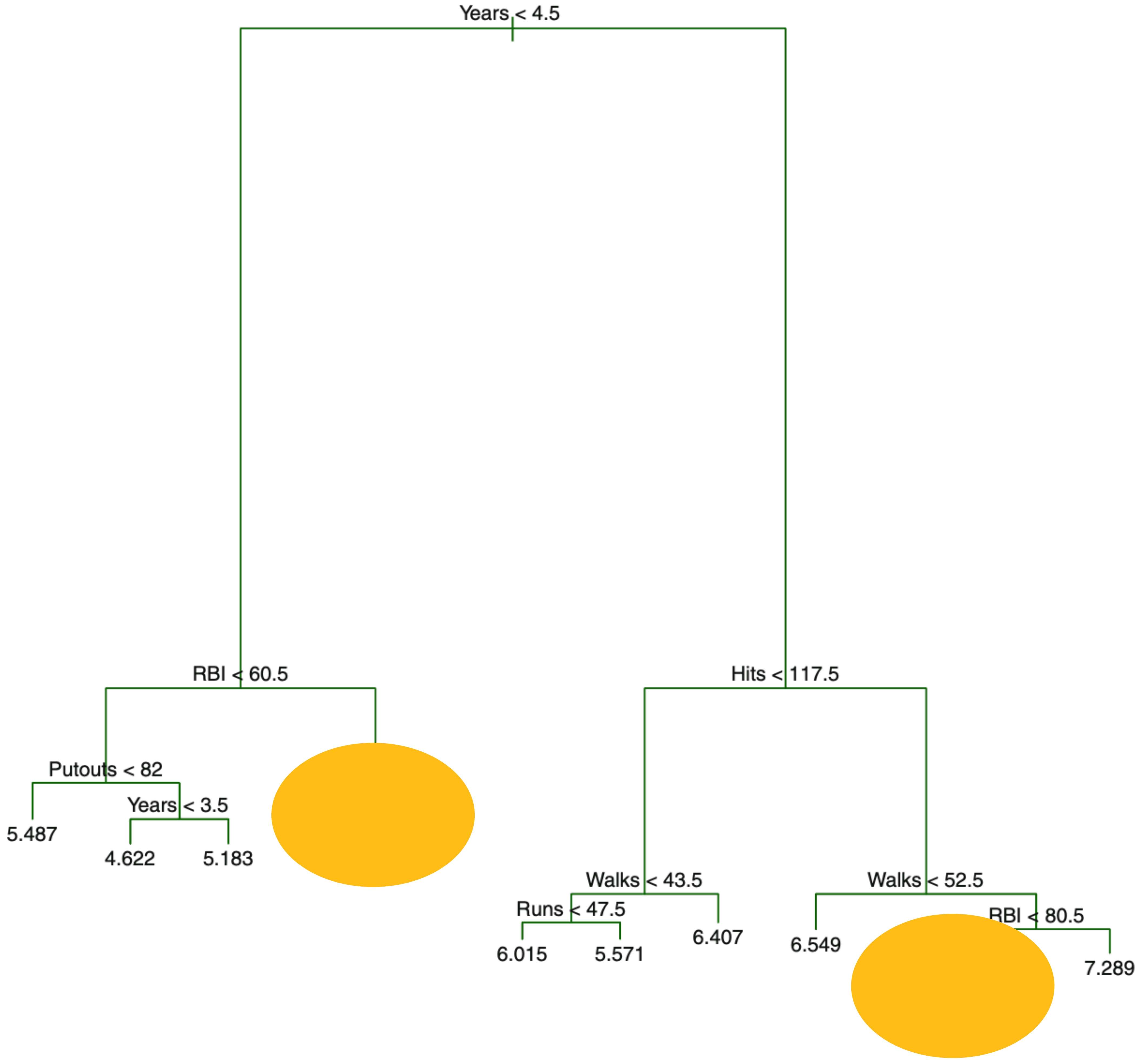


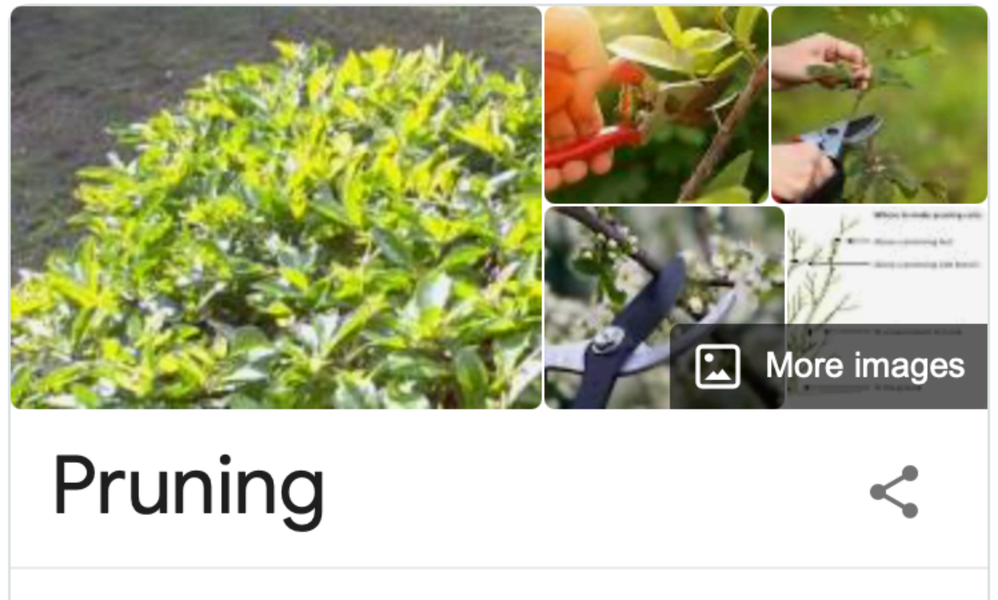


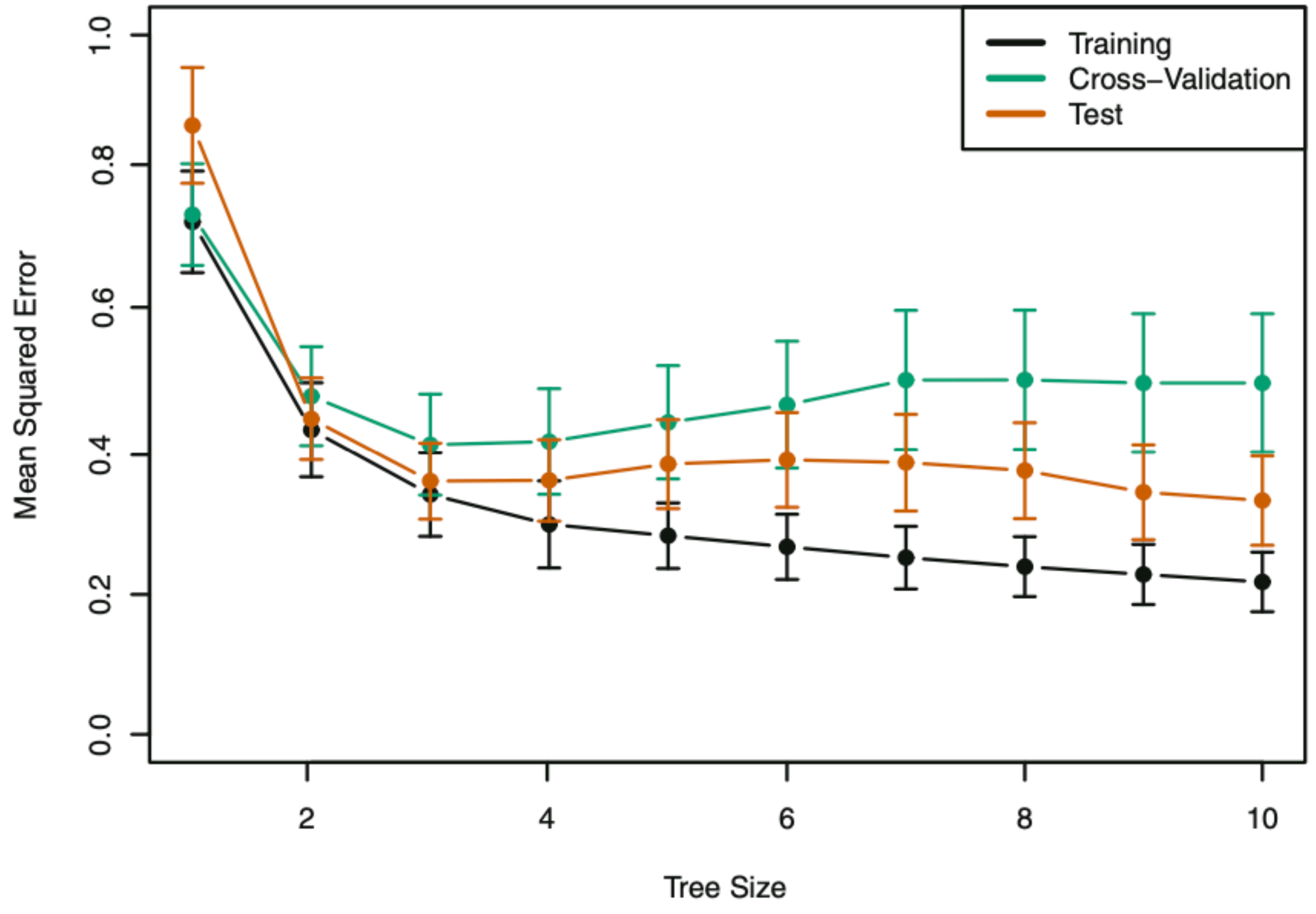
Pruning, less complex, less overfitting!

Deeper trees are overfitting

Deeper not always better







Pruning, less complex, less overfitting!