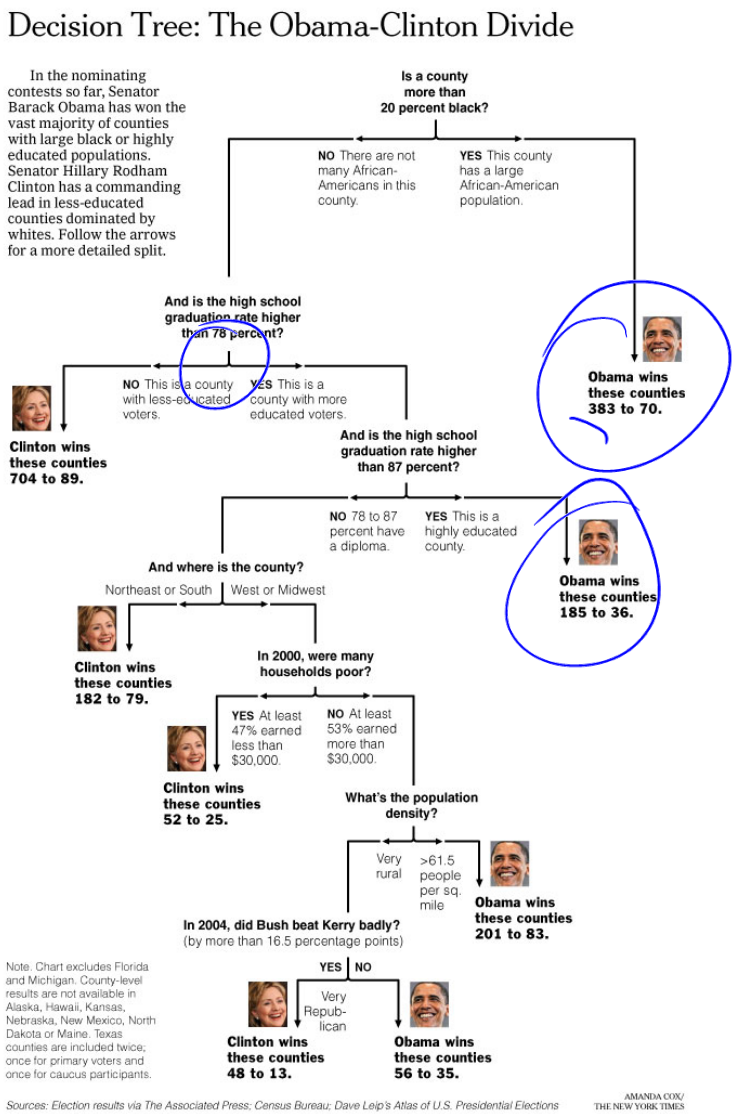
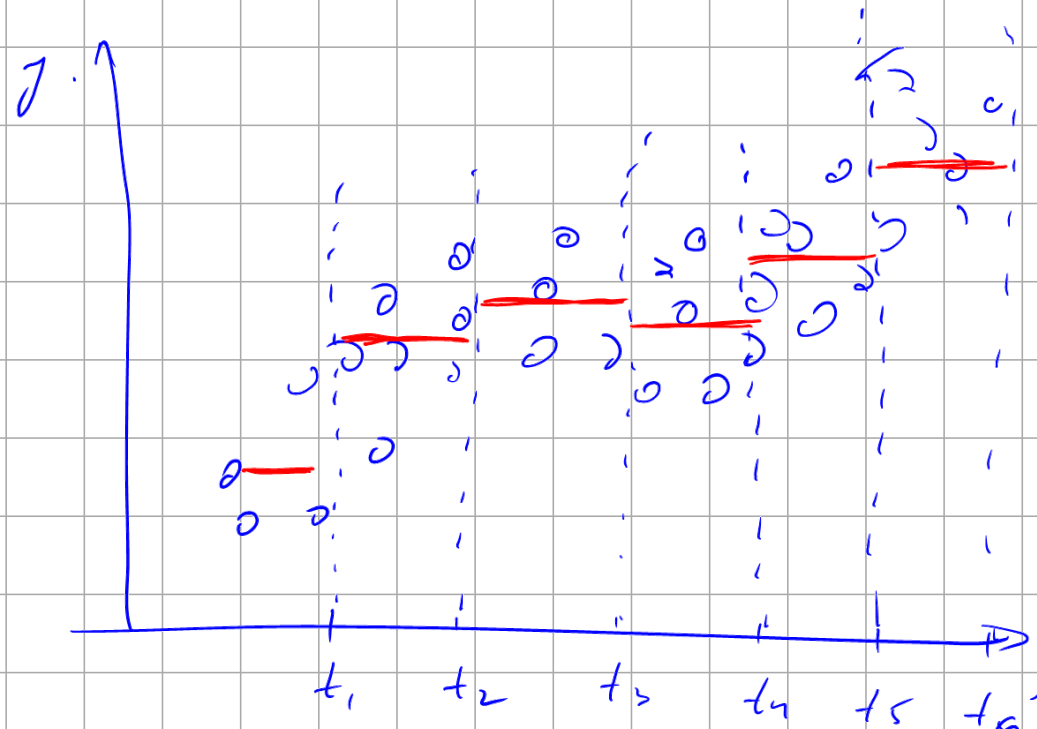
Regression trees: CART



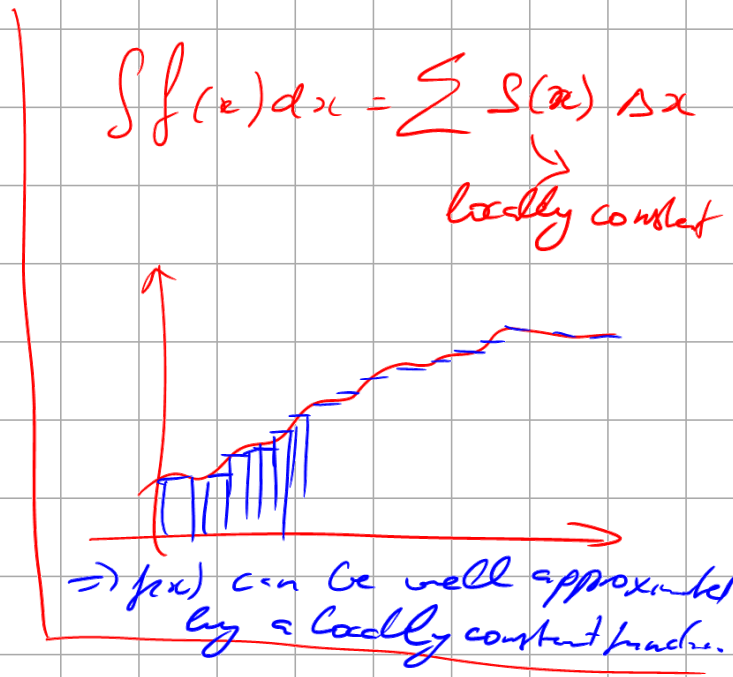
Note: a simple linear regression is too restrictive for large data sets.

Regression trees offer a flexible technique with results, which are easy to interpret



$$\int f(x)\,dx = \sum S(x)\,\Delta x$$

locally constant



$\Rightarrow f(x)$ can be well approximated by a locally constant function.

$$y = m(x) + \varepsilon;$$

$$m(x) = \begin{cases} c_1, & x < t_1 \\ c_2, & t_1 \le x < t_2 \\ \vdots \\ c_6, & x > t_5 \end{cases}$$

averages of the subset of the points falling into that interval

① t_i - ? how to determine the splitting points?

② why c_i are averages?

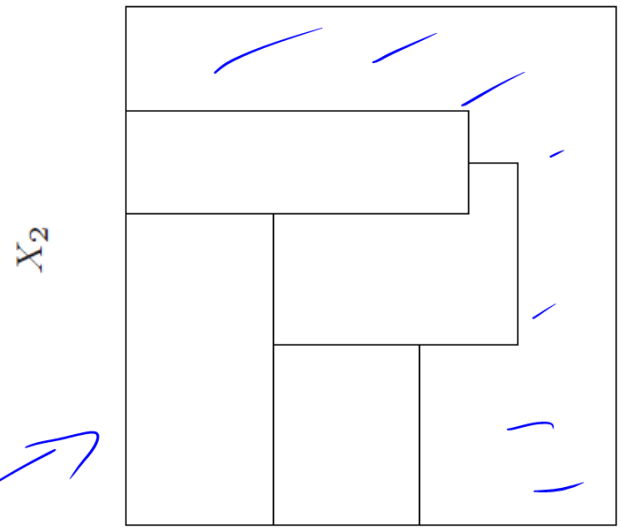
③ what happens if we have more features?

$J = 1 \Rightarrow$ one feature $\Rightarrow$ intervals

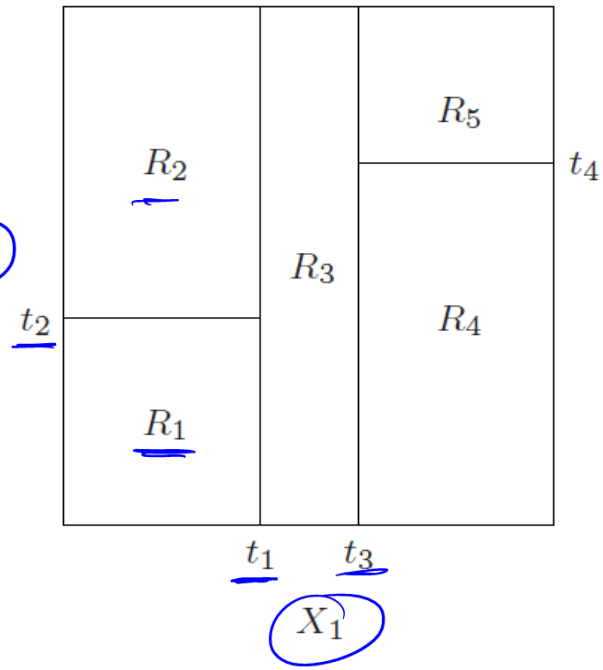
General strategy:

- The values of the explanatory variables are split into P disjunct regions (rectangles) $R_1, \ldots, R_P$: binary splitting

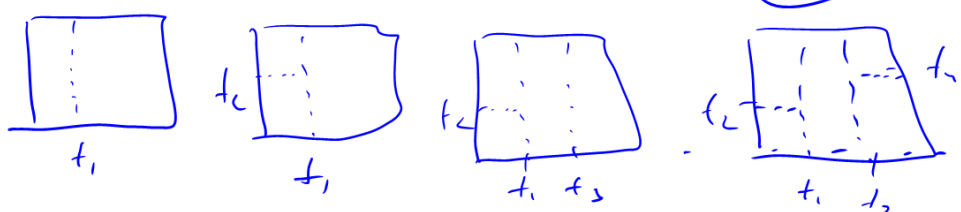
two features



not possible.



Source: Hastie et al. (2001)

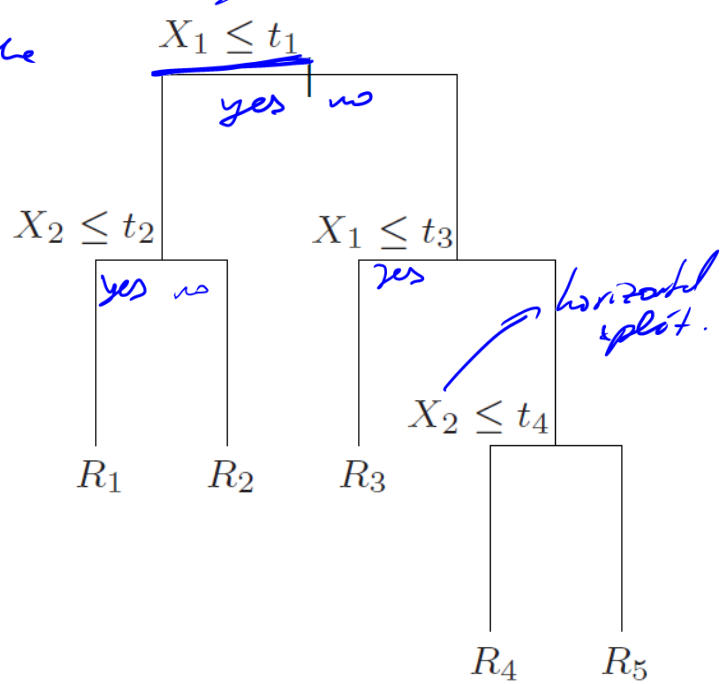


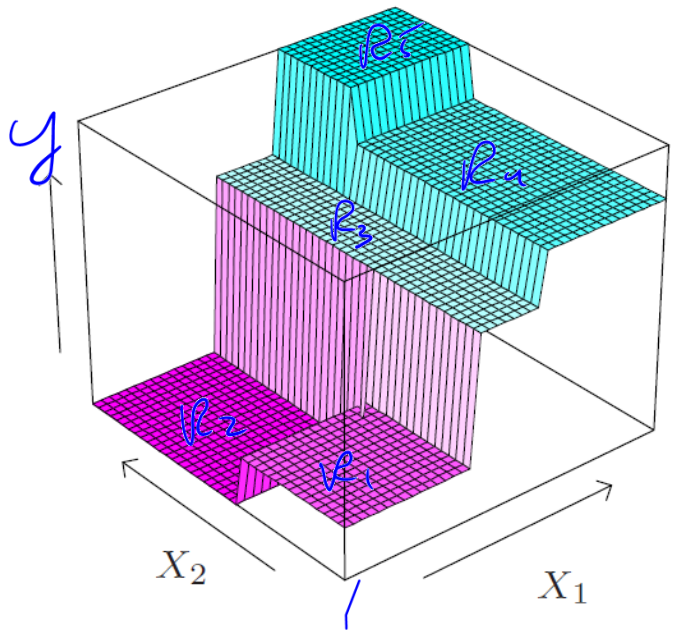
$\Rightarrow$ stepwise procedure

at every step one rectangle is split into two rectangles

- In each rectangle we fit a simple model

  e.g. a constant, i.e. the forecast in rectangle R_p is the mean of all Y-values falling into this rectangle.





Source: Hastie et al. (2001)

Question: how to determine the regions?

OLS method:

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{b})^2 \longrightarrow min, \quad w.r.t. \quad \boldsymbol{b}.$$

For regression trees:

$$\sum_{p=1}^{P}\sum_{i\in R_p}(y_i - \hat{y}_{R_p})^2 \longrightarrow min, \quad w.r.t. \quad R_1, \ldots, R_p,$$

where $\hat{y}_{R_p}$ is the mean of observations in the p-th rectangle.

Note: direct optimization is hardly possible $\rightsquigarrow$ recursive binary splitting

R_p are fixed.

$$\sum_{p=1}^{P} \sum_{i \in R_p} (y_i - c_i)^2 \to \min \quad \text{w.r.t } c_1 \dots c_p$$

$$\sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2 + \dots + \sum_{i \in R_p} (y_i - c_p)^2 \to \min$$

$c_1 \dots c_p$

$\downarrow$ min, w.r.t. c_1 $\quad$ $\downarrow$ min, w.r.t c_2 $\quad$ $\swarrow$ min w.r.t. c_p

$$\frac{\partial}{\partial c_1} = -\sum_{i \in R_1} 2 (y_i - c_1) \stackrel{!}{=} 0 \quad \Rightarrow \quad \hat{c}_1 = \frac{1}{|R_1|} \cdot \sum_{i \in R_1} y_i$$

$\Rightarrow \hat{c}_1$ is simply the average of all y's for point falling into R_1.

Step 1

- Find the variable X_j and the splitting point s, which separates the space into two regions:

$$R_1(j,s) = \{\boldsymbol{X} | X_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{\boldsymbol{X} | X_j > s\}.$$

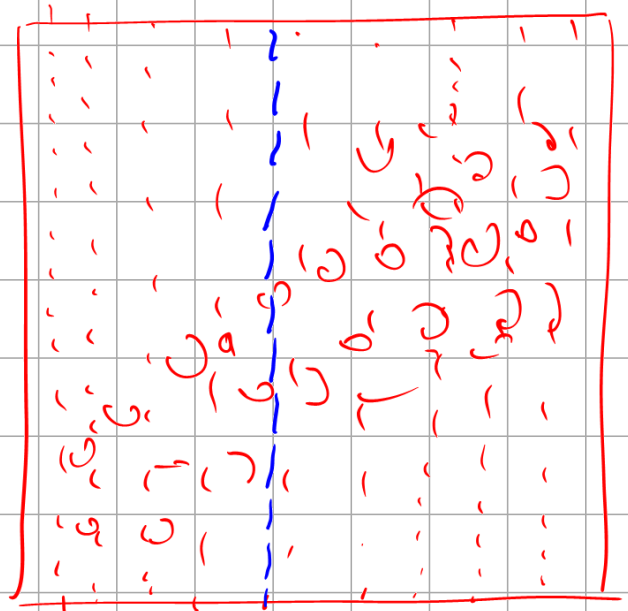
- j and s are determined using the following objective function

$$\sum_{i:\boldsymbol{x}_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\boldsymbol{x}_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are averages in R_1 and R_2.

Step 2

Repeat Step 1 to split regions R_1 and R_2 recursively.



$$\sum_{i:x_i<s} (y_i - \hat{y}_1)^2 + \sum_{i:x_i>s} (y_i - \hat{y}_2)^2$$

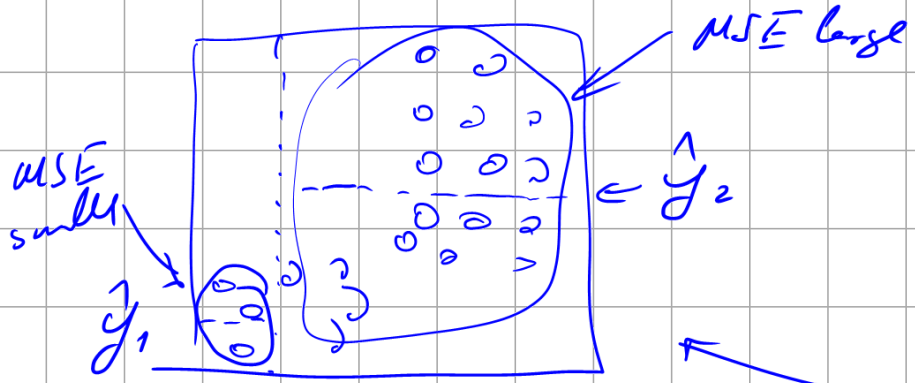
variance of the points to the left from s

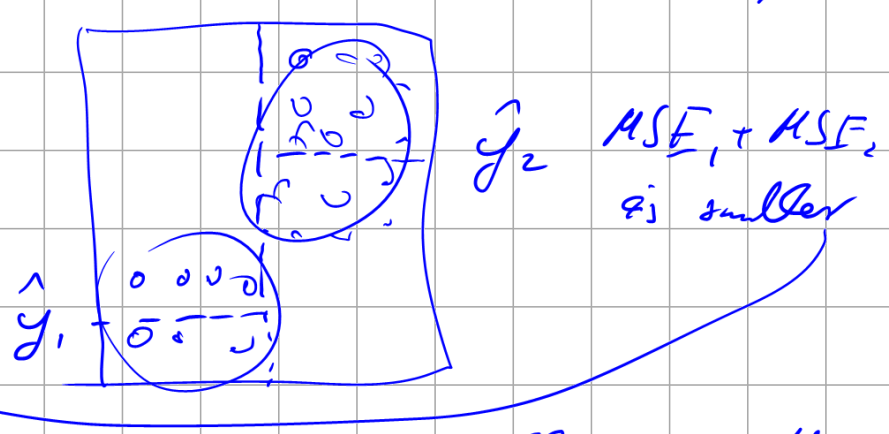
variance/MSE to the right.

$\Rightarrow$ min. w.r.t. s. $\Rightarrow$ simple because of the grid.

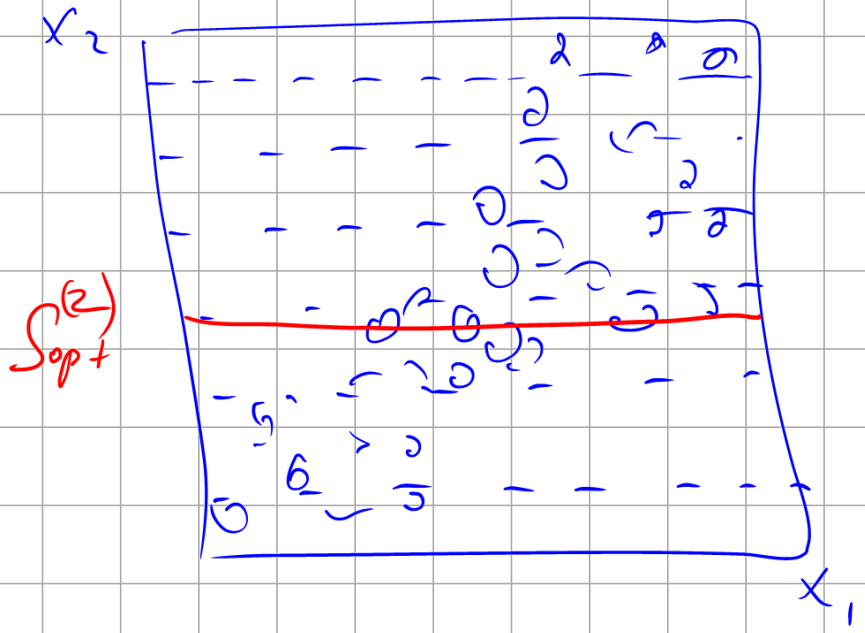
s_{opt}.

$\Rightarrow$ a grid for X_1-values.





$\Rightarrow$ we pick such s, that the variation of points in $\hat{y}$ is as small as possible

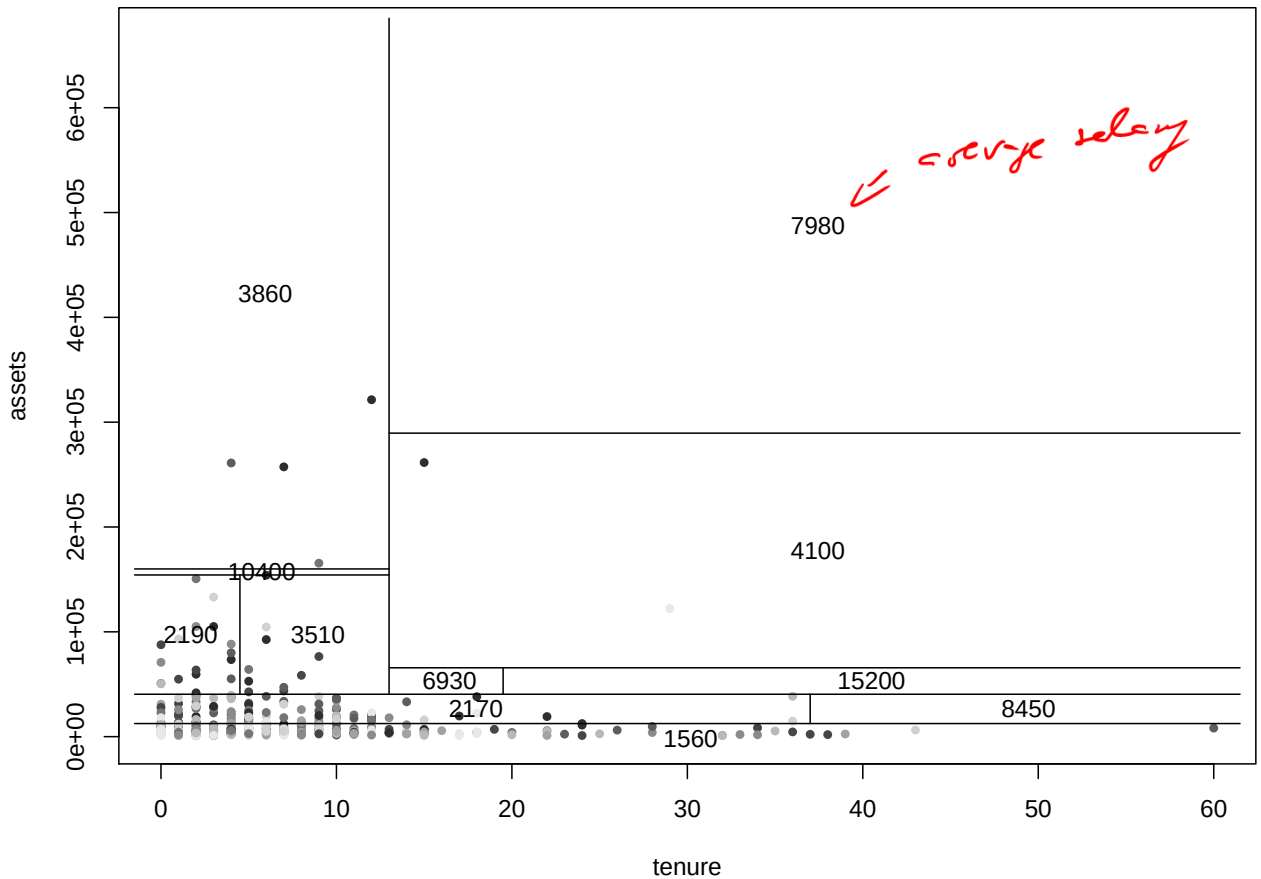


similarly - grid for x_2

$\Rightarrow$ $S_{opt}^{(2)}$ - optimal splitting for x_2.

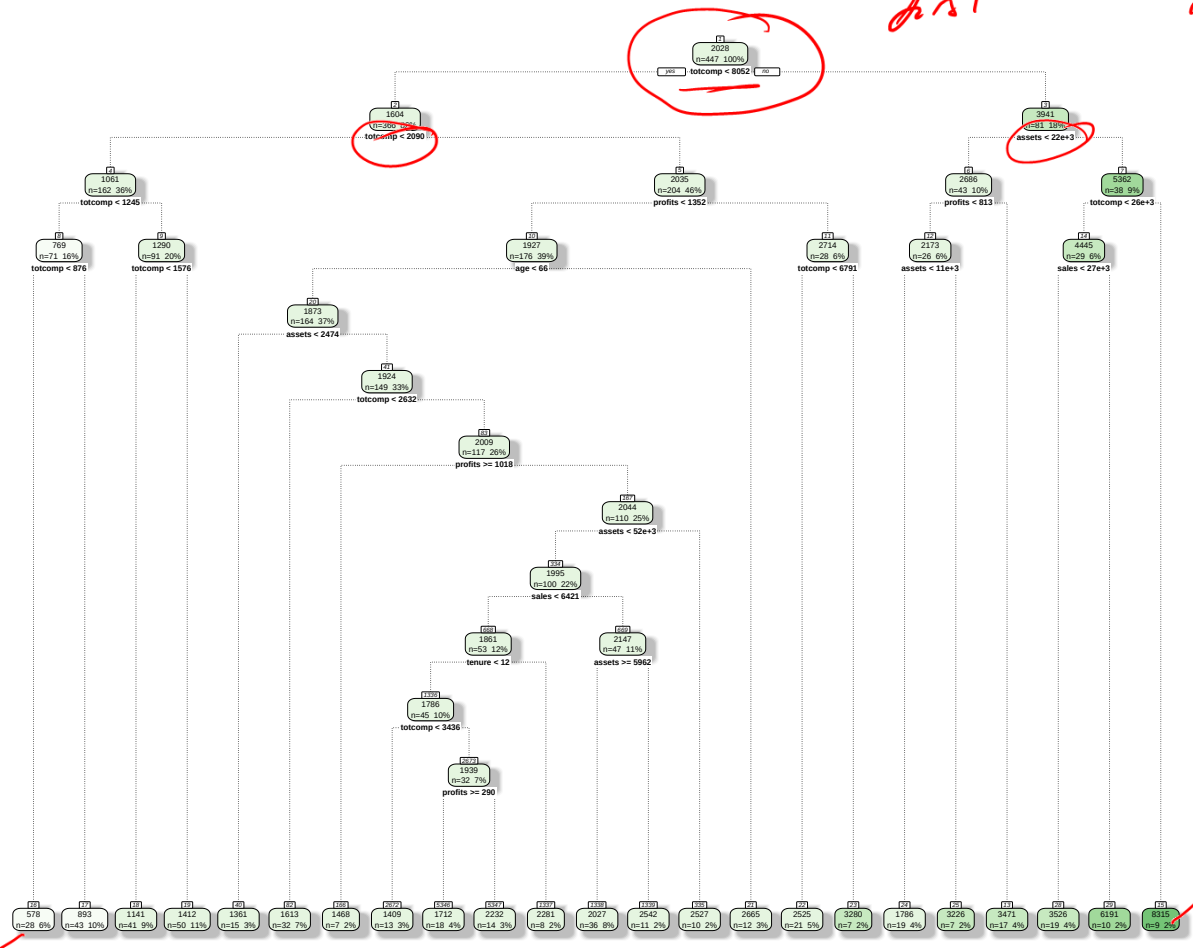
$\Rightarrow$ look which $S_{opt}^{(1)}$ or $S_{opt}^{(2)}$ is better (leads to a smaller value of the objective function)!

```
> library("tree")
> tree.ceo = tree(salary ~ tenure + assets, data=ceo)
> plot(ceo$tenure,ceo$assets, type="p", pch=20, xlab="tenure", ylab="assets")
> partition.tree(tree.ceo, ordvars=c("tenure","assets"), add=TRUE)
```



```
> library("rpart")
> rpart.ceo = rpart(salary ~ .,data=ceo,control=rpart.control(cp = 0.001))
> fancyRpartPlot(rpart.ceo)
```

n = 447 full data set
first variable to split is "totcomp" => most important variable
$t_1 = 8052$



Rattle 2017-Nov-16 09:30:27 okhrinya

28 poorest CEO's with average salary 578 = $\hat{y}_1$

terminal node with 2% of all observations
n = 9.
=> 9 richest CEO's with the average salary of 8315
$\hat{y}_9$

Problems:

1. Does an additional split improve the model?
2. Which criteria shall be used to stop tree growing? $\rightsquigarrow$ complexity parameter, prunning
3. how to determine the importance of the variables?
4. Extensions of CART? $\rightsquigarrow$ bagging, random forest, boosting

Problem 1: Does an additional split improve the model?

- The improvement of the model can be quantified using mean-squared error (MSE)

$$\text{Node 1} : \quad MSE_1 = \frac{1}{K_1}\sum_{k=1}^{K_1}(y_k - \bar{y}_{\text{Node 1}})^2$$

$$\text{Node 2} : \quad MSE_2 = \frac{1}{K_2}\sum_{k=1}^{K_2}(y_k - \bar{y}_{\text{Node 2}})^2$$

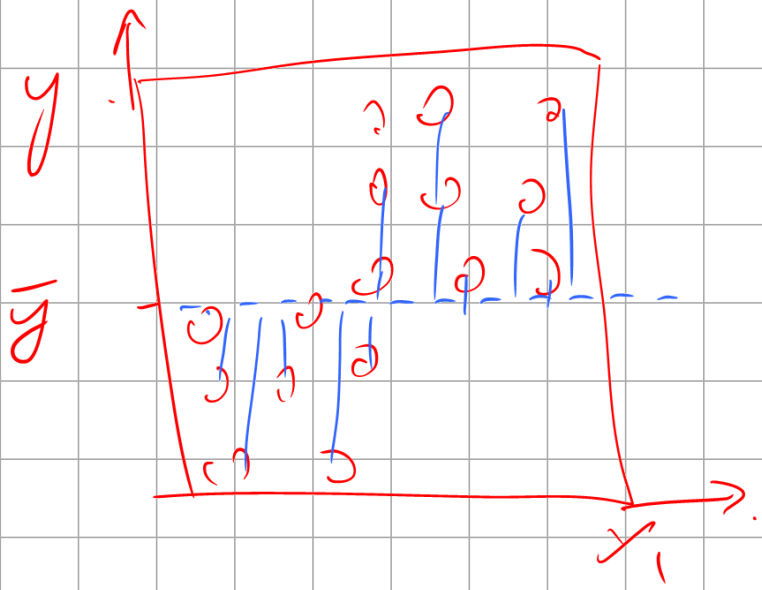
$$\text{Node 3} : \quad MSE_3 = \frac{1}{K_3}\sum_{k=1}^{K_3}(y_k - \bar{y}_{\text{Node 3}})^2$$

before the split — after the split.

$$\text{Improvement} = \frac{MSE_1 \cdot K_1 - [MSE_2 \cdot K_2 + MSE_3 \cdot K_3]}{MSE_1 \cdot K_1} > 0.$$

before

⇒ to set a % improvement in MSE.

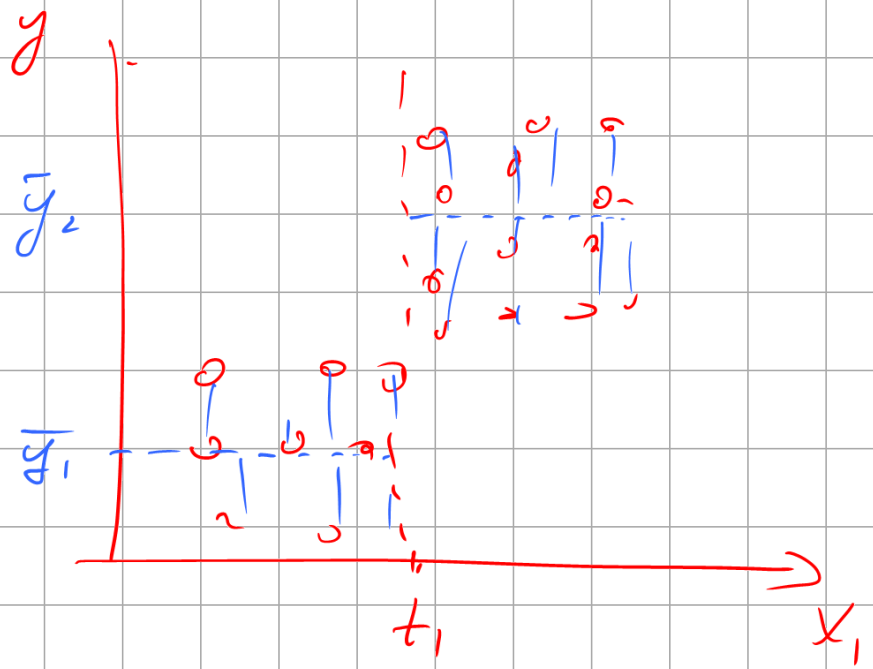


Step 1: don't split $\Rightarrow$

MSE: $\sum_{i=1}^{K_1} (y_i - \bar{y})^2$

$\rightarrow$ total mean.

$\rightarrow$ large.



step 1: 1 split

MSE_{total}: $\sum_{i=1}^{K_2} (y_i - \bar{y}_1) + \sum_{i=1}^{K_3} (y_i - \bar{y}_2)^2$

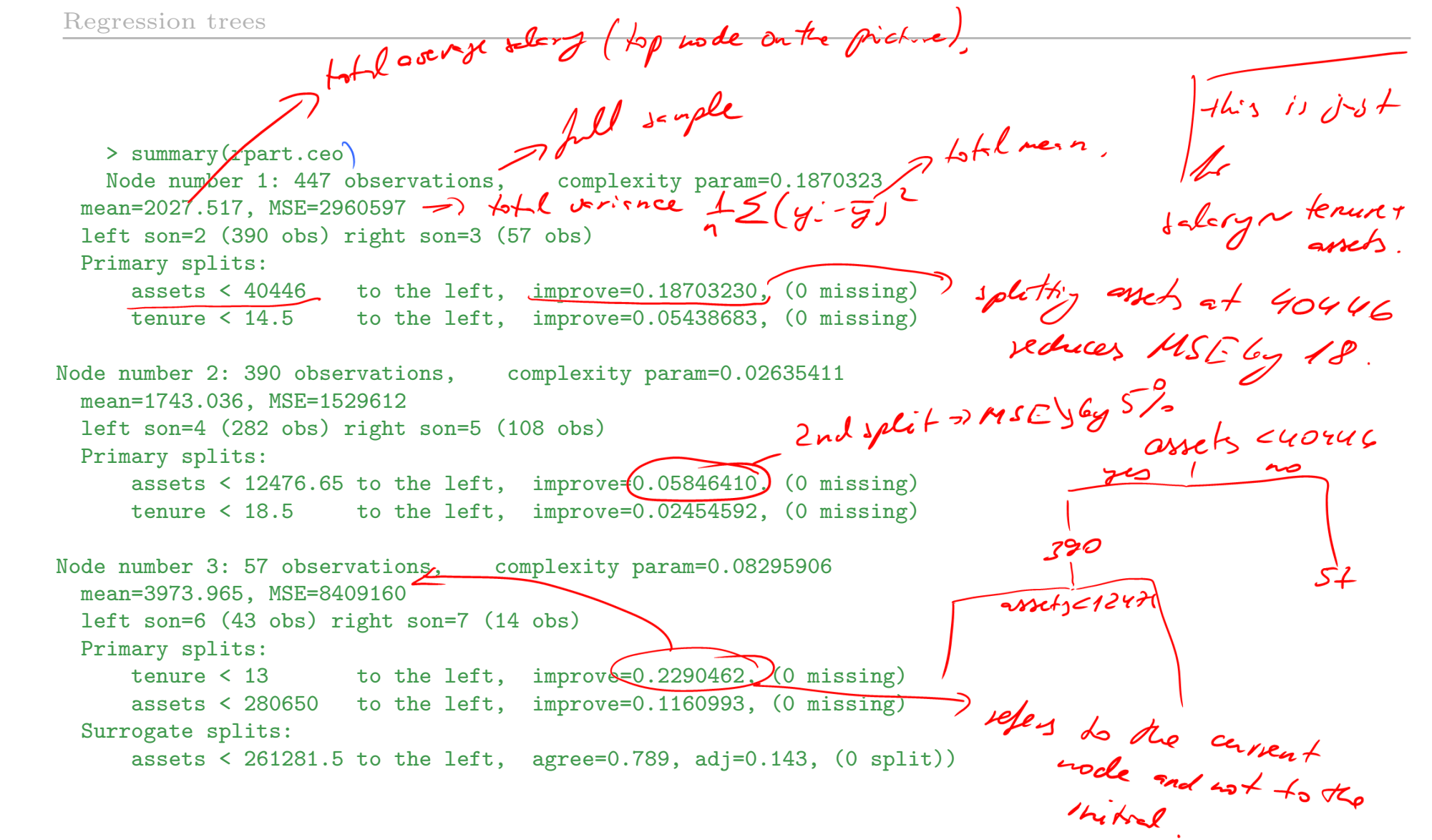
$\Rightarrow$ smaller because deviations are much smaller

$\Rightarrow$

```
> summary(rpart.ceo)
Node number 1: 447 observations,    complexity param=0.1870323
  mean=2027.517, MSE=2960597
  left son=2 (390 obs) right son=3 (57 obs)
  Primary splits:
      assets < 40446    to the left,  improve=0.18703230, (0 missing)
      tenure < 14.5     to the left,  improve=0.05438683, (0 missing)

Node number 2: 390 observations,    complexity param=0.02635411
  mean=1743.036, MSE=1529612
  left son=4 (282 obs) right son=5 (108 obs)
  Primary splits:
      assets < 12476.65 to the left,  improve=0.05846410, (0 missing)
      tenure < 18.5     to the left,  improve=0.02454592, (0 missing)

Node number 3: 57 observations,    complexity param=0.08295906
  mean=3973.965, MSE=8409160
  left son=6 (43 obs) right son=7 (14 obs)
  Primary splits:
      tenure < 13       to the left,  improve=0.2290462, (0 missing)
      assets < 280650   to the left,  improve=0.1160993, (0 missing)
  Surrogate splits:
      assets < 261281.5 to the left,  agree=0.789, adj=0.143, (0 split))
```



Thus the MSE reduces in the first node for 19%.

Problem 2: Using CART we can grow the tree to saturation.

- Fix the maximal number of splittings and a lower bound for the number of observations per region.
- Fix the minimal change in the objective function.
- tree prunning: after the optimal tree is found, it is shortened

Give $1-R^2$.

$$R_\alpha(T) = \frac{1}{\sum_i (y_i - \bar{y})^2} \sum_{m=1}^{|T|} \sum_{i:\boldsymbol{x}_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

where $|T|$ is the number of terminal nodes in a tree and α is the complexity parameter.

α - small $\Rightarrow$ $|T|$ can be bigger $\Rightarrow$ tree is longer

α - large $\Rightarrow$ $|T|$ is smaller $\Rightarrow$ tree is shorter

Key properties of CARTs

- For given α it is possible to determine the tree $T(\alpha)$ with the smallest $R_\alpha(T)$ uniquely
- If $\alpha > \beta$ then $T(\alpha) = T(\beta)$ or $T(\alpha)$ is a strict subtree of $T(\beta)$.

```
> printcp(rpart.ceo)
> printcp(rpart.ceo)

Regression tree:
rpart(formula = salary ~ ., data = ceo, control = rpart.control(cp = 0.001, xval = 10))
Variables actually used in tree construction:
[1] age     assets  profits sales   tenure  totcomp
Root node error: 1323386794/447 = 2960597
n= 447

         CP nsplit rel error  xerror    xstd
1  0.2738266      0   1.00000 1.00703 0.19979
2  0.1091070      1   0.72617 0.77938 0.13638
3  0.0777412      2   0.61707 0.83535 0.15451
4  0.0646524      3   0.53933 0.77968 0.15159
5  0.0351651      4   0.47467 0.70797 0.15182
6  0.0130789      5   0.43951 0.73676 0.17353
7  0.0113130      6   0.42643 0.78445 0.17517
8  0.0081763      7   0.41512 0.79755 0.17644
9  0.0080167      8   0.40694 0.79045 0.17654
10 0.0052976      9   0.39892 0.78938 0.17620
11 0.0032733     10   0.39363 0.78670 0.17590
12 0.0029769     11   0.39035 0.77301 0.17377
13 0.0022593     12   0.38737 0.77146 0.17367
14 0.0017931     13   0.38512 0.77539 0.17310
15 0.0016171     15   0.38153 0.77520 0.17313
16 0.0016099     17   0.37829 0.77603 0.17313
17 0.0012678     20   0.37346 0.77635 0.17313
18 0.0012449     21   0.37220 0.77348 0.17303
19 0.0010000     22   0.37095 0.77403 0.17301
```

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

$\frac{\sum \sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - R^2$.

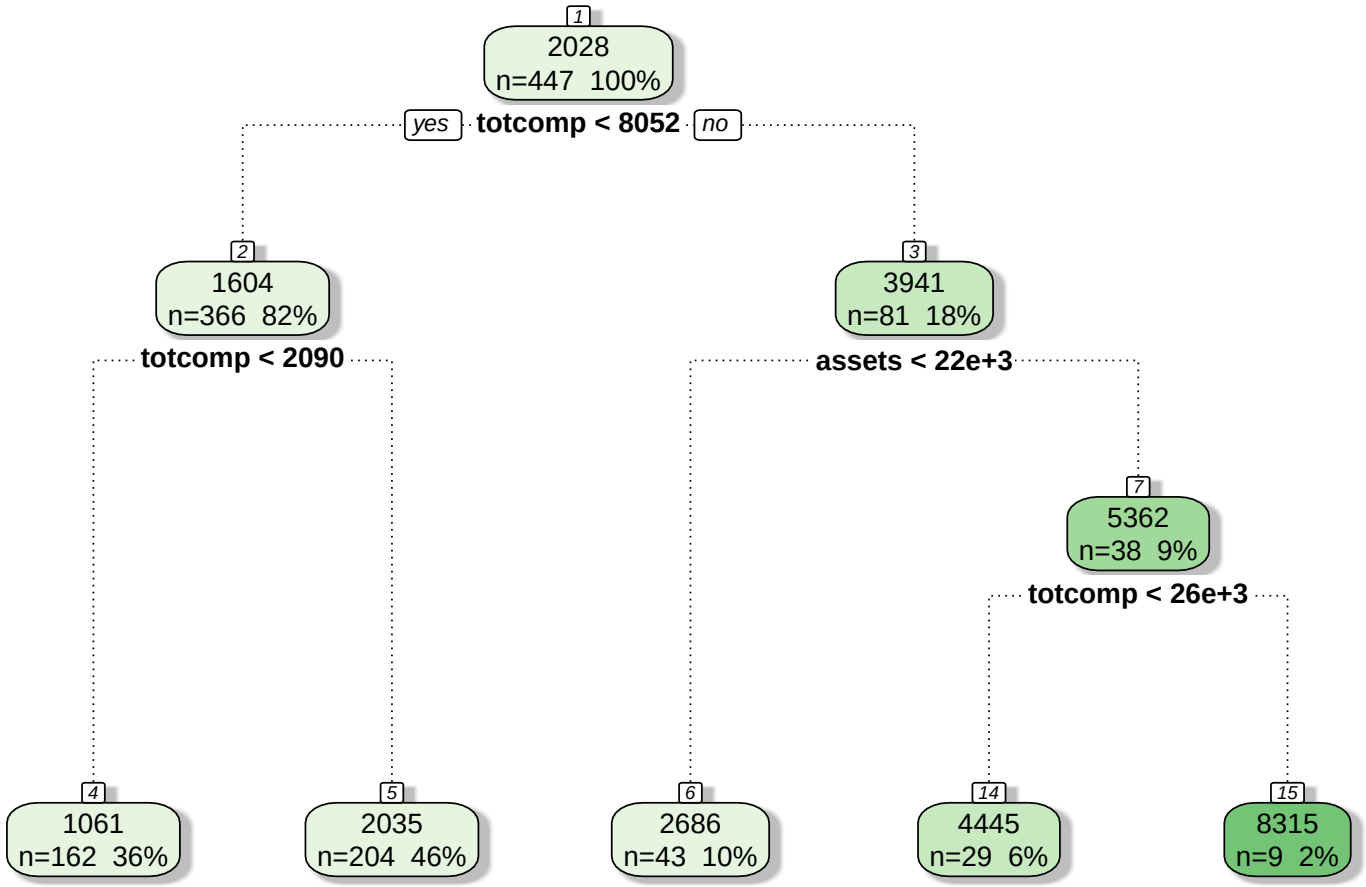
like $R^2 = 1 - 0{,}726 = \approx 0{,}27$.

the lowest error based on cross-validation

$\Rightarrow$ 4 splits.

22 splits $\Rightarrow R^2 \approx 0{,}63$.

based on training data $\Rightarrow$ decreasing with n splits.

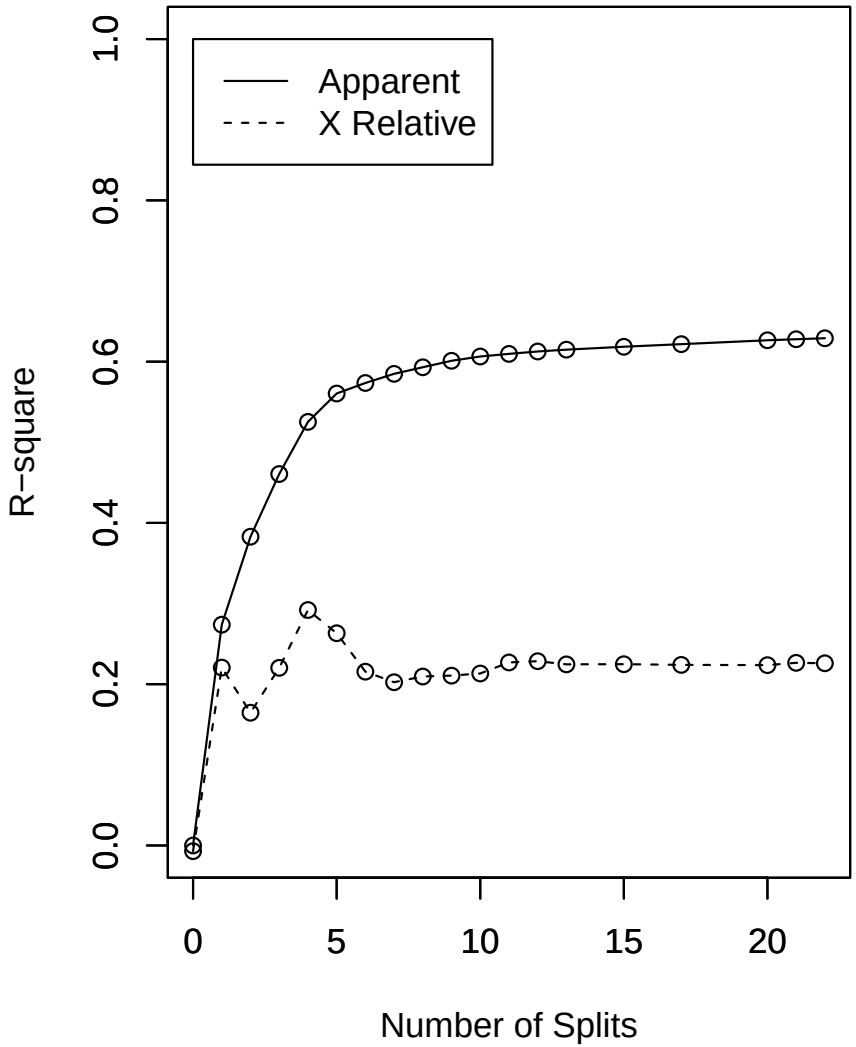


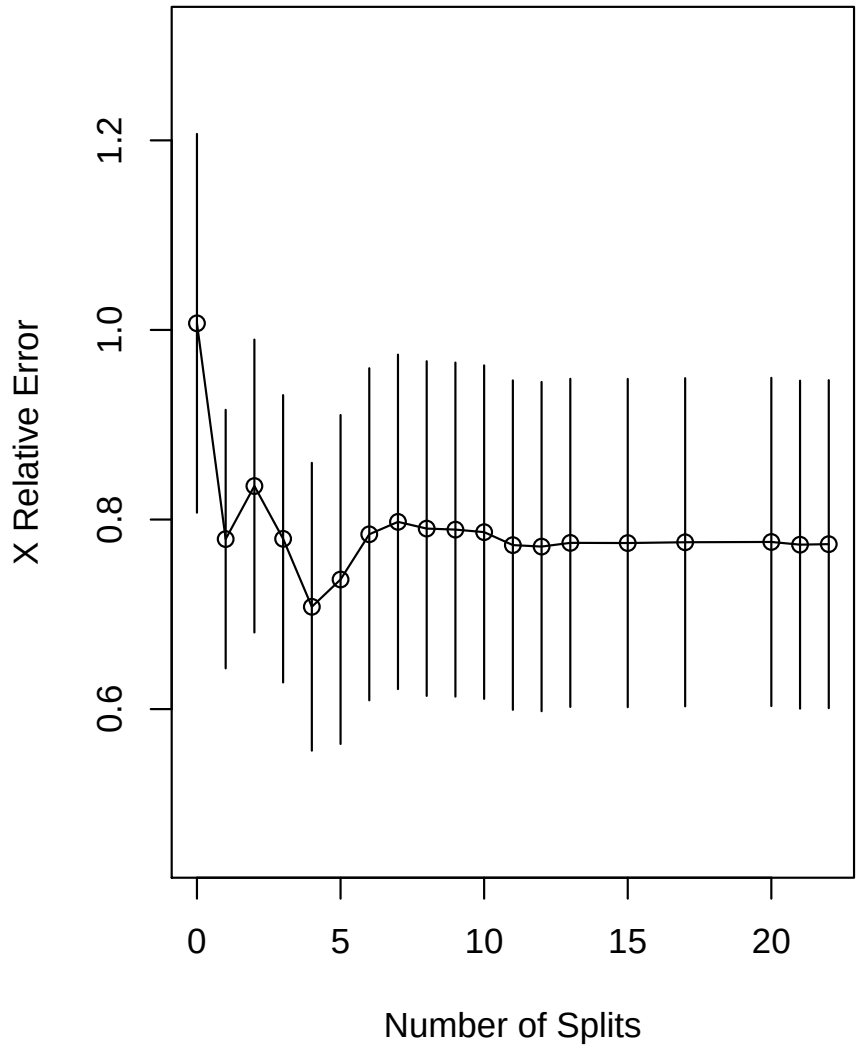
Q: How to choose the overall optimal α or subtree? $\rightsquigarrow$ cross-validation

- The sequence of trees T_0 (no splits) to T_m (m splits) uniquely determines the sequence of possible α's

$$\infty, \alpha_1, \ldots, \alpha_{m-1}, \alpha_{min}$$

- Any α between $(\alpha_i, \alpha_{i+1}]$ leads to the same optimal subtree
- Define $\beta_i = \sqrt{\alpha_i \alpha_{i+1}}$ as an "average" CP for every interval
- Split the data into B subsets $G_1, \ldots, G_B$ (10 by default)
  - For every subset excluding the G_i's determine $T_{\beta_1}, \ldots T_{\beta_m}$
  - Compute the relative MSE as the forecast loss for elements in G_i
- Compute the average loss over all G_i's and choose β (and thus the optimal subtree) which corresponds to the smallest one.





Problem 3: how to determine the importance of the variables and the marginal effects?

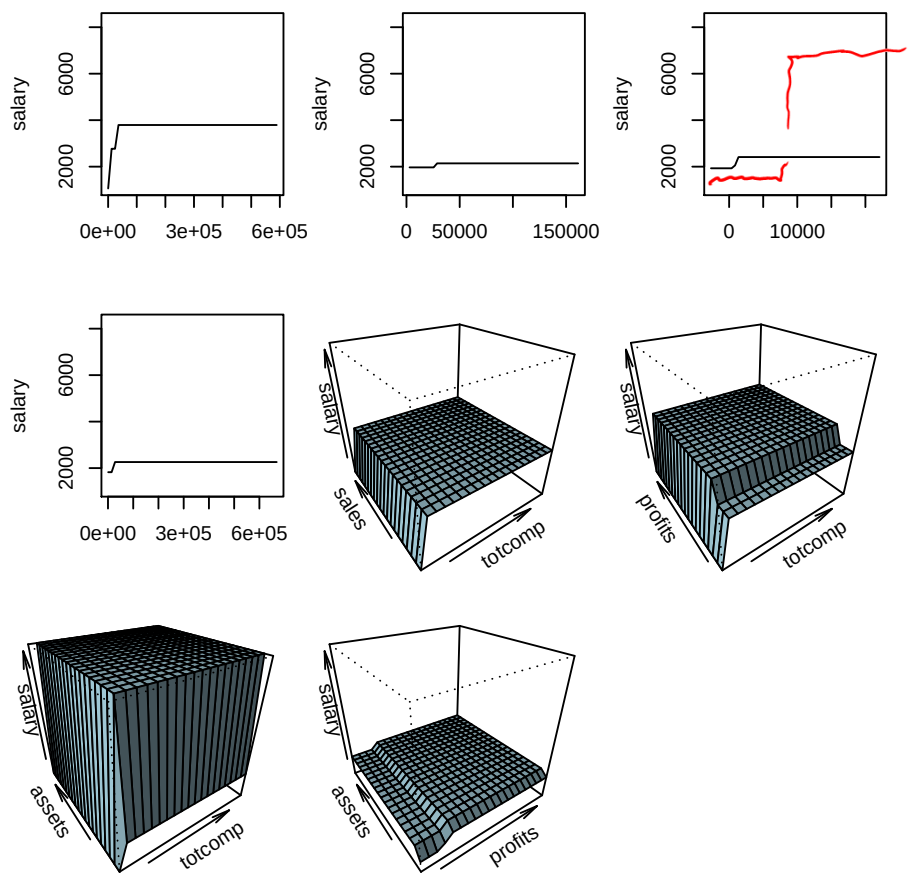
- The importance canNOT be determined via standardisation or via removing variables from the model.
- In `rpart` the importance is computed as the sum of improvement values over all nodes where the corresponding variable is used for splitting.

```
> cp.min = which.min(rpart.ceo$cptable[,4]);
> rpart.ceo.prune=prune(rpart.ceo, cp=rpart.ceo$cptable[cp.min,1])
> rpart.ceo.prune$variable.importance/sum(rpart.ceo.prune$variable.importance)
   totcomp     assets      sales    profits     tenure        age
0.52984007 0.18398873 0.10229360 0.09962347 0.05557637 0.02867777
```

The marginal effects can be visualized as follows

- Let $\hat{f}(\text{age}, \text{assets}, ..., \text{totcompt})$ be the function, which replicates the tree.
- Then $\hat{f}(x, \overline{\text{assets}}, ..., \overline{\text{totcomp}})$ is the marginal effect of `age`

```
> plotmo(rpart.ceo)
```

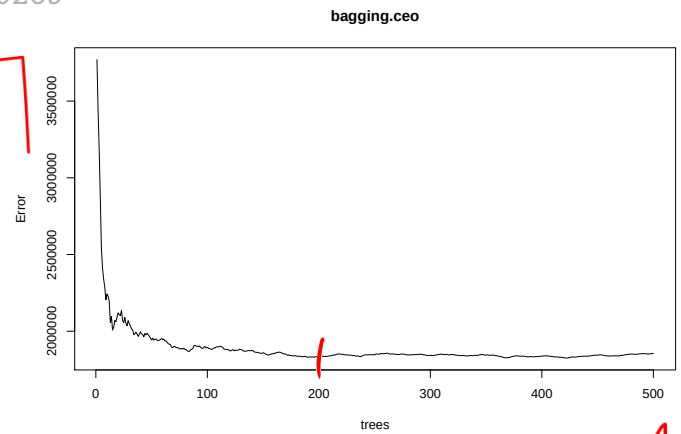


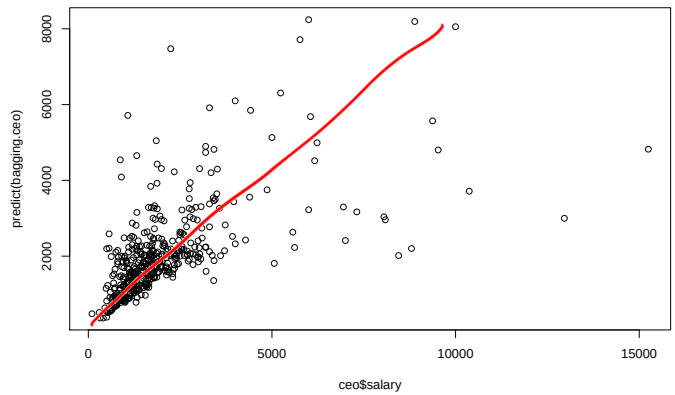
Problem 4: Generalizations

- Bagging: if you use for CART just a subsample, then you obtain a completely different tree.
  - Fit a CART to B random subsamples (*bootstrap*).
  - The error is measured on the remaining observations out-of-bag.
  - The final forecast is:

$$\hat{f}_{avr}(\boldsymbol{x}_0) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(\boldsymbol{x}_0).$$

```
> bagging.ceo= randomForest(salary ~ ., data=ceo, mtry=6)
> predict(bagging.ceo);
> cor(ceo$salary,predict(bagging.ceo))
[1] 0.6180269
```





Random Forests: is a generalization of *Bagging*

- For each splitting you consider not all explanatory variables bus just a subset of size $M \approx \sqrt{J}$
- ... this makes the trees more heterogenous and "uncorrelated" in terms of forecasts
- Each tree is grown on a bootstrap sample (as for bagging)
- The importance of a variable is measured by increase in (a) MSE ; (b) in node impurity over the out-of-bag sample if the variable is permuted

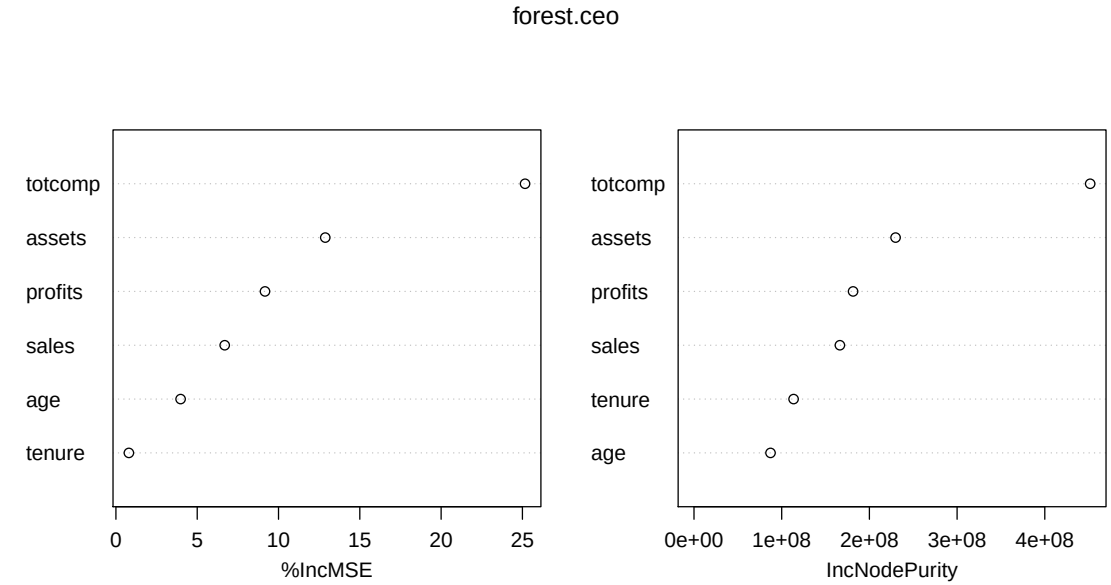
$$\Delta MSE_{j,b} = \frac{1}{|\overline{\mathcal{B}}_b|} \sum_{k \in \overline{\mathcal{B}}_b} \hat{u}^2(x_{1k}, \ldots, x_{Jk}) - \frac{1}{|\overline{\mathcal{B}}_b|} \sum_{k \in \overline{\mathcal{B}}_b} \tilde{\hat{u}}_k^2(x_{1k}, ..., x_{j-1,k}, \tilde{x}_{jk}, x_{j+1,k}, \ldots, x_{Jk}),$$

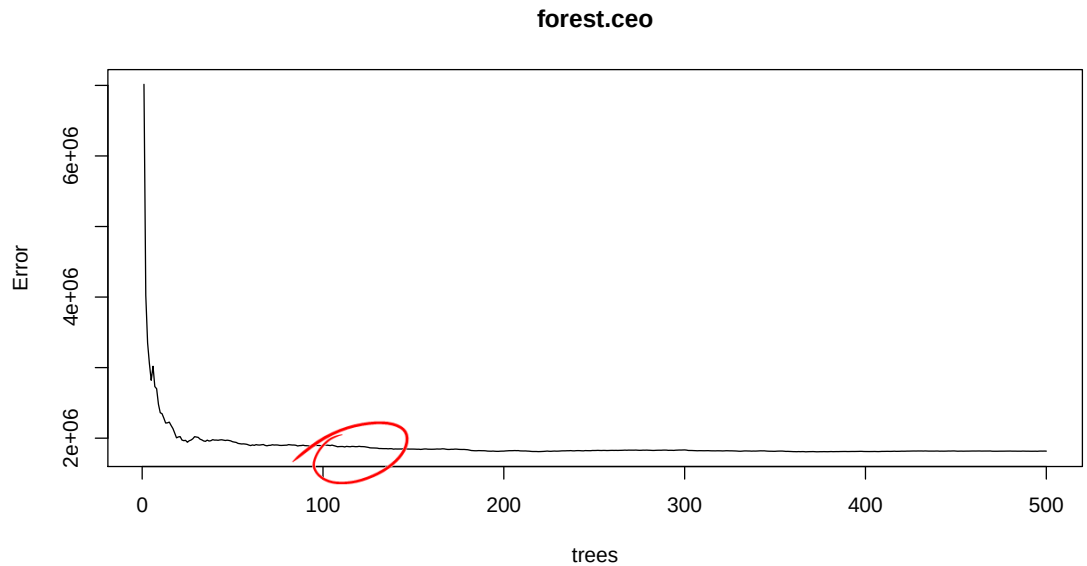
use original

use permuted

where $\tilde{\boldsymbol{x}}_j$ are the randomly permuted (reordered) observations on the jth variable and $\overline{\mathcal{B}}_b$ is the bth out-of-bag subsample.

```
> forest.ceo= randomForest(salary ~ ., data=ceo, importance=T)
> varImpPlot(forest.ceo)
```

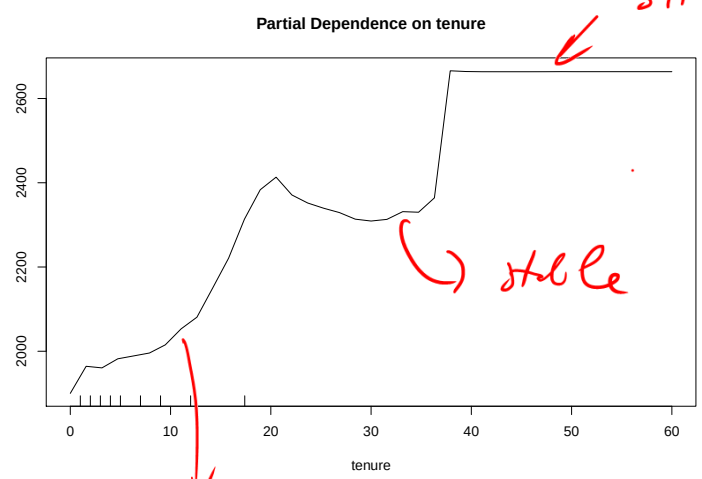




Partial dependence plots: visualize the marginal impact of a variable/feature

$$\tilde{f}_j(x) = \frac{1}{K}\sum_{k=1}^{K} \hat{f}(x_{1k}, ..., x_{j-1,k}, x, x_{j+1,k}, \dots, x_{Jk})$$

```
> partialPlot(forest.ceo, pred.data=ceo, x.var=tenure)
```



Regression trees: CHAID

- An alternative approach is CHAID (Chi-square Automatic Interaction Detectors): allows not only for binary splitting and is similar to ANOVA.
- Analysis is a generalization of two-sample test for the mean.
- Idea: let G be the number of splittings for variable X. We test if there is a significant difference between the means of Y in different regions.

Total sum of squares

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{g=1}^{G}\sum_{i:\boldsymbol{x}_i \in R_g}(y_i - \bar{y})^2$$

Within sum of squares

$$WSS = \sum_{g=1}^{G}\sum_{i:\boldsymbol{x}_i \in R_g}(y_i - \bar{y}_{R_g})^2$$

Between sum of squares

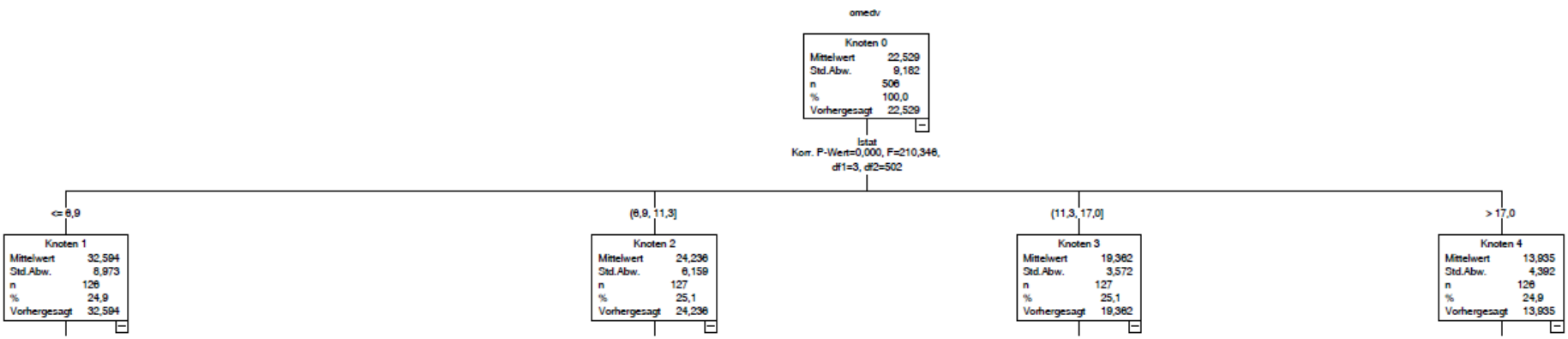
$$BSS = TSS - WSS = \sum_{g=1}^{G}|R_g|(\bar{y}_{R_g} - \bar{y})^2.$$

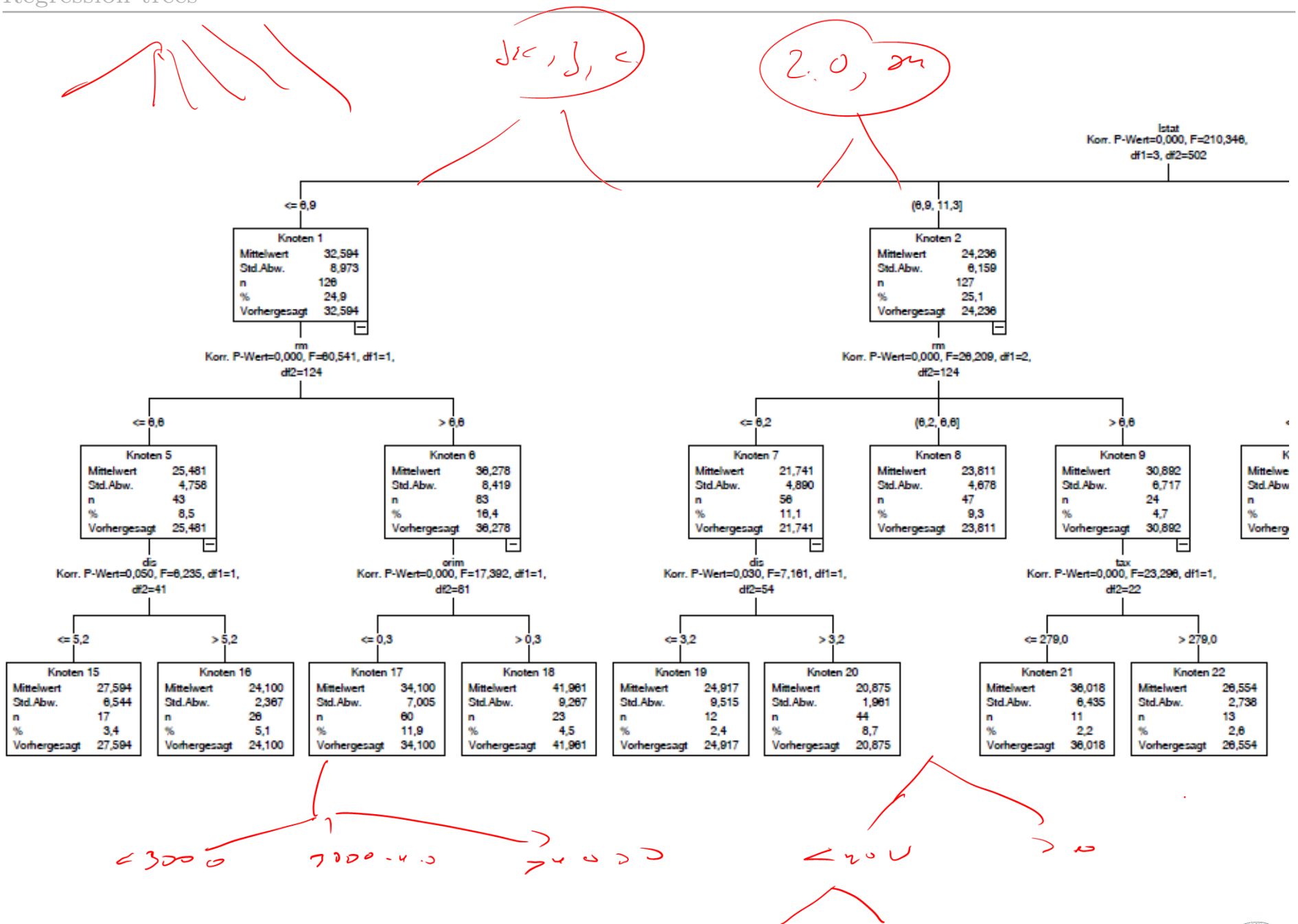
$$H_0 : \mu_1 = \cdots = \mu_G \quad \text{vs} \quad H_1 : \mu_i \neq \mu_j \text{ for at least one pair } i, j$$

$$\text{Test statistic:} \quad F = \frac{BSS/(G-1)}{WSS/(n-G)} \sim F_{G-1,n-G}$$

Idea:

- For each predictor we determine the optimal splitting, i.e. the regions with the smallest p-value of the test.
- The p-values should be corrected due to multiple testing (Bonferroni correction).
- The predictor with the smallest corrected p-value is used for splitting.





Baumtabelle

Knoten	Mittelwert	Standardabweichung	N	Prozent	Vorhergesager Mittelwert	Übergeordneter Knoten	Primäre unabhängige Variable					
							Variable	Sig.[a]	F	df1	df2	Werte aufteilen
0	22,53	9,182	506	100,0%	22,53							
1	32,59	8,973	126	24,9%	32,59	0	lstat	,000	210,346	3	502	<= 6,9
2	24,24	6,159	127	25,1%	24,24	0	lstat	,000	210,346	3	502	(6,9, 11,3]
3	19,36	3,572	127	25,1%	19,36	0	lstat	,000	210,346	3	502	(11,3, 17,0]
4	13,93	4,392	126	24,9%	13,93	0	lstat	,000	210,346	3	502	> 17,0
5	25,48	4,758	43	8,5%	25,48	1	rm	,000	60,541	1	124	<= 6,6
6	36,28	8,419	83	16,4%	36,28	1	rm	,000	60,541	1	124	> 6,6
7	21,74	4,890	56	11,1%	21,74	2	rm	,000	26,209	2	124	<= 6,2
8	23,81	4,678	47	9,3%	23,81	2	rm	,000	26,209	2	124	(6,2, 6,6]
9	30,89	6,717	24	4,7%	30,89	2	rm	,000	26,209	2	124	> 6,6
10	22,59	3,393	17	3,4%	22,59	3	tax	,000	18,192	1	125	<= 279,0
11	18,86	3,345	110	21,7%	18,86	3	tax	,000	18,192	1	125	> 279,0
12	19,31	2,881	14	2,8%	19,31	4	nox	,000	36,153	2	123	<= ,5
13	15,72	3,623	44	8,7%	15,72	4	nox	,000	36,153	2	123	(,5, ,6]
14	11,67	3,554	68	13,4%	11,67	4	nox	,000	36,153	2	123	> ,6
15	27,59	6,544	17	3,4%	27,59	5	dis	,050	6,235	1	41	<= 5,2
16	24,10	2,367	26	5,1%	24,10	5	dis	,050	6,235	1	41	> 5,2
17	34,10	7,005	60	11,9%	34,10	6	crim	,000	17,392	1	81	<= ,3
18	41,96	9,267	23	4,5%	41,96	6	crim	,000	17,392	1	81	> ,3
19	24,92	9,515	12	2,4%	24,92	7	dis	,030	7,161	1	54	<= 3,2
20	20,88	1,961	44	8,7%	20,88	7	dis	,030	7,161	1	54	> 3,2
21	36,02	6,435	11	2,2%	36,02	9	tax	,000	23,296	1	22	<= 279,0
22	26,55	2,738	13	2,6%	26,55	9	tax	,000	23,296	1	22	> 279,0
23	19,51	2,847	75	14,8%	19,51	11	nox	,007	9,596	1	108	<= ,6
24	17,47	3,911	35	6,9%	17,47	11	nox	,007	9,596	1	108	> ,6
25	17,24	3,698	24	4,7%	17,24	13	rad	,022	11,566	1	42	2,0; 5,0; 6,0; 24
26	13,90	2,596	20	4,0%	13,90	13	rad	,022	11,566	1	42	4,0
27	10,85	3,087	56	11,1%	10,85	14	dis	,000	22,769	1	66	<= 2,1
28	15,53	3,094	12	2,4%	15,53	14	dis	,000	22,769	1	66	> 2,1