Regression trees

Myra O'Regan

October 26, 2017

Regression trees

- What do we do when we have a continuous output variable?
- What is going to change?
- There are again a few choices.
- We can always define our own splitting criteria
- We are going to look at a very simple situation
- We are going to look at sums of squares approach called anova in the package **rpart**

For One Node t

- We want to find best split
- Again look at all splits and each split produces two nodes A and B
- We use the same idea as before
- For split s at node t
- $\Delta(t, s) = i(t) - i(t_A) - i(t_B)$
- $i(t) = \sum_{i \in t}(y_i - \bar{y}_t)^2$
- Often called Sums of Squares

ANOVA view

- Here we look at SUMS of squares $SS_t - SS_A - SS_B$
- Want to maximise this
- SS_t is calculated on parent mode
- Called deviance in **rpart** output
- Wonderful **rpart** prints out improvement
- Improvement $= \frac{\Delta(t,s)}{i(t)}$
- MSE (Mean Squared Error) for node t

$$= \frac{SS_t}{no.\ of\ obs.\ in\ node\ t}$$

Calculations for animal sleep data

- Root Node: deviance = 1624.066 = 19.57*83 (MSE*n)
- $\Delta(t, s) = 1624.066 - 9 * (0.874) - 74 * 15.26 = 1129.24$
- R Improvement $= \frac{\Delta(t,s)}{i(t)} = \frac{1129.24}{1624.066} = 0.2998$
- $\alpha = \frac{R(t) - R(T_t)}{|T| - 1}$
- $cp = \frac{\alpha}{R(0)}$

Problems

- Outliers
- Scale of data

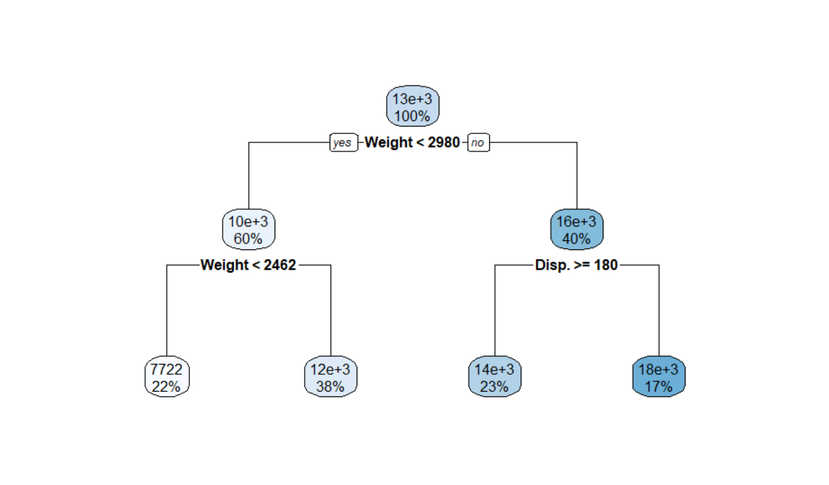
Choosing a tree

- We do the same as before
- We choose the cp value at the minimum xerror value
- Again we can look at values corresponding to min(xerror) - 1 xstd
- How do we calculate relative error (relative to root node)
- We calculate

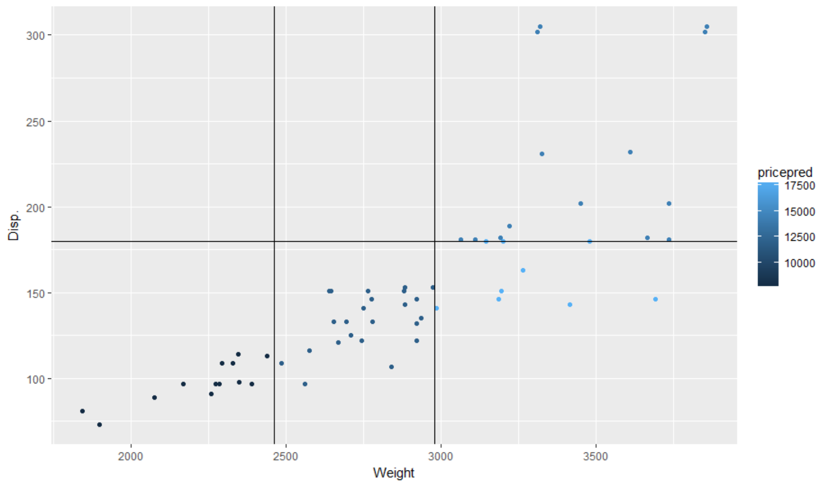
$$\sum_{terminal \ \ nodes} \sum_{i \in t} (y_i - \bar{y}_t)^2$$

- Divide by the SS for the root node.
- It could also be thought of as 1-R^2

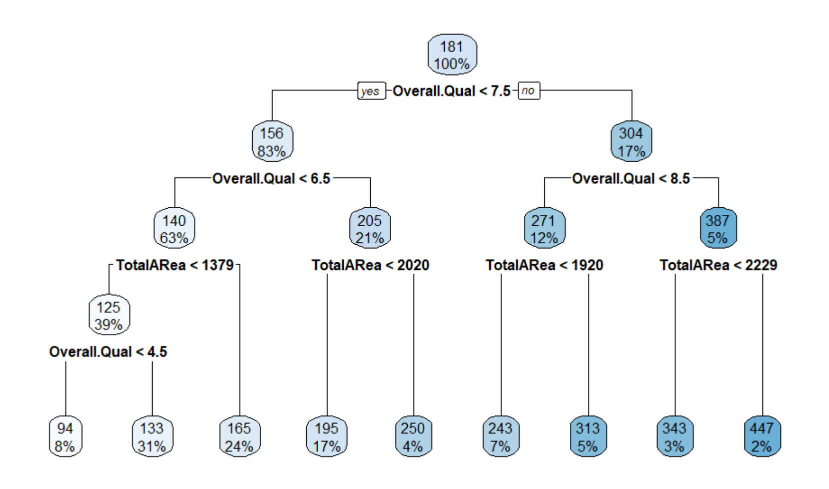
Regression Tree



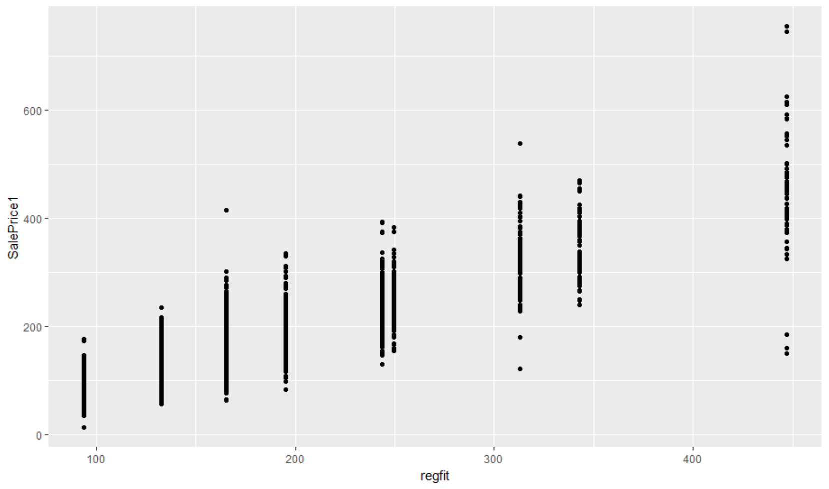
Regression Tree



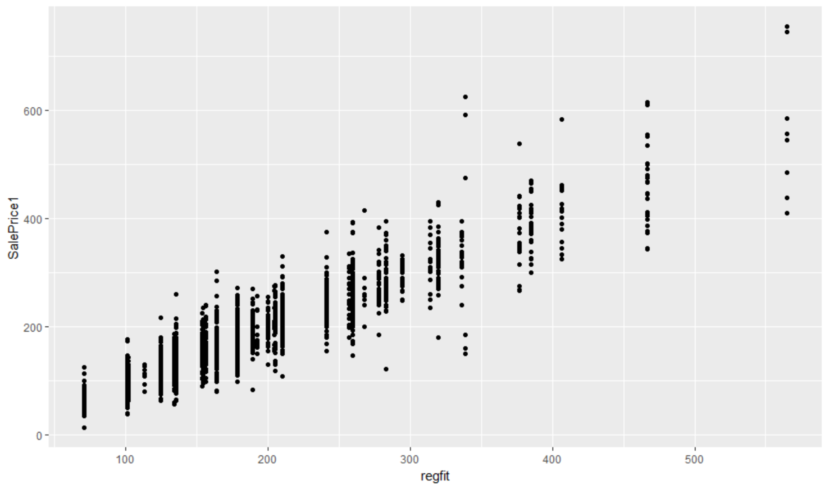
Regression Tree-Another example



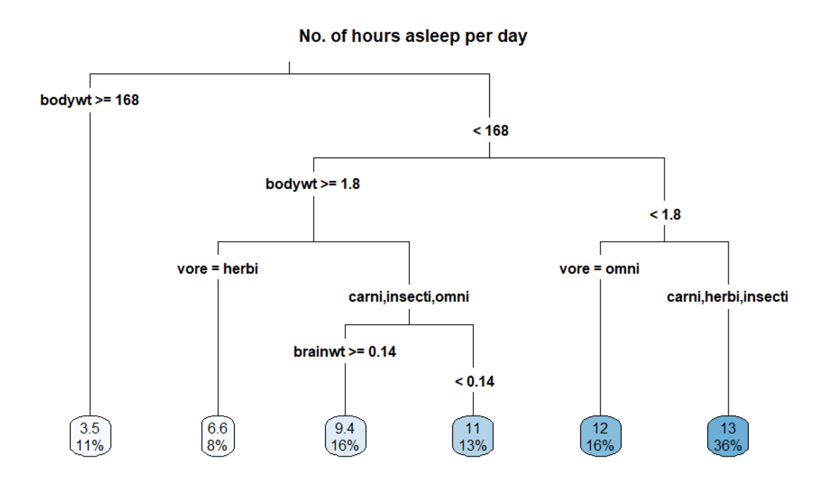
How good a tree?



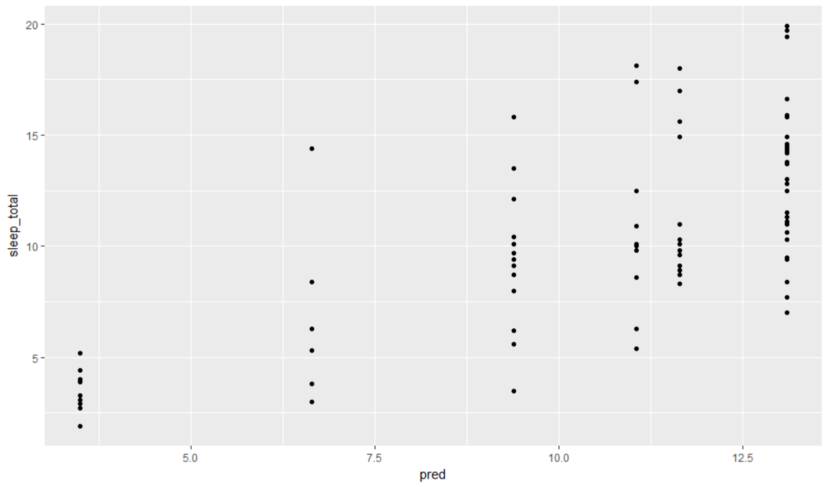
A better tree?



Regression Tree-Sleepy animals



Regression Tree-Sleepy animals



How well does the model fit?

- Plot predicted P_i vs Observed O_i for test set
- Calculate correlation r between them
- r^2 is the R^2 we all know about.
- Calculate root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}$$

- Same units as original predicted variable
- Absolute measure of fit
- Look for large differences between P_i and O_i