Regression Trees

Predicting a continuous target attribute

Ricco RAKOTOMALALA

Modeling the relationship between Y and X_1, X_2, …, X_J

Determine a mathematical function which enables to explain/predict as well as possible the values of Y from X_1, X_2, …

$$Y = f\left(X_1, \ldots, X_J\right) + \varepsilon$$

Response /target variable
Continuous

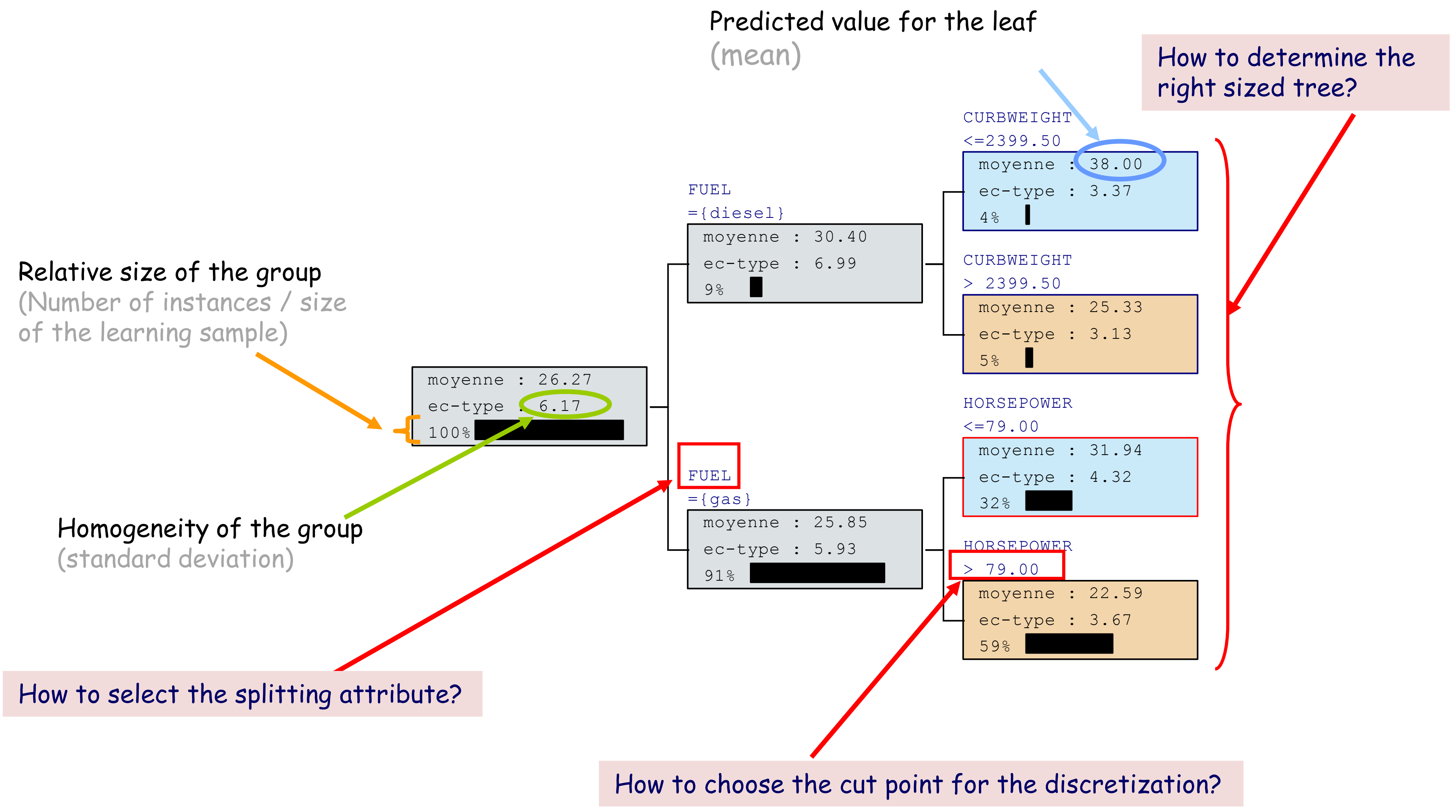
Input variables, descriptors, …
Continuous or/and discrete

Error: the part of Y which is not explained by the function

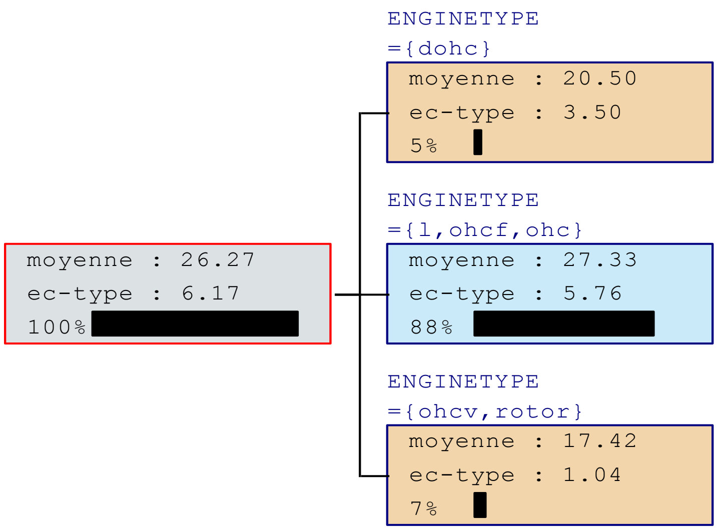
We must:
(1) Define the mathematical function f()
(2) Estimate the parameters of the function f() from a sample (learning set)
(3) Choose an evaluation criterion which enables to establish the quality of the model

A possible solution: REGRESSION TREE
(1) Model: prediction tree, that we can transform in a set of rules
(2) Partition the space into regions as homogenous as possible regarding Y
(3) Least squares approach i.e. minimizing the sum of squared residuals

Example of a regression tree: main features and issues



Splitting criterion



The splitting attribute

(1) Makes the conditional means as different as possible between groups.

Or (this is the same thing)

(2) Makes the variance (or the standard deviation) within the groups the smallest.

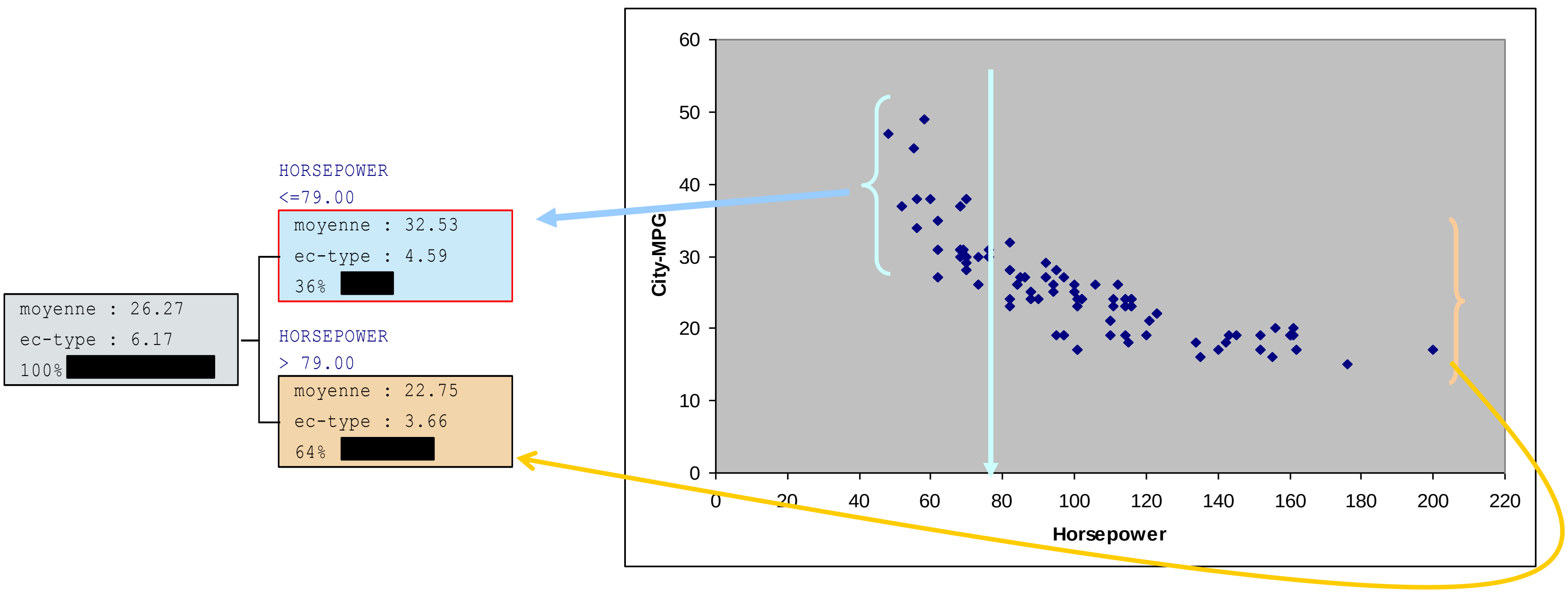
Variance decomposition: TSS = BSS + WSS

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Sum of squares total}} = \underbrace{\sum_{l=1}^{L} n_l(\bar{y}_l - \bar{y})^2}_{\text{SS. Between Groups}} + \underbrace{\sum_{l=1}^{L}\sum_{i=1}^{n_l}(y_{il} - \bar{y}_l)^2}_{\text{SS. Within Groups}}$$

Splitting attribute selection

$$X_{j^*} = \arg\max_j BSS(X_j)$$

Determining the "best" cut point for continuous attribute



Selecting the cut point that maximizes the BSS

$$BSS(X) = n_1 \times (\bar{y}_1 - \bar{y})^2 + n_2 \times (\bar{y}_2 - \bar{y})^2$$

Or, equivalently

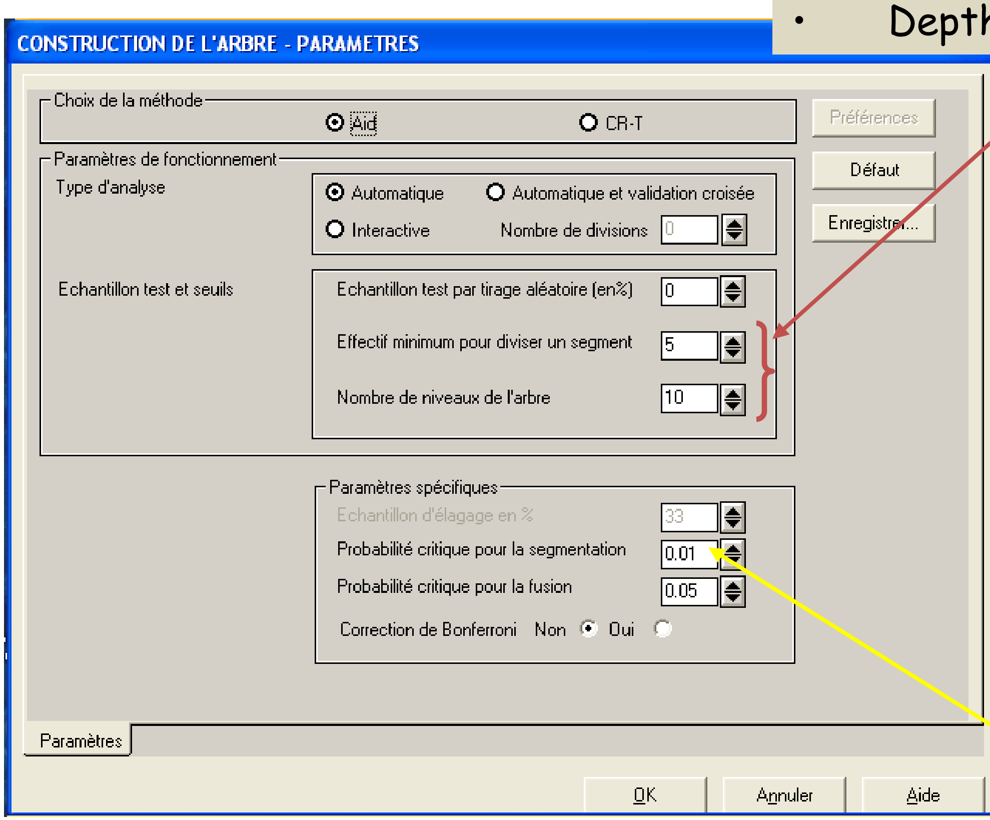
$$BSS(X) = \frac{n_1 \times n_2}{n_1 + n_2} \times (\bar{y}_1 - \bar{y}_2)^2$$

⇒ The Ward's minimum variance criterion

Determining the right sized tree
Pre-pruning (variant of AID)

Empirical stopping rule

- Size of nodes (support criterion)
- Depth of the tree



Statistical criterion (AID) : The significance test for the ANOVA
i.e. H0 : The conditional means are the same whatever the group
If p-value of the test is lower than a predefined threshold, the split is performed

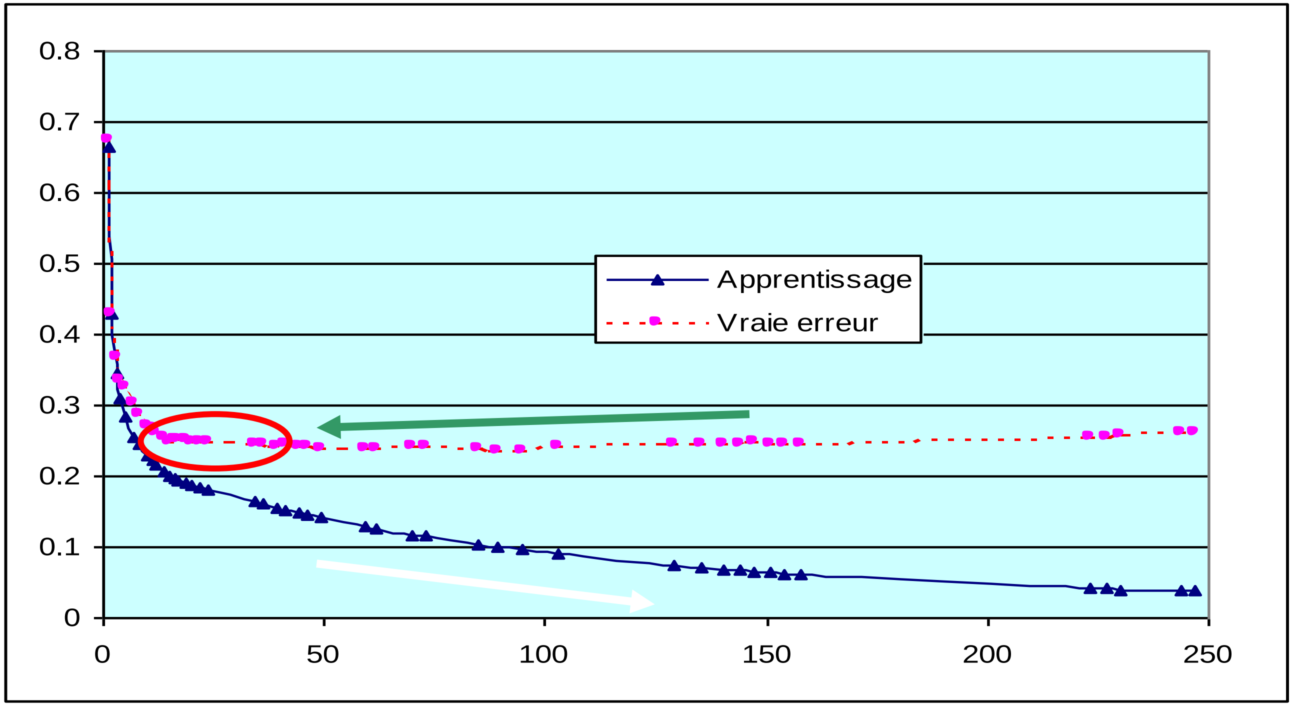
Post-pruning (CART)

Prediction from the leaf for which belongs the instance

Two steps in the learning process

(1) Growing → maximizing the homogeneity of the groups

(2) (Post) pruning → minimizing the sum of squares residuals ➔ $E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$



The post-pruning strategy is the same than for the classification tree algorithm:

- Defining a sequence of trees with the same complexity cost
- Choosing the one which minimizes the sum of squares residuals on the pruning set
- Possibly, applying the preference to the simplicity via the 1-SE rule

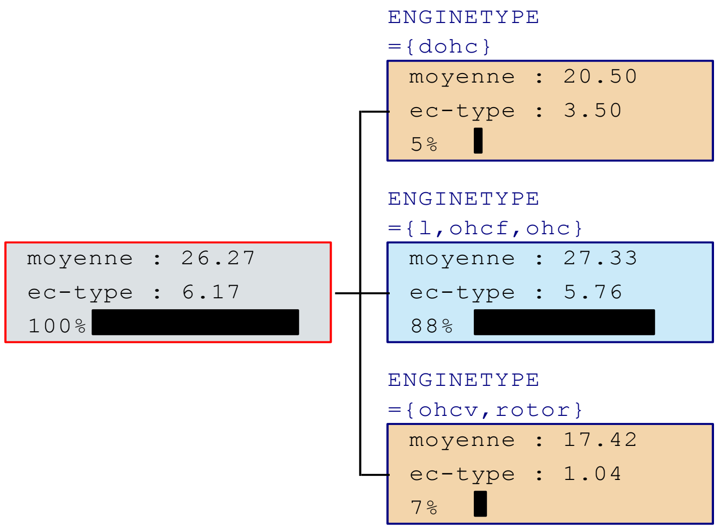
Merging the leaves from the same parent node (to avoid the data fragmentation in the multiway splits)

Two different approaches according CART and AID

(1) CART : always binary tree → Finding the binary gathering which maximizes the BSS

(2) AID : merging in m "best" leaves → Merging the leaves for which the conditional means are not significantly different

- Merging the leaves for which the means are the most similar
- Continue until there are pair of means which are not significantly different according the significance level alpha.



No merging

Merging of 3 leaves

Merging of 2 leaves

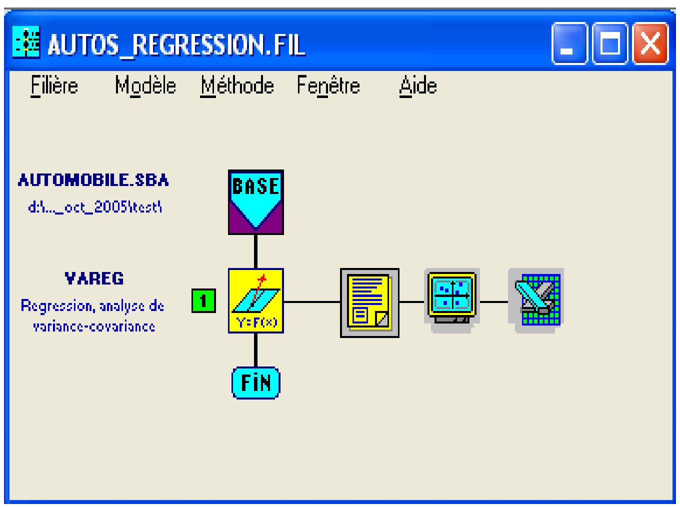
Linear regression

An alternative approach for predicting continuous target attribute

MULTIPLE LINEAR REGRESSION

(1) Linear combination of input variables
(2) Least squares estimation
(3) Minimizing the sum of squares residuals

$$Y = a_0 + a_1 X_1 + \cdots + a_J X_J + \varepsilon$$

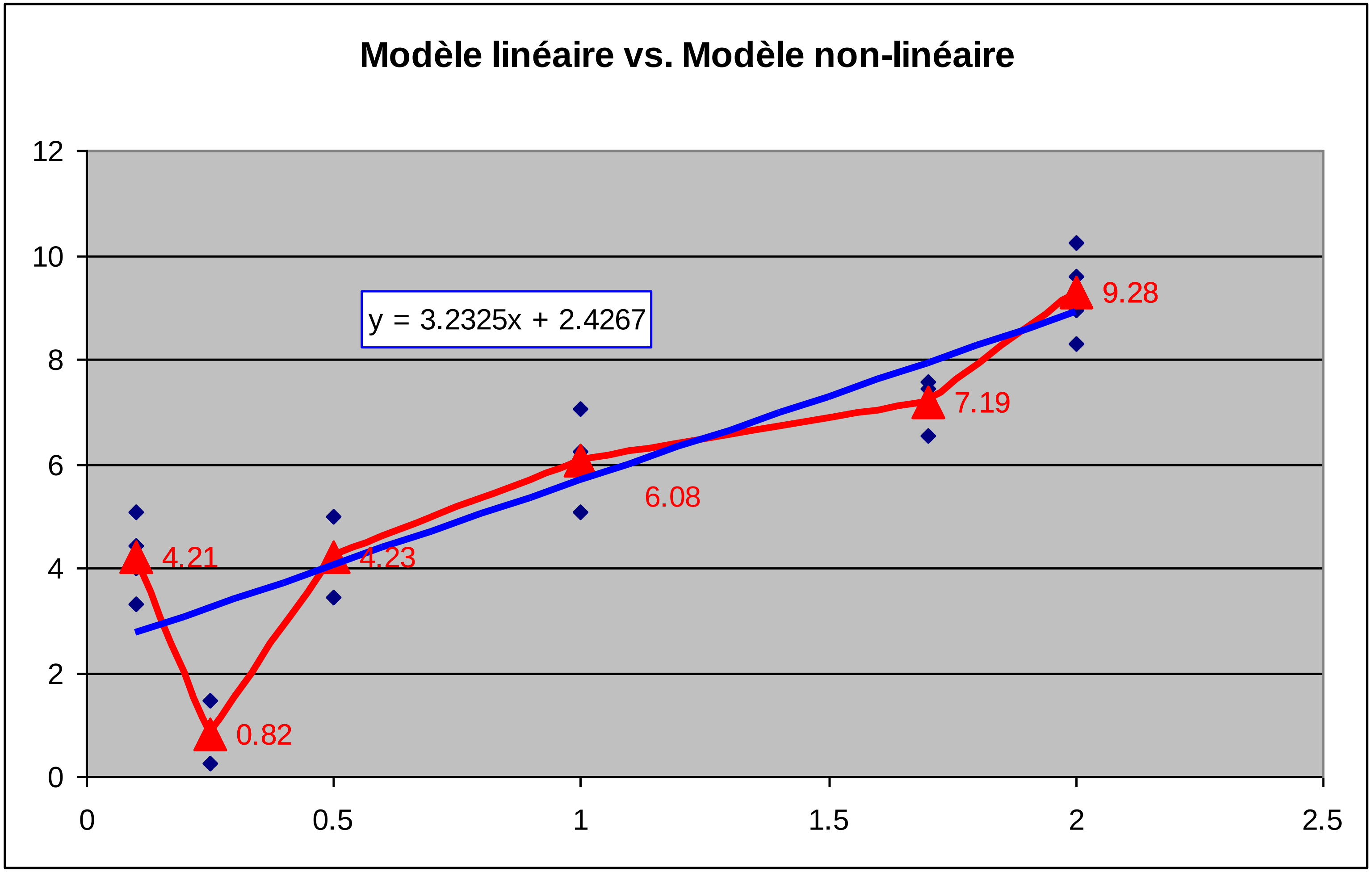


Regression coefficients — Significance of reg. coef.

```
ESTIMATION / COEFFICIENTS
AJUSTEMENT DES MOINDRES CARRES (AVEC TERME CONSTANT)
  164 INDIVIDUS,  4 PARAMETRES (CONSTANTE EN QUEUE).
  IDEN   LIBELLE               COEFFICIENT  ECART-TYPE STUDENT  PROBA. V.TEST
                                                         160
CRITERE(S)
  C13  - CURBWEIGHT              -0.0066      0.001    5.793   0.000  -5.51

  C16  - ENIGNESIZE               0.0829      0.017    4.875   0.000   4.70

  C18  - HORSEPOWER              -0.1519      0.014   10.975   0.000  -9.46

         CONSTANTE               47.3431      1.389   34.087   0.000  18.36
TEST D'AJUSTEMENT GLOBAL
SOMME DES CARRES DES ECARTS ........... SCE =        1632.5071
COEFFICIENT DE CORRELATION MULTIPLE ... R   =      0.8596    R2 =       0.7389
VARIANCE ESTIMEE DES RESIDUS ...... S2  =      10.2032    S  =       3.1942
TEST DE NULLITE SIMULTANEE DES COEFFICIENTS DES  3 VARIABLES :
   FISHER =   150.924          DEG.LIB =  3  160
   P.CRIT =    0.0000          V.TEST  =   14.26
```

Global significance of the regression

Comparison between nonlinear and linear regressions



Conclusion

Prediction performance

In practice, the regression trees are not (always) better than standard linear models.

As an exploration tool

The trees are interesting because they enable to identify 'areas' where observations are homogeneous according the target attribute Y. Then, we can make a local estimation of the distribution parameters of Y.

Reference

Breiman, Friedman, Olshen and Stone – « Classification and Regression Trees », Chapman & Hall, 1984.