Arbres de Régression

Prédiction d'une variable continue

Ricco RAKOTOMALALA

Principes de la régression

Prédiction d'une variable continue à partir d'une série de variables de type quelconque

$$Y = f(X_1, \ldots, X_J)$$

Variable à prédire, endogène
Forcément continue

Variable prédictives, enxogènes, descripteurs
Continues et/ou discrètes

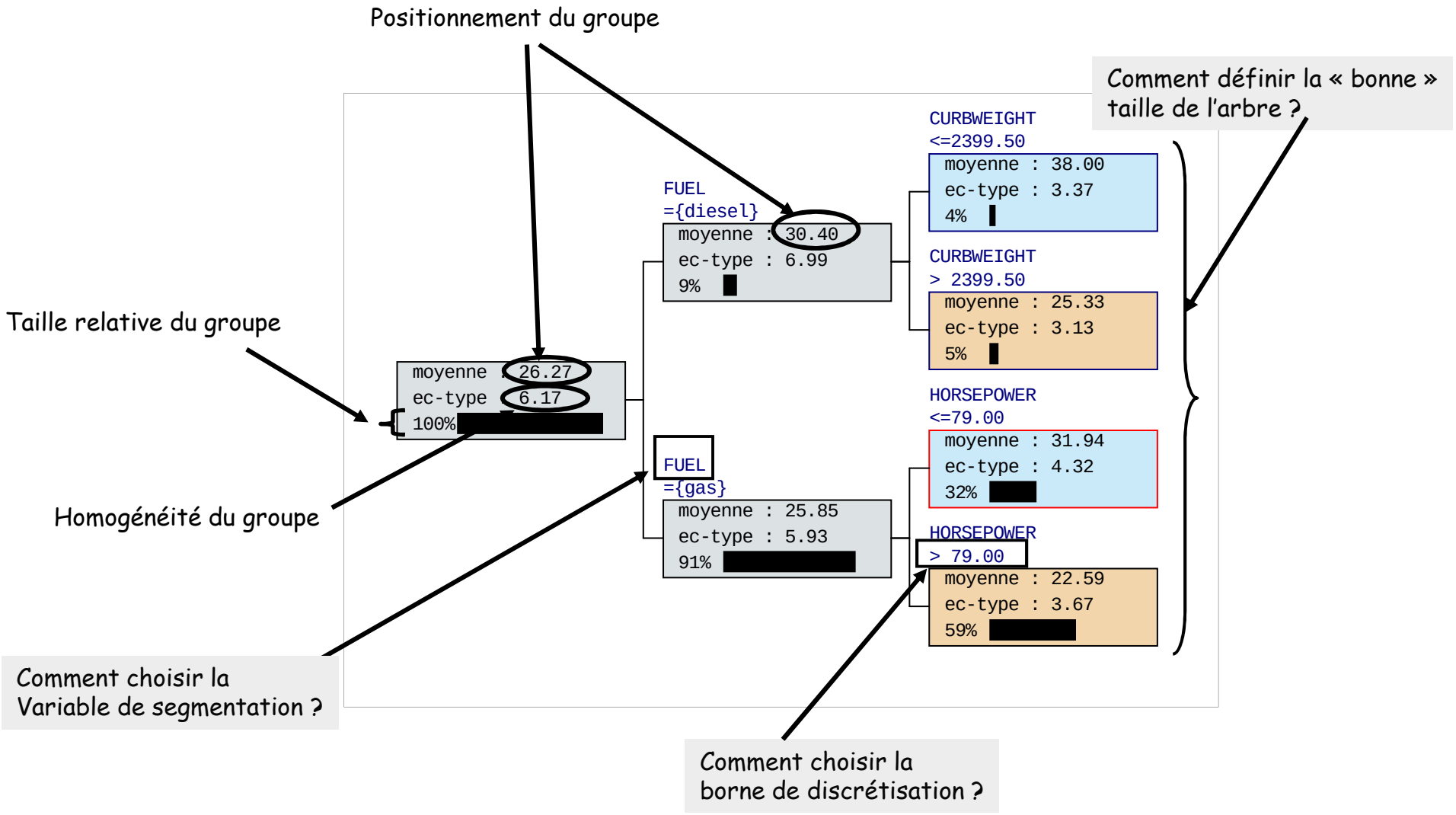
A définir :
(1) La forme de la fonction f
(2) L'estimation de ses paramètres à partir de l'échantillon d'apprentissage
(3) Le critère d'évaluation de la qualité de l'estimation

→

Solution : ARBRE DE REGRESSION
(1) Un arbre logique
(2) Segmentation de manière à obtenir des groupes « purs » sur Y
(3) Critère des moindres carrés

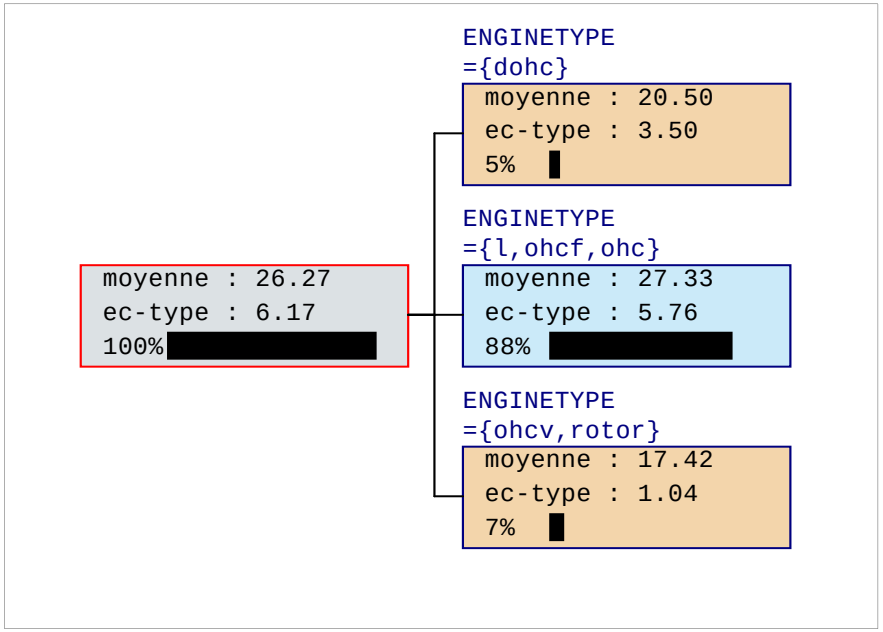
Arbres de régression

Structure générale et éléments clés



Arbres de régression

Critère pour la segmentation – L'équation d'ANOVA



Choisir la segmentation de manière à ce que

(1) Les moyennes soient le plus disparates possibles entre les groupes

ou (de manière équivalente)

(2) Les valeurs soient le plus proches possibles dans les groupes

Équation d'analyse de variance : TSS = BSS + WSS

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{l=1}^{L} n_l (\overline{y}_l - \overline{y})^2 + \sum_{l=1}^{L} \sum_{i=1}^{n_l} (y_{il} - \overline{y}_l)^2$$

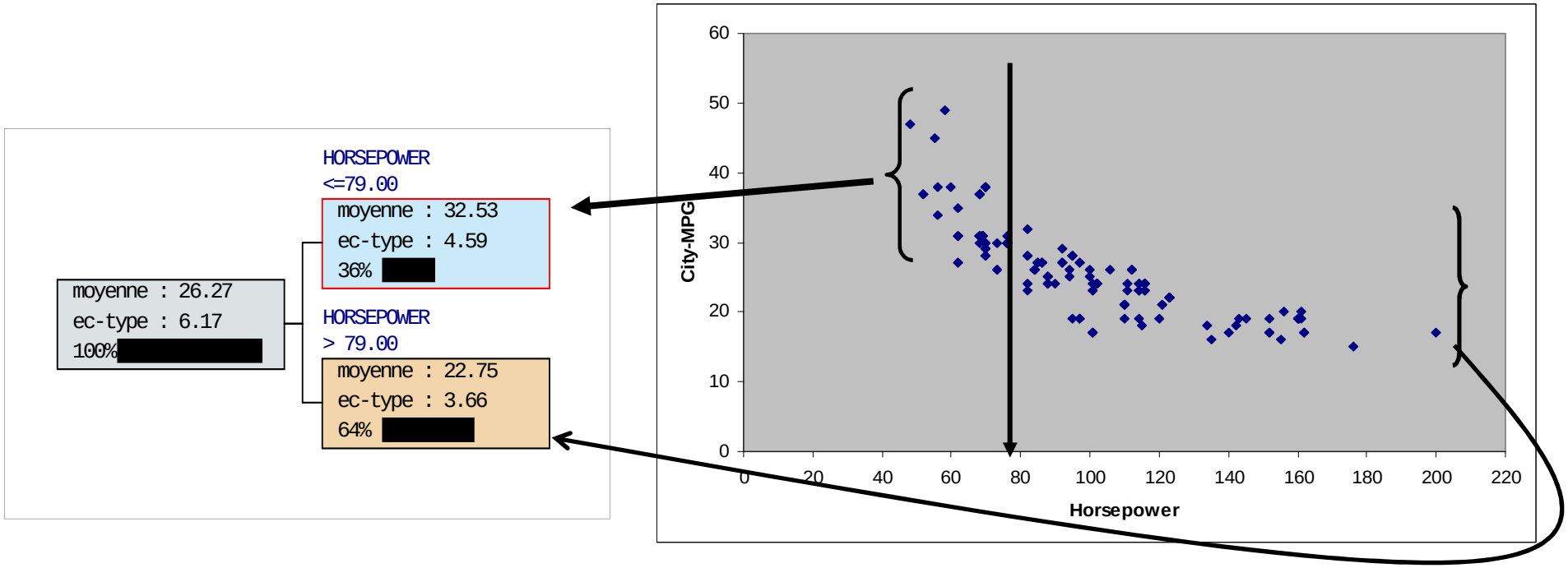
n x V.Totale n x V. Inter-classes n x V. Intra-classes

Choix de la variable de segmentation

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

Arbres de régression

Traitement des variables continues



Trouver le point de coupure (discrétisation) sur X tel que BSS est maximum

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

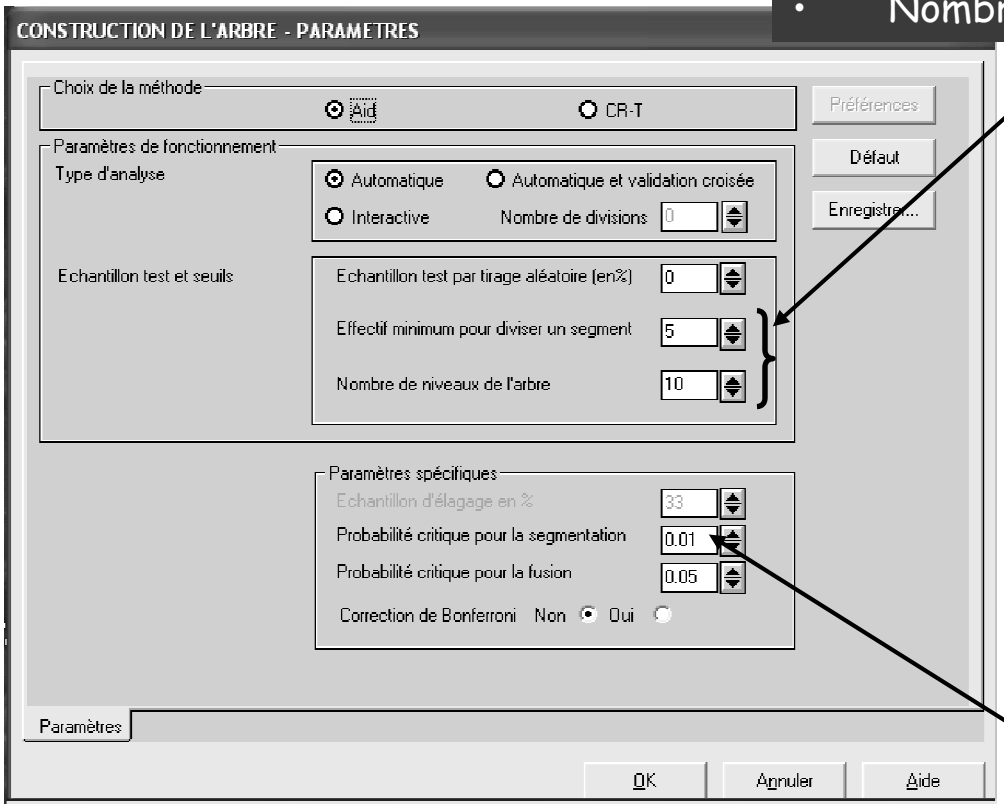
Ou, de manière équivalente

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

Arbres de régression

Règles d'arrêt – Pre-pruning

Critères empiriques pour contrôler la taille de l'arbre
- Effectif minimum pour segmenter
- Nombre de niveaux de l'arbre



Critère statistique (AID) : probabilité critique pour la segmentation
Si p-value de l'ANOVA est inférieure au seuil, on segmente

Arbres de régression

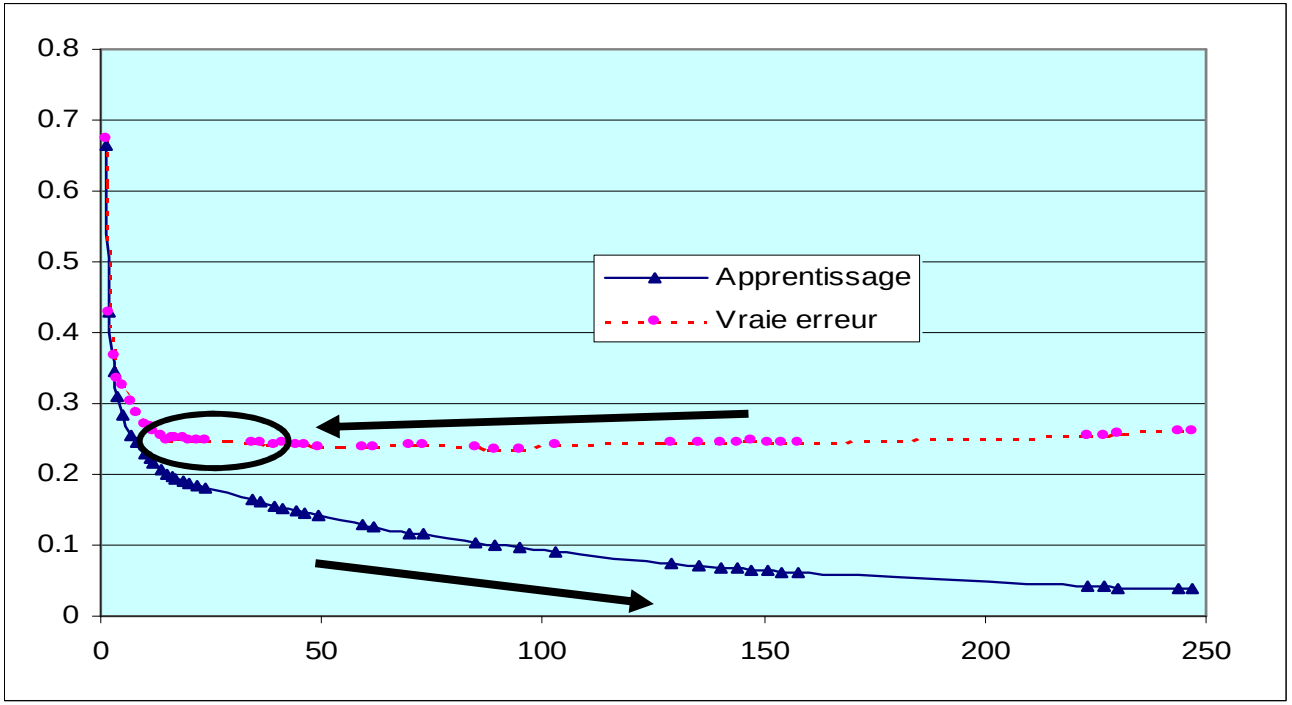
Post-pruning avec CR-T

Apprentissage en deux phases

(1) Expansion [growing] → maximiser l'homogénéité des groupes

(2) Elagage [pruning] → minimiser l'erreur de prédiction au sens des moindres carrés

$$E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$



La stratégie de l'élagage est la même que pour la discrimination :

- Définir une séquence d'arbres de coût-complexité équivalents
- Choisir dans la séquence, celle qui minimise l'erreur sur un fichier d'élagage
- *Éventuellement, donner une préférence à la simplicité en introduisant la règle de l'écart-type*

Arbres de régression

Regroupement des feuilles issues d'une segmentation

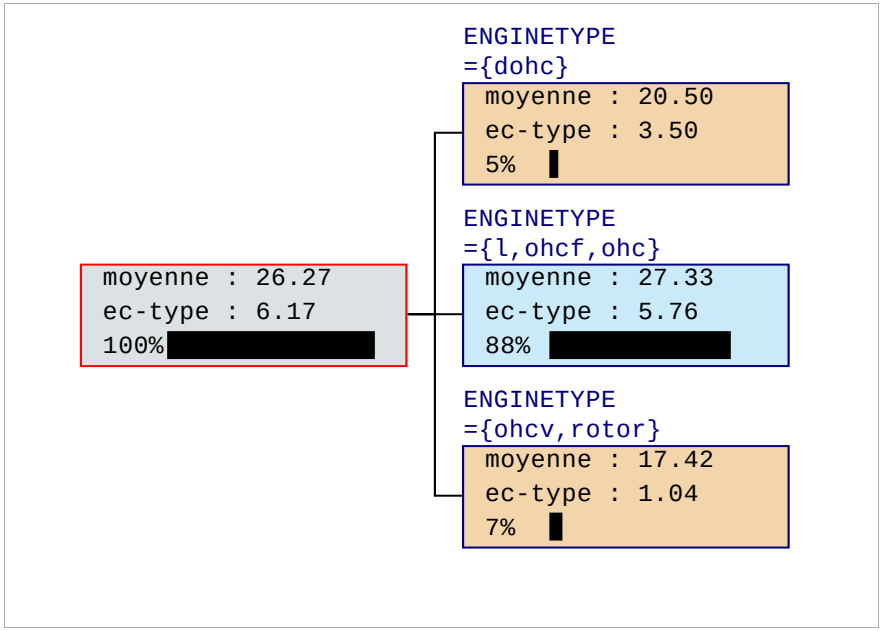
2 approches différentes selon C-RT et AID

(1) C-RT : arbre toujours binaire → trouver le regroupement qui maximise BSS

(2) AID : arbre m-aire → regrouper les feuilles très proches au sens de Y

- On fusionne les 2 feuilles les plus proches (comparaison de moyennes – test de Student)
- On réitère l'opération tant que la p-value est supérieure à la probabilité critique pour la fusion

Remarque : il est tout à fait possible que toutes les feuilles soient regroupées en une feuille unique



Pas de fusion

Fusion de 3 feuilles

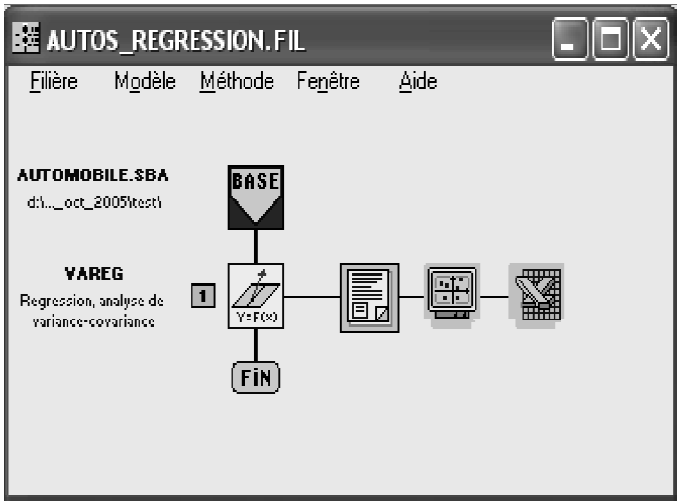
Fusion de 2 feuilles

Régression linéaire multiple

Une technique alternative

Solution : REGRESSION LINEAIRE MULTIPLE
(1) Une combinaison linéaire des variables exogènes
(2) Méthodes des moindres carrés
(3) Critère des moindres carrés

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

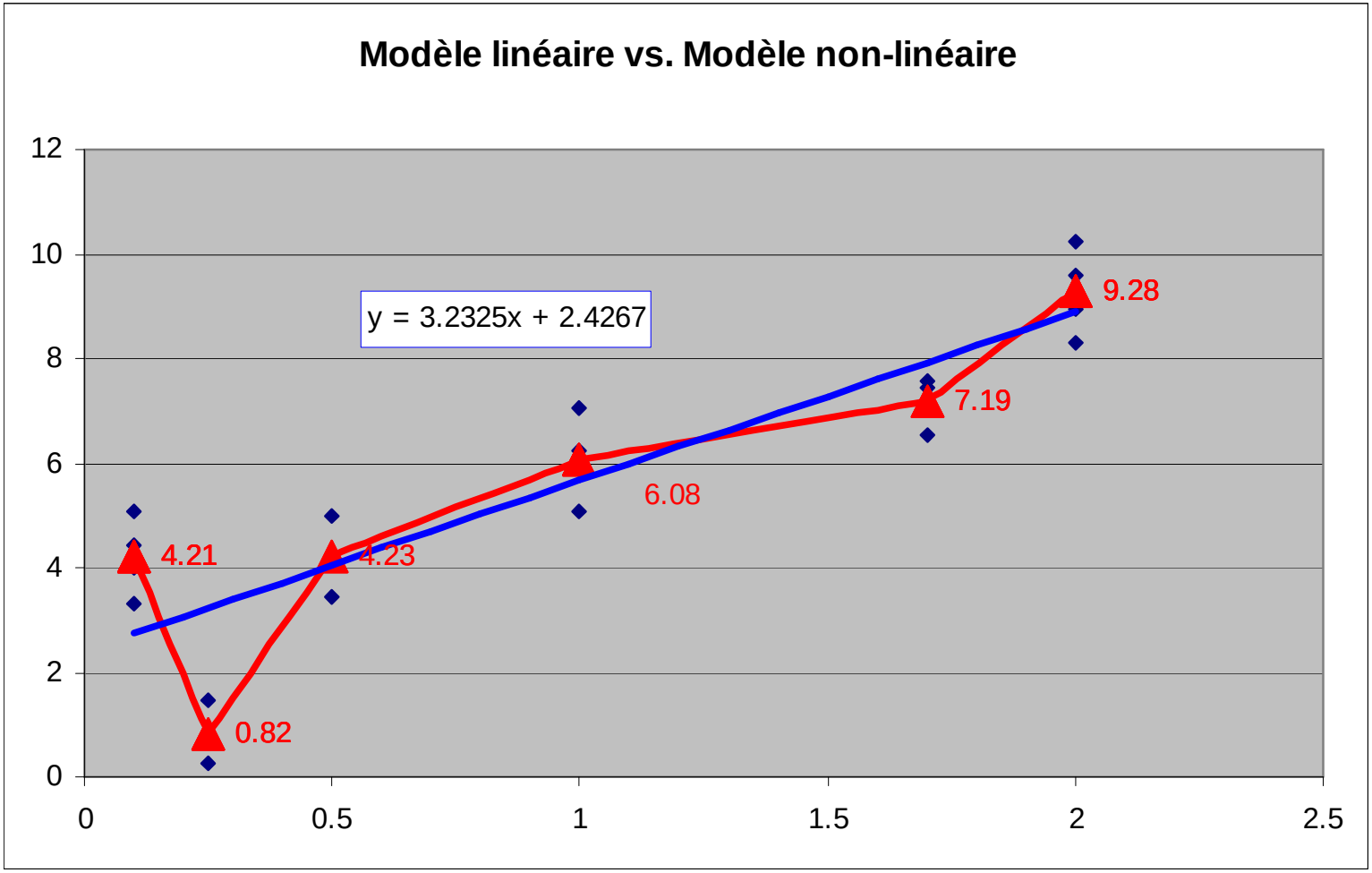


Coefficients — Évaluation des coefficients

```
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
```

Évaluation globale de la régression

Comparaison Linéaire vs. Non-linéaire



Conclusion

En termes de performances
Dans la pratique, les arbres de régression ne se démarquent pas de la régression linéaire

En matière d'exploration
Les arbres sont à privilégier, ils permettent d'identifier des « zones » où les observations sont homogènes, et procéder alors une estimation locale des paramètres de distribution de Y

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