

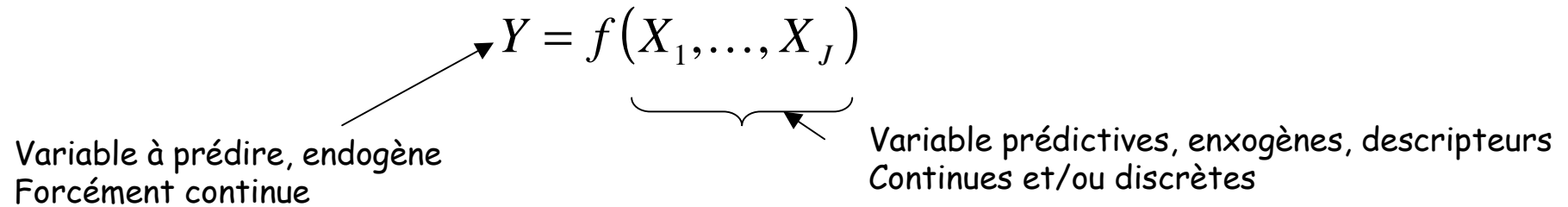
# Arbres de Régression

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



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

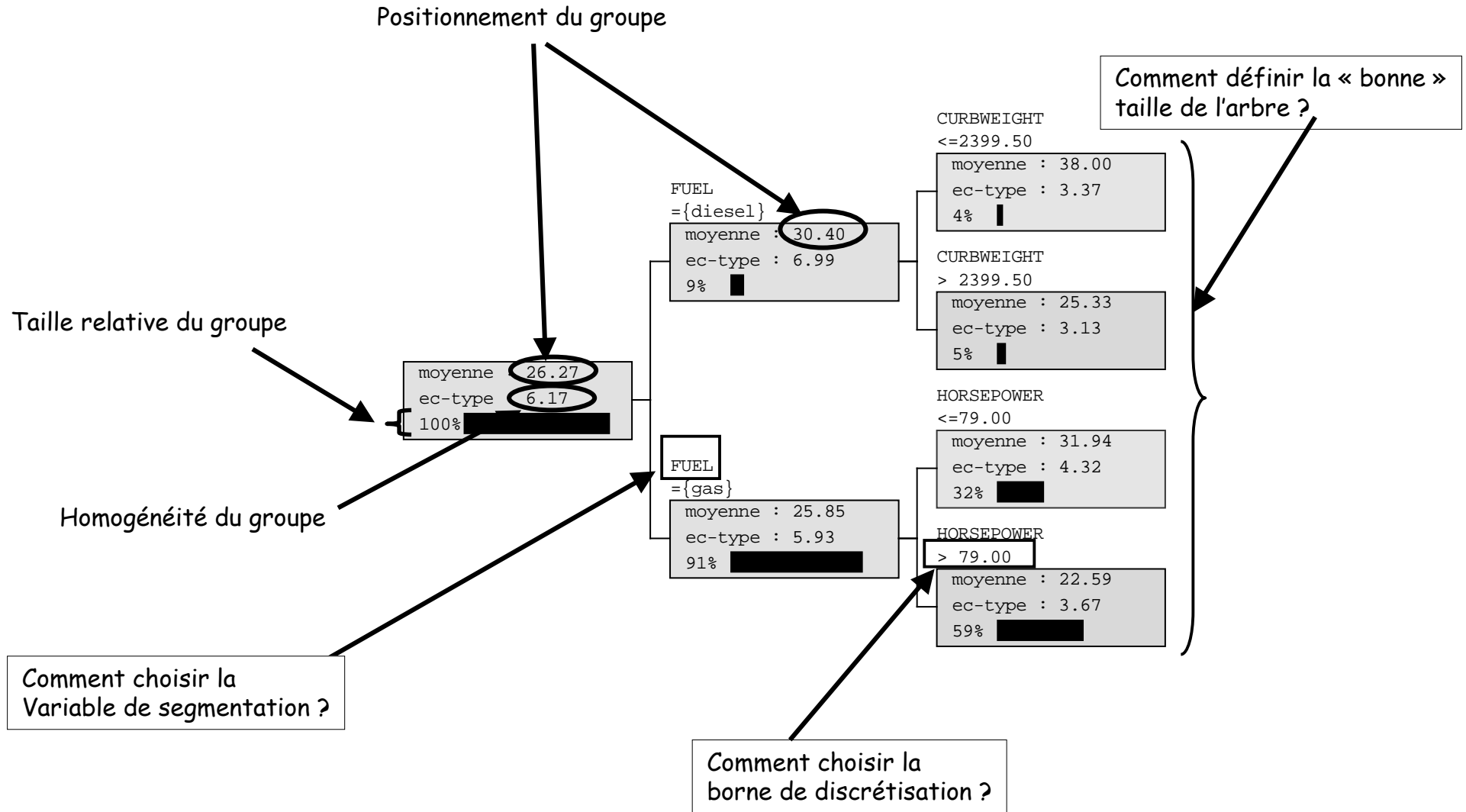


Solution : ARBRE DE REGRESSION

- (1) Un arbre logique
- (2) Segmentation de manière à obtenir des groupes « purs » sur  $Y$
- (3) Critère des moindres carrés

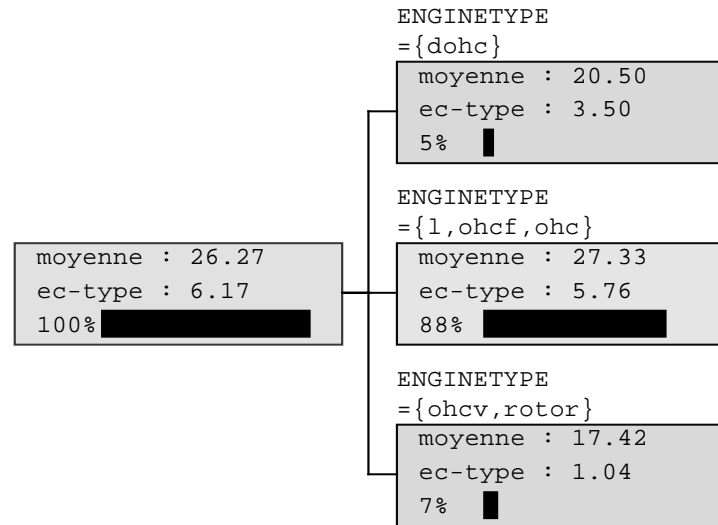
# Arbres de régression

## 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 :  $TSS = BSS + WSS$

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

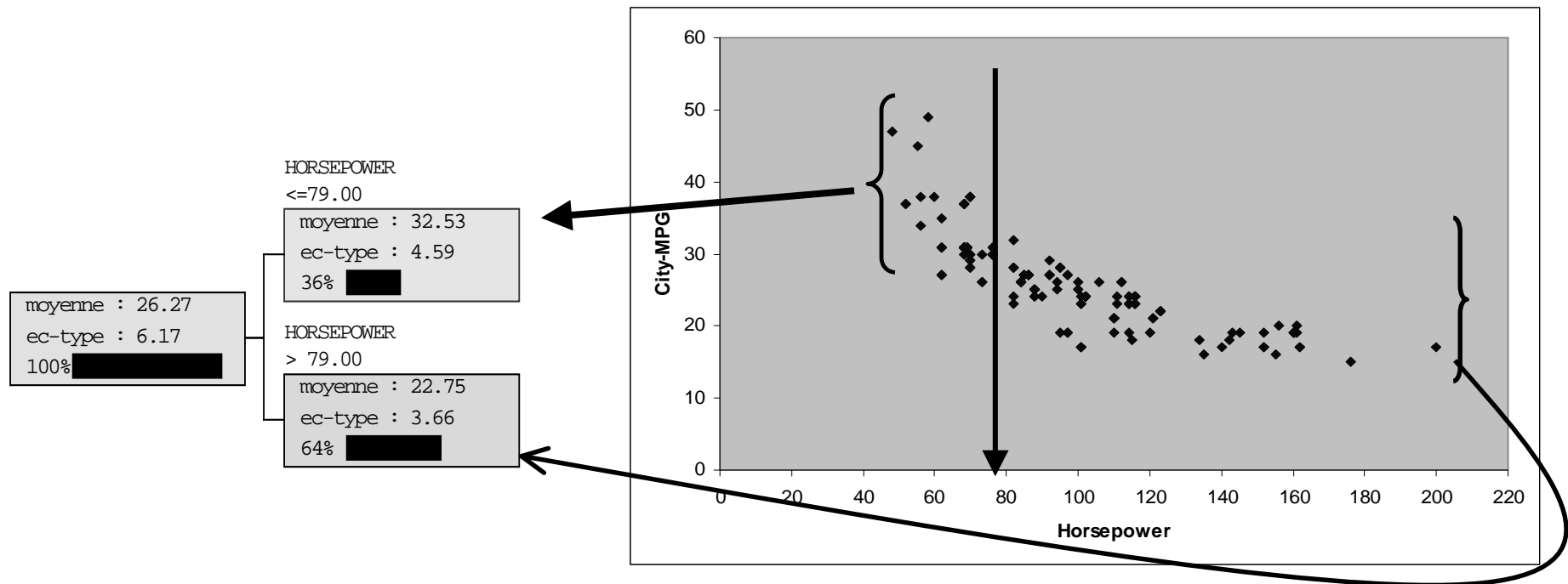
$\underbrace{\hspace{10em}}_{n \times V. Totale}$ 
 $\underbrace{\hspace{10em}}_{n \times V. Inter-classes}$ 
 $\underbrace{\hspace{10em}}_{n \times V. Intra-classes}$

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

CONSTRUCTION DE L'ARBRE - PARAMETRES

Choix de la méthode:  Aid  CR-T

Paramètres de fonctionnement:

Type d'analyse:  Automatique  Automatique et validation croisée  Interactive

Nombre de divisions: 0

Echantillon test et seuils:

Echantillon test par tirage aléatoire (en%): 0

Effectif minimum pour diviser un segment: 5

Nombre de niveaux de l'arbre: 10

Paramètres spécifiques:

Echantillon d'élagage en %: 33

Probabilité critique pour la segmentation: 0.01

Probabilité critique pour la fusion: 0.05

Correction de Bonferroni: Non  Oui

Préférences

Défaut

Enregistrer...

Paramètres

OK Annuler Aide

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

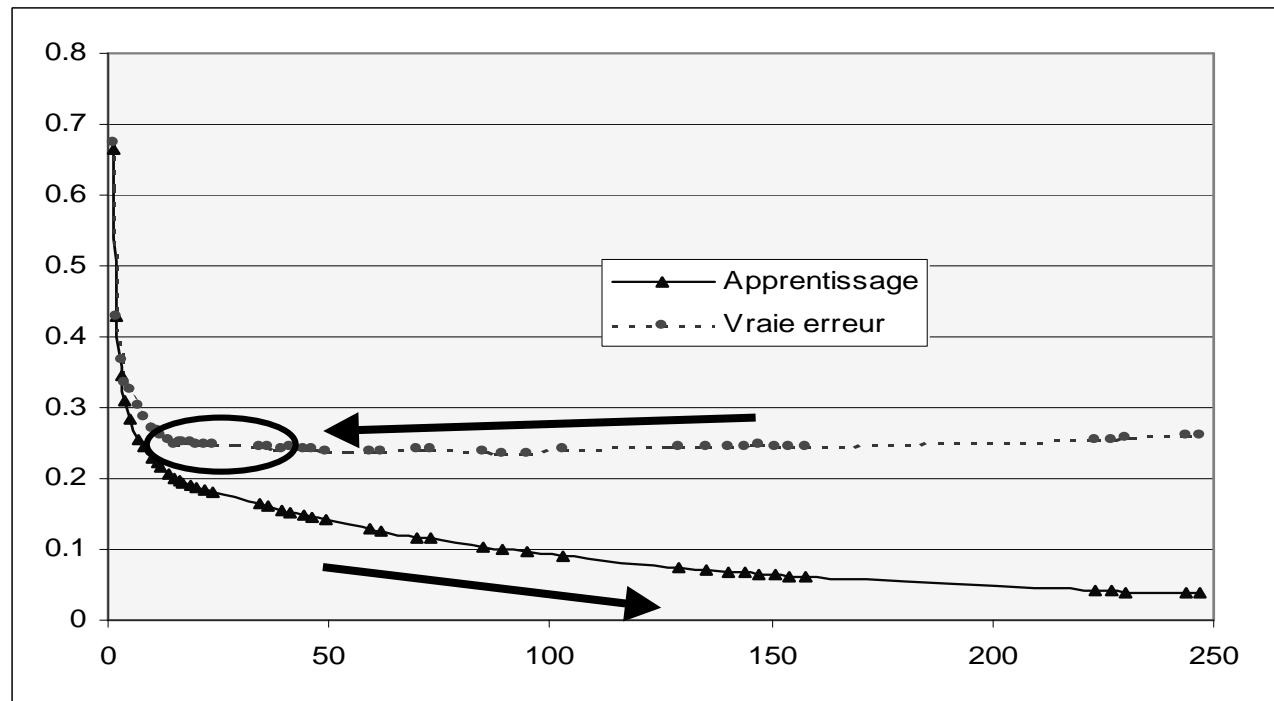
# Arbres de régression

## 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

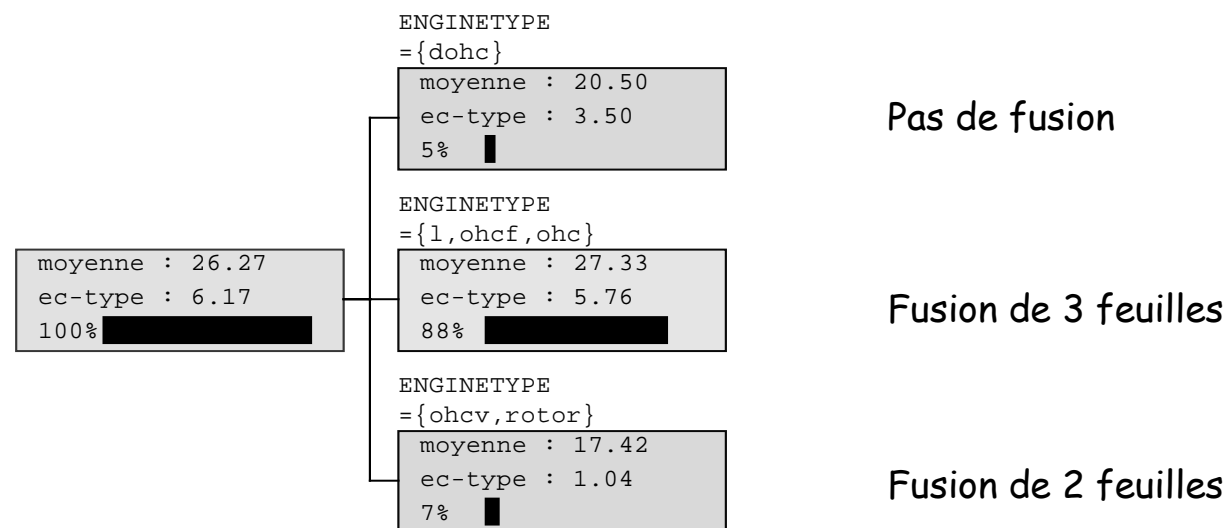
# Arbres de régression

## *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



# Régression linéaire multiple

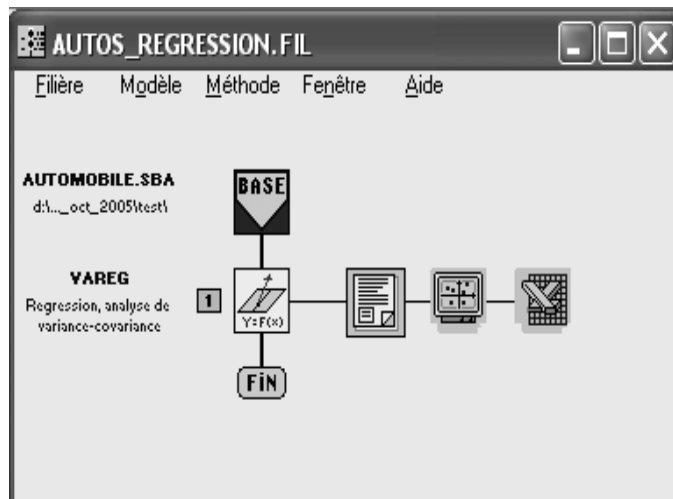
*Une technique alternative*

Solution : REGRESSION LINEAIRE MULTIPLE

- (1) Une combinaison linéaire des variables exogènes
- (2) Méthodes des moindres carrés
- (3) Critère des moindres carrés

$$Y = a_0 + a_1X_1 + \dots + a_JX_J + \varepsilon$$

Coefficients      Évaluation des coefficients



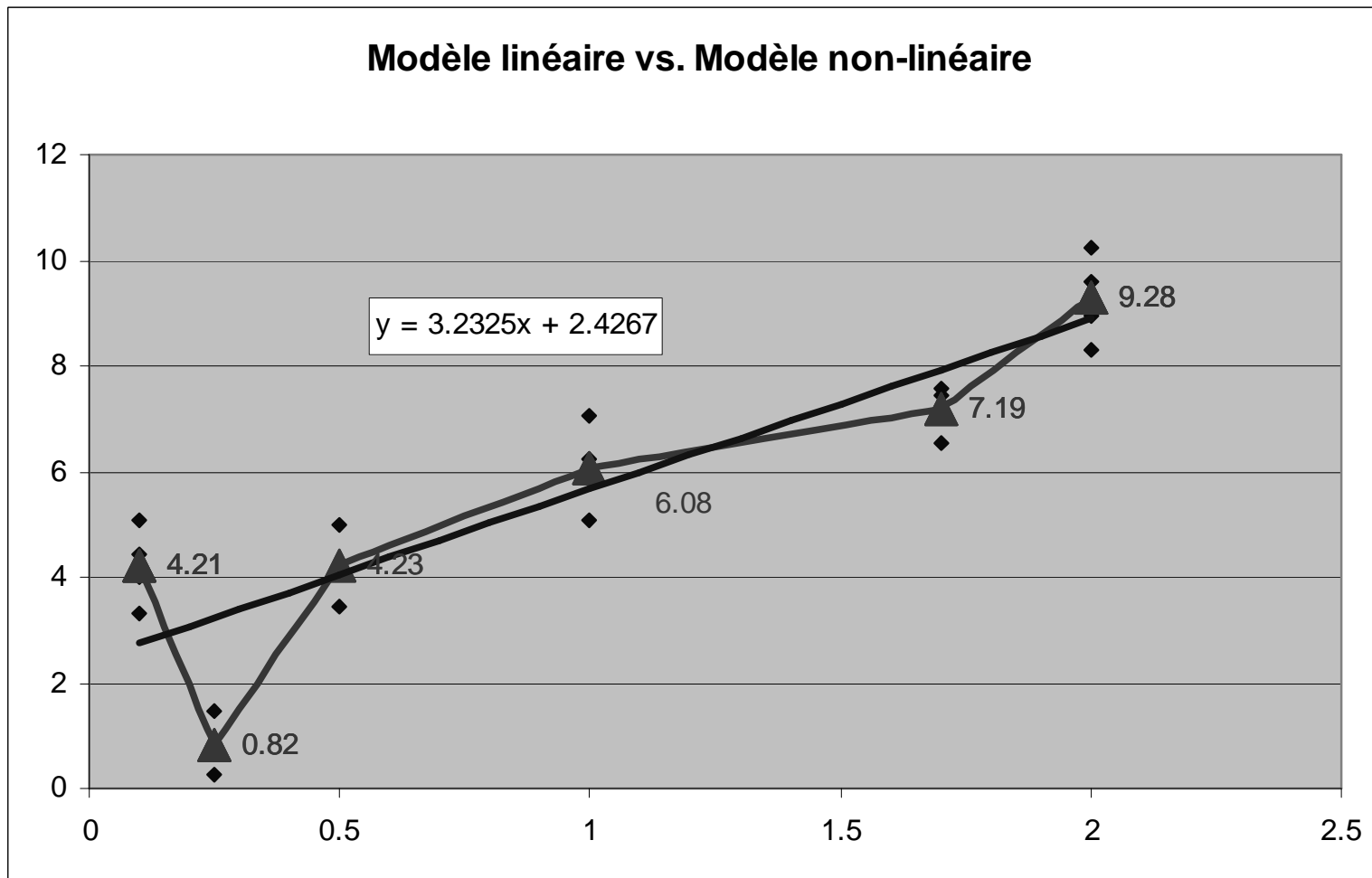
```

ESTIMATION / COEFFICIENTS
AJUSTEMENT DES MOINDRES CARRÉS (AVEC TERME CONSTANT)
164 INDIVIDUS, 4 PARAMETRES (CONSTANTE EN QUEUE).
IDEN  LIBELLE                COEFFICIENT  ECART-TYPE  STUDENT  PROBA.  V.TEST
                                           160
CRITERE ($)
C13 - CURBWEIGHT             -0.0066     0.001     5.793    0.000   -5.51
C16 - ENIGNEZSIZE            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 CARRÉS DES ECARTS ..... SCE =          1632.5071
COEFFICIENT DE CORRELATION MULTIPLE ... R =          0.8596  R2 =          0.7389
VARIANCE ESTIMÉE DES RESIDUS ..... S2 =          10.2032  S =          3.1942
TEST DE NULLITE SIMULTANÉE 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.