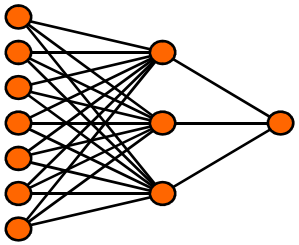
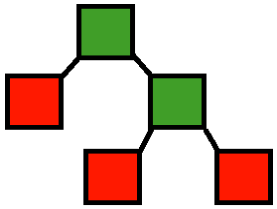
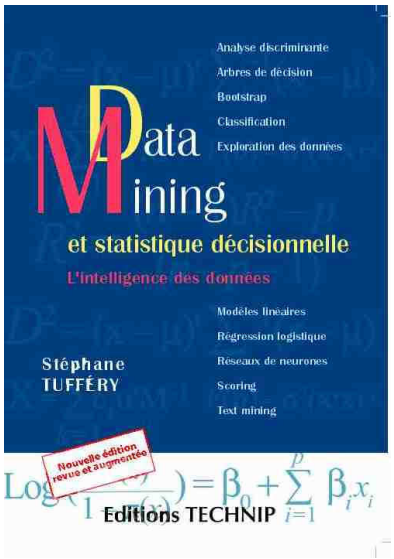
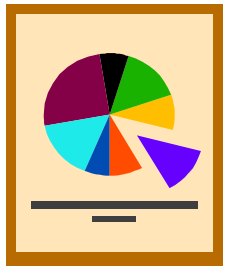
Stéphane Tufféry

DATA MINING
& STATISTIQUE DÉCISIONNELLE









Plan du cours

- Qu'est-ce que le data mining ?
- A quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs - Consulting
- L'analyse et la préparation des données
- Techniques descriptives de data mining
- *Techniques prédictives de data mining*
- Logiciels de statistique et de data mining
- Informatique décisionnelle et de gestion
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

Techniques prédictives
Points forts et points faibles

Techniques prédictives de data mining :

Généralités

Les 2 grandes familles : Classement et prédiction

- **Classement** : la variable à expliquer est *qualitative*
  - on parle aussi de **classification** (dans l'école anglo-saxonne) ou de **discrimination**
  - **scoring** : classement appliqué à une problématique d'entreprise

- **Prédiction** : la variable à expliquer est *continue*
  - on parle aussi de **régression**
  - ou d'**apprentissage supervisé** (réseaux de neurones)

Classement ≠ classification

- Le **classement** consiste à placer chaque individu de la population dans une *classe*, parmi plusieurs classes prédéfinies, en fonction des caractéristiques de l'individu indiquées comme variables explicatives
- Le résultat du classement est un algorithme permettant d'affecter chaque individu à la meilleure classe
- Le plus souvent, il y a 2 classes prédéfinies (« sain » et « malade », par exemple)
- La **classification** consiste à regrouper les individus d'une population en un nombre limité de *classes* qui :
  - ne sont pas prédéfinies mais déterminées au cours de l'opération (même leur nombre n'est pas toujours prédéfini)
  - regroupent les individus ayant des caractéristiques similaires et séparent les individus ayant des caractéristiques différentes (forte inertie interclasse ⇔ faible inertie intraclasse)

Prédiction

- La **prédiction** consiste à estimer
  - la valeur d'une variable continue (dite « à expliquer », « cible », « réponse », « dépendante » ou « endogène »)
  - en fonction de la valeur d'un certain nombre d'autres variables (dites « explicatives », « de contrôle », « indépendantes » ou « exogènes »)
- Cette variable « cible » est par exemple :
  - le poids (en fonction de la taille)
  - la taille des ailes d'une espèce d'oiseau (en fonction de l'âge)
  - le prix d'un appartement (en fonction de sa superficie, de l'étage et du quartier)
  - la consommation d'électricité (en fonction de la température extérieure et de l'épaisseur de l'isolation)

Choix d'une méthode : nature des données

explicatives ➔ **⬇ à expliquer**	**1 quantitative (covariable)**	**n quantitatives (covariables)**	**1 qualitative (facteur)**	**n qualitatives (facteurs)**	**mélange**
1 quantitative	rég. linéaire simple, régression robuste, arbres de décision	rég. linéaire multiple, rég. robuste, PLS, arbres, réseaux de neurones	ANOVA, arbres de décision	ANOVA, arbres de décision, réseaux de neurones	ANCOVA, arbres de décision, réseaux de neurones
***n* quantitatives (représentent des quantités ≠)**	régression PLS2	régression PLS2, réseaux de neurones	MANOVA	MANOVA, réseaux de neurones	MANCOVA, réseaux de neurones
1 qualitative nominale ou binaire	ADL, régression logistique, arbres de décision	ADL, rég. logistique, reg. logistique PLS, arbres, réseaux de neurones, SVM	régression logistique, DISQUAL, arbres	régression logistique, DISQUAL, arbres, réseaux de neurones	régression logistique, arbres, réseaux de neurones
1 discrète (comptage)	modèle linéaire généralisé (régression de Poisson, modèle log-linéaire)				
1 quantitative asymétrique	modèle linéaire généralisé (régressions gamma et log-normale)				
1 qualitative ordinale	régression logistique ordinale (au moins 3 niveaux)				
***n* quantitatives ou qualitatives**	modèle à mesures répétées (les *n* variables représentent des mesures répétées d'une même quantité)				

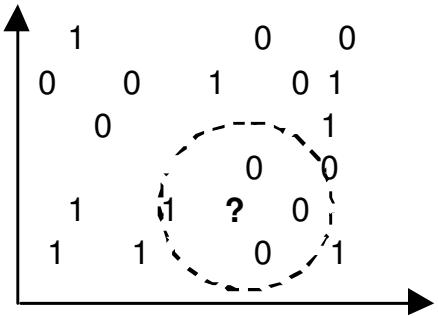
Techniques inductives et transductives

- Dans les **techniques inductives :**
  - une phase d'apprentissage (phase inductive) pour élaborer un modèle, qui résume les relations entre les variables
  - et qui peut ensuite être appliqué à de nouvelles données pour en déduire un classement ou une prédiction (phase déductive)

- Les **techniques transductives**
  - ne comprennent qu'une seule étape (éventuellement réitérée), au cours de laquelle chaque individu est directement classé (ou objet d'une prédiction) par référence aux autres individus déjà classés
  - il n'y a pas élaboration d'un modèle

k-plus proches voisins

- La plus connue des *techniques transductives*
- Le classement (prédiction) de chaque individu s'opère en regardant, parmi les individus déjà classés, la classe des *k* individus qui sont les plus proches voisins (ou en calculant la moyenne dans le voisinage de la variable à prédire)
- La valeur de *k* sera choisie en sorte d'obtenir le meilleur classement (prédiction) possible :
  - ce choix est la principale difficulté de cet algorithme !
- Ainsi, dans l'exemple ci-contre, l'individu « **?** » est classé en « 0 », car entouré en majorité de « 0 »



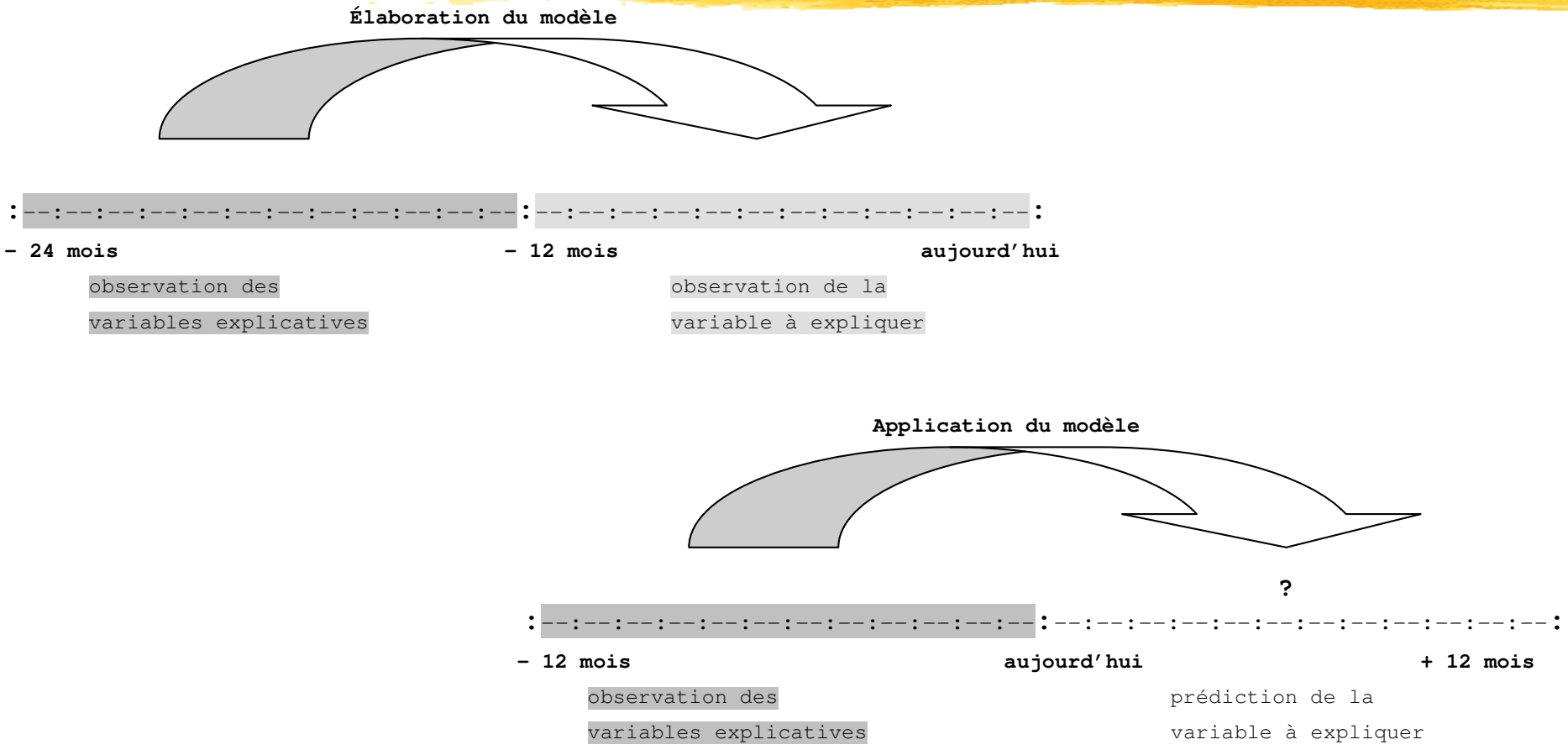
? est classé en "0"

Limites des méthodes transductives

- Une technique inductive *résume dans un modèle l'information* contenue dans les données
  - ce qui permet d '*appliquer rapidement* ce modèle à de nouvelles données
- Une technique transductive *manipule l'ensemble des individus* déjà classés, pour tout nouveau classement
  - ce qui nécessite donc une *grande puissance de stockage et de calcul*
- **On utilise surtout les techniques inductives.**

- Une méthode transductive, comme les k-NN, peut être utilisée dans une étape préalable de détection et de mise à l'écart des individus hors norme, des « outliers ».

Méthodes inductives : schéma



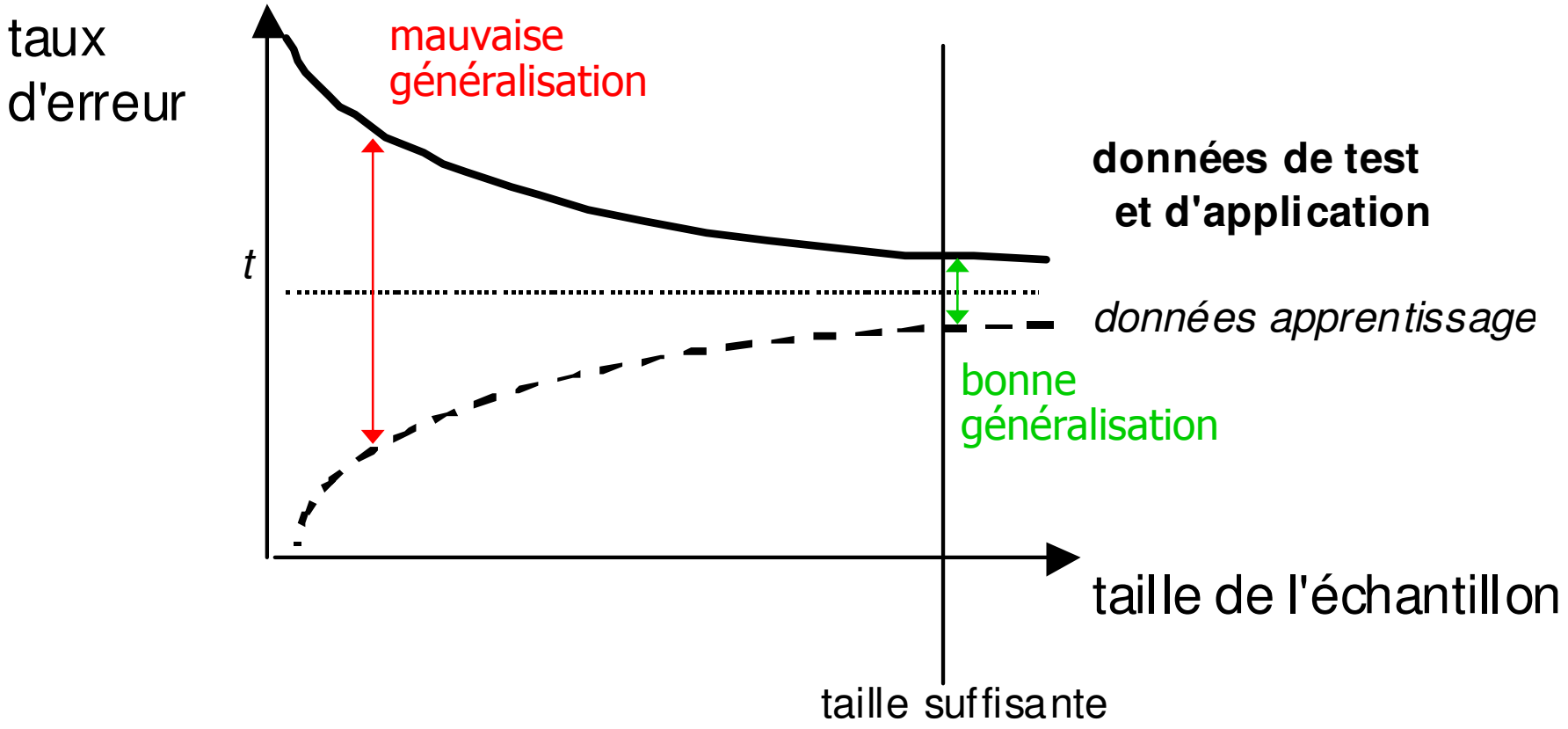
Le **modèle** sera par exemple une fonction *f* telle que :

Probabilité(variable cible = x) = *f*(variables explicatives)

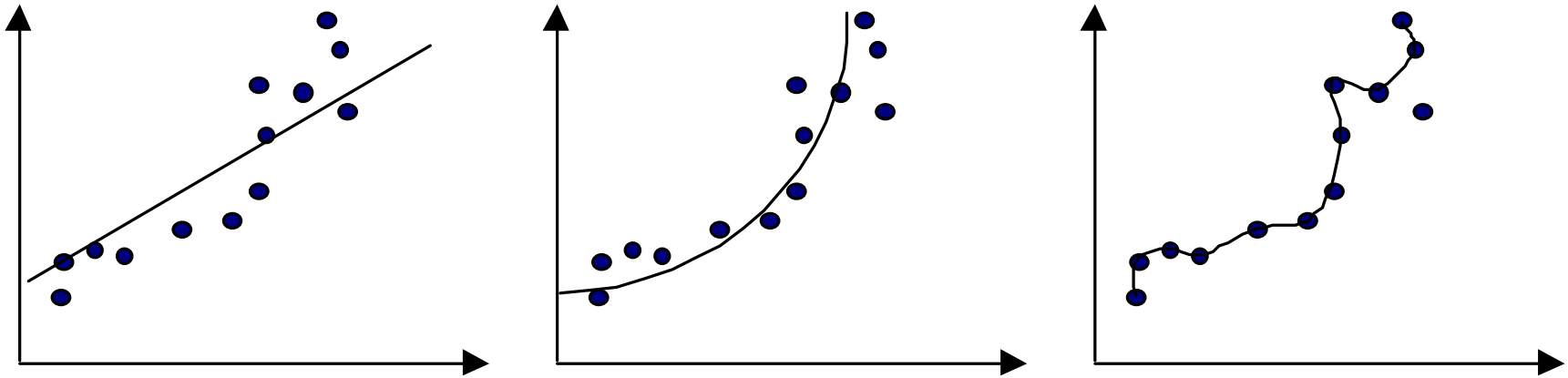
Méthodes inductives : 4 étapes

- Apprentissage : **construction du modèle** sur un 1er échantillon pour lequel on connaît la valeur de la variable cible
- Test : **vérification du modèle** sur un 2d échantillon pour lequel on connaît la valeur de la variable cible, que l'on compare à la valeur prédite par le modèle
  - si le résultat du test est insuffisant (d'après la *matrice de confusion* ou l'*aire sous la courbe ROC*), on recommence l'apprentissage
- Eventuellement, **validation du modèle** sur un 3e échantillon, pour avoir une idée du taux d'erreur non biaisé du modèle
- **Application du modèle** à l'ensemble de la population à scorer, pour déterminer la valeur de la variable cible de chaque individu

Courbes du taux d'erreur en apprentissage et en test



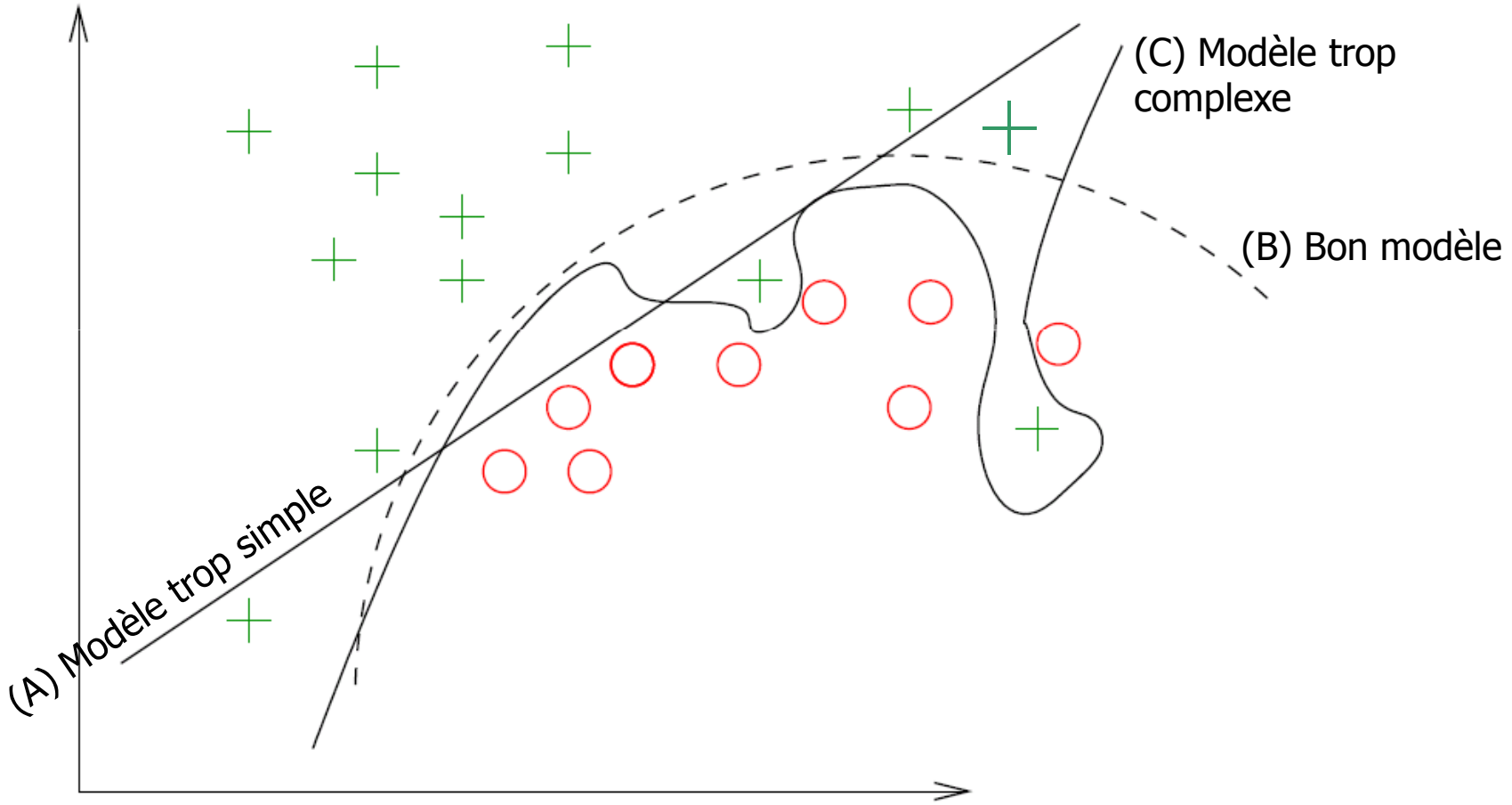
Sur-apprentissage en régression



(A) Modèle trop simple (B) Bon modèle (C) Modèle trop complexe

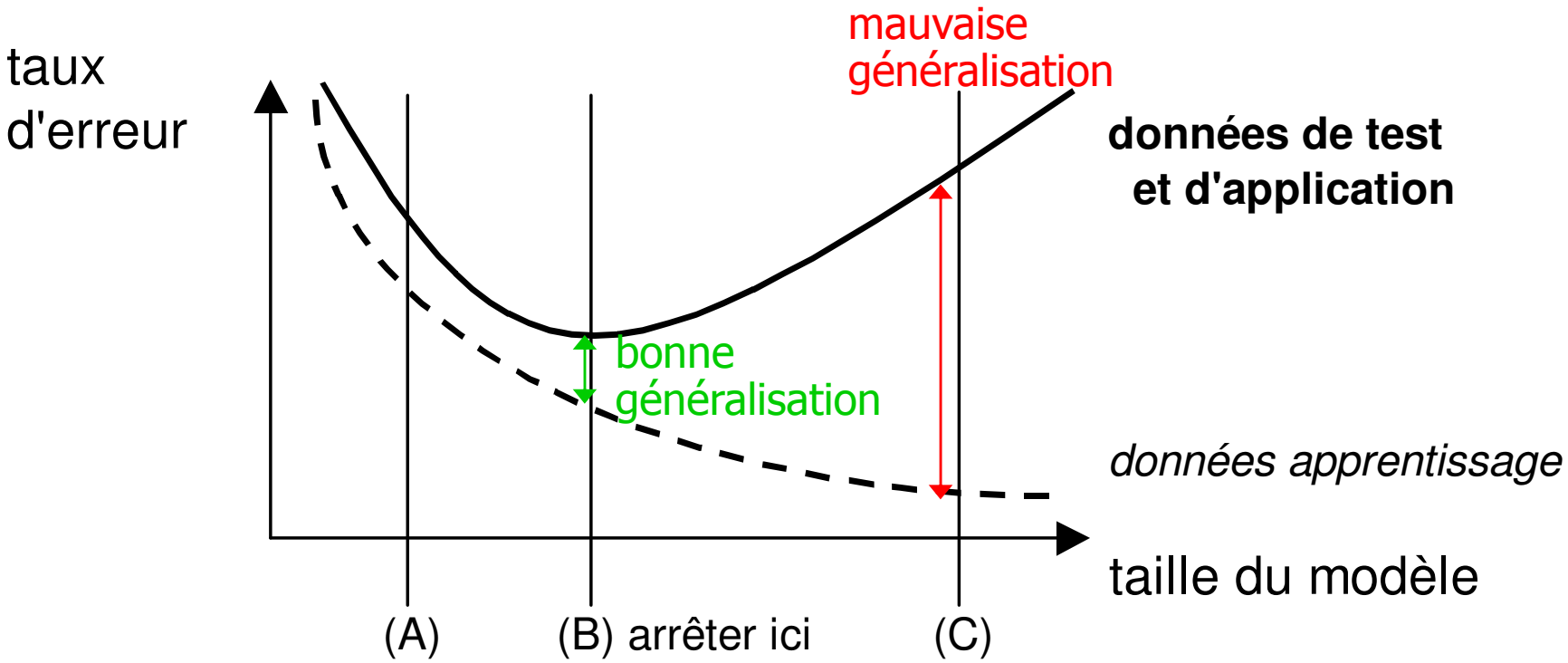
- Un modèle trop poussé dans la phase d'apprentissage :
  - épouse toutes les fluctuations de l'échantillon d'apprentissage,
  - détecte ainsi de fausses liaisons,
  - et les applique à tort sur d'autres échantillons

Sur-apprentissage en classement

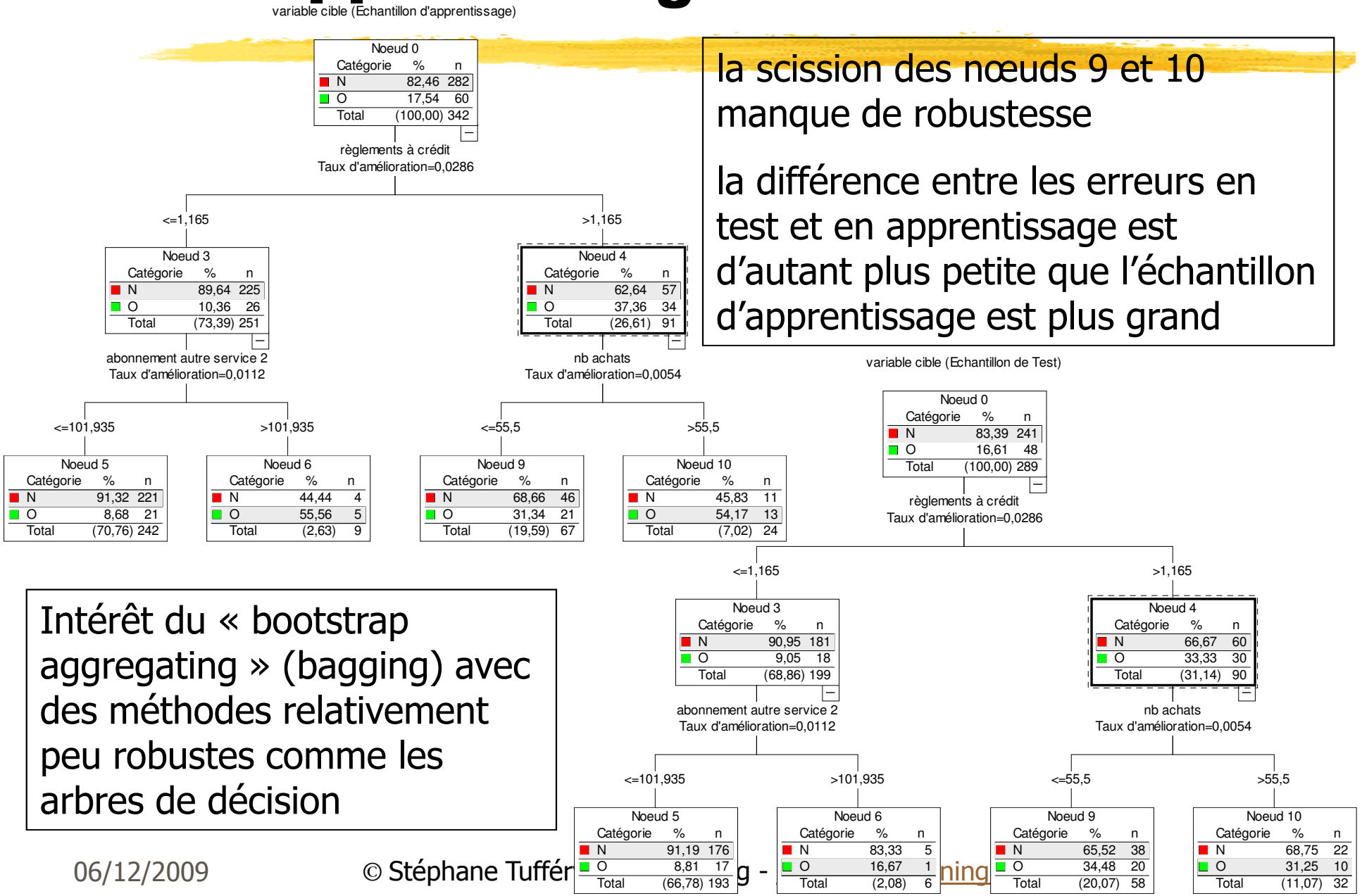


Source : Olivier Bousquet

Taux d'erreur en fonction de la complexité du modèle



Sur-apprentissage dans un arbre



la scission des nœuds 9 et 10 manque de robustesse

la différence entre les erreurs en test et en apprentissage est d'autant plus petite que l'échantillon d'apprentissage est plus grand

Intérêt du « bootstrap aggregating » (bagging) avec des méthodes relativement peu robustes comme les arbres de décision

Méthodes de sélection

- Pas à pas : Ascendante (« forward »)
  - aucune variable au départ : on ajoute 1 à 1 celles qui contribuent le plus au modèle (en un sens pouvant varier selon les cas : R^2, maximum de vraisemblance…)
- Pas à pas : Descendante (« backward »)
  - toutes les variables au départ : on rejette 1 à 1 celles qui sont insuffisamment corrélées à la cible et contribuent le moins au modèle
- Pas à pas : Mixte (« stepwise »)
  - comme « Ascendante », mais on peut retrancher une variable à chaque étape si son pouvoir discriminant est contenu dans une combinaison des nouvelles variables
- Globale : Algorithme de Furnival et Wilson (si 2 groupes)
  - cherche à ajuster le R^2 en comparant une partie de tous les modèles possibles (élimine les moins intéressants *a priori*)

Validation des modèles

- Etape très importante car des modèles peuvent :
  - donner de faux résultats (données non fiables)
  - mal se généraliser dans l'espace (autre échantillon) ou le temps (échantillon postérieur)
    - sur-apprentissage
  - être peu efficaces (déterminer avec 2 % d'erreur un phénomène dont la probabilité d'apparition = 1 % !)
  - être incompréhensibles ou inacceptables par les utilisateurs
    - souvent en raison des variables utilisées
  - ne pas correspondre aux attentes
- Principaux outils de comparaison :
  - matrices de confusion, courbes ROC, de lift, et indices associés

Matrice de confusion

valeur prédite ➔ valeur réelle ➔	A	B	TOTAL
A	1800	200	
B	300	1700	
TOTAL			4000

- Taux d'erreur = (200 + 300) / 4000 = 12,5 %

Q_{PRESS}

- Pour vérifier que le % d'individus correctement classés est significativement meilleur que par un classement aléatoire, on calcule la quantité suivante :

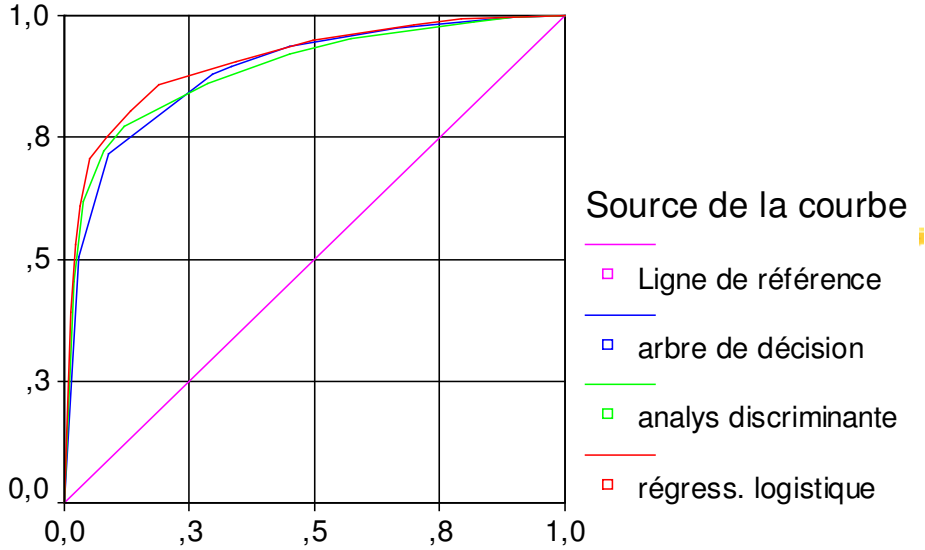
$$Q_{PRESS} = \frac{\left(n - \left(c \times k\right)\right)^2}{n \times \left(k - 1\right)}$$

  - n = taille échantillon
  - k = nb de groupes
  - c = nb d'individus bien classés
- Q_{PRESS} suit un χ^2 à 1 degré de liberté
  - valeur critique : 10,8 à 0,1 % - 6,63 à 1 % - 3,84 à 5 %
- Ici on a : $Q_{PRESS} = (4000 - 7000)^2/4000 = 2250$

Sensibilité et spécificité

- Pour un score devant discriminer un groupe A (les positifs ; ex : les risqués) par rapport à un autre groupe B (les négatifs ; ex : les non risqués), on définit 2 fonctions du seuil de séparation s du score :
  - sensibilité = $\alpha(s)$ = Proba(score $\geq$ s / A) = probabilité de bien détecter un positif
  - spécificité = $\beta(s)$ = Proba(score < s / B) = probabilité de bien détecter un négatif
- Pour un modèle, on cherche s qui maximise $\alpha(s)$ tout en minimisant les faux positifs 1 - $\beta(s)$ = Proba(score $\geq$ s / B)
  - faux positifs : négatifs considérés comme positifs à cause du score
- Le meilleur modèle : permet de capturer le plus possible de vrais positifs avec le moins possible de faux positifs

Courbe ROC

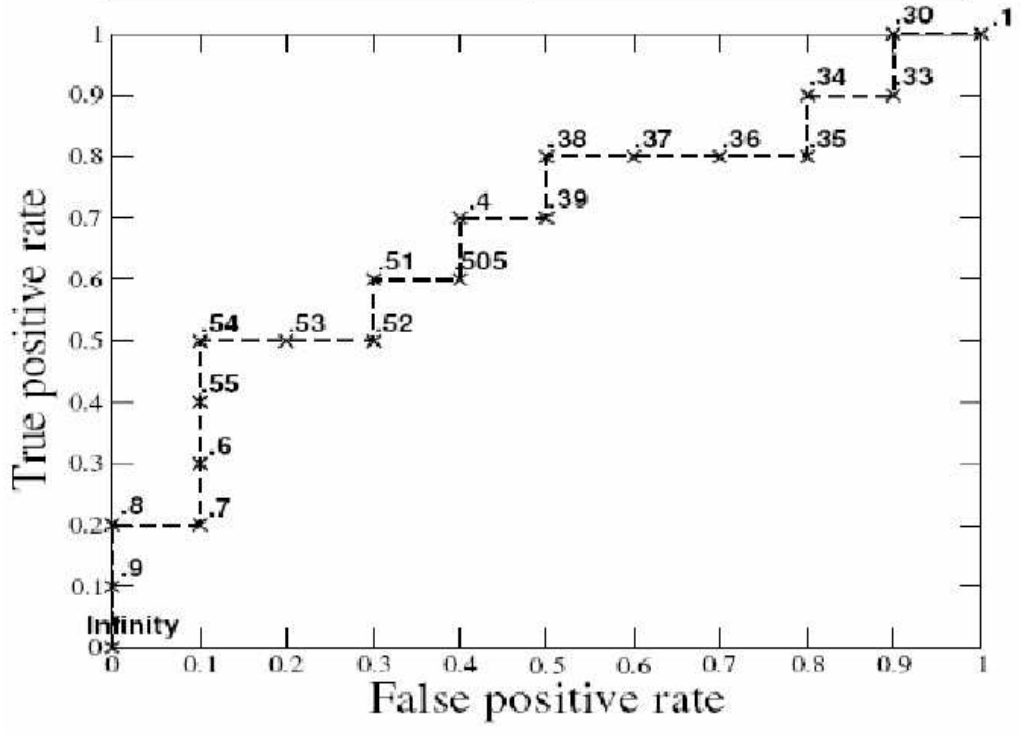


- La courbe ROC

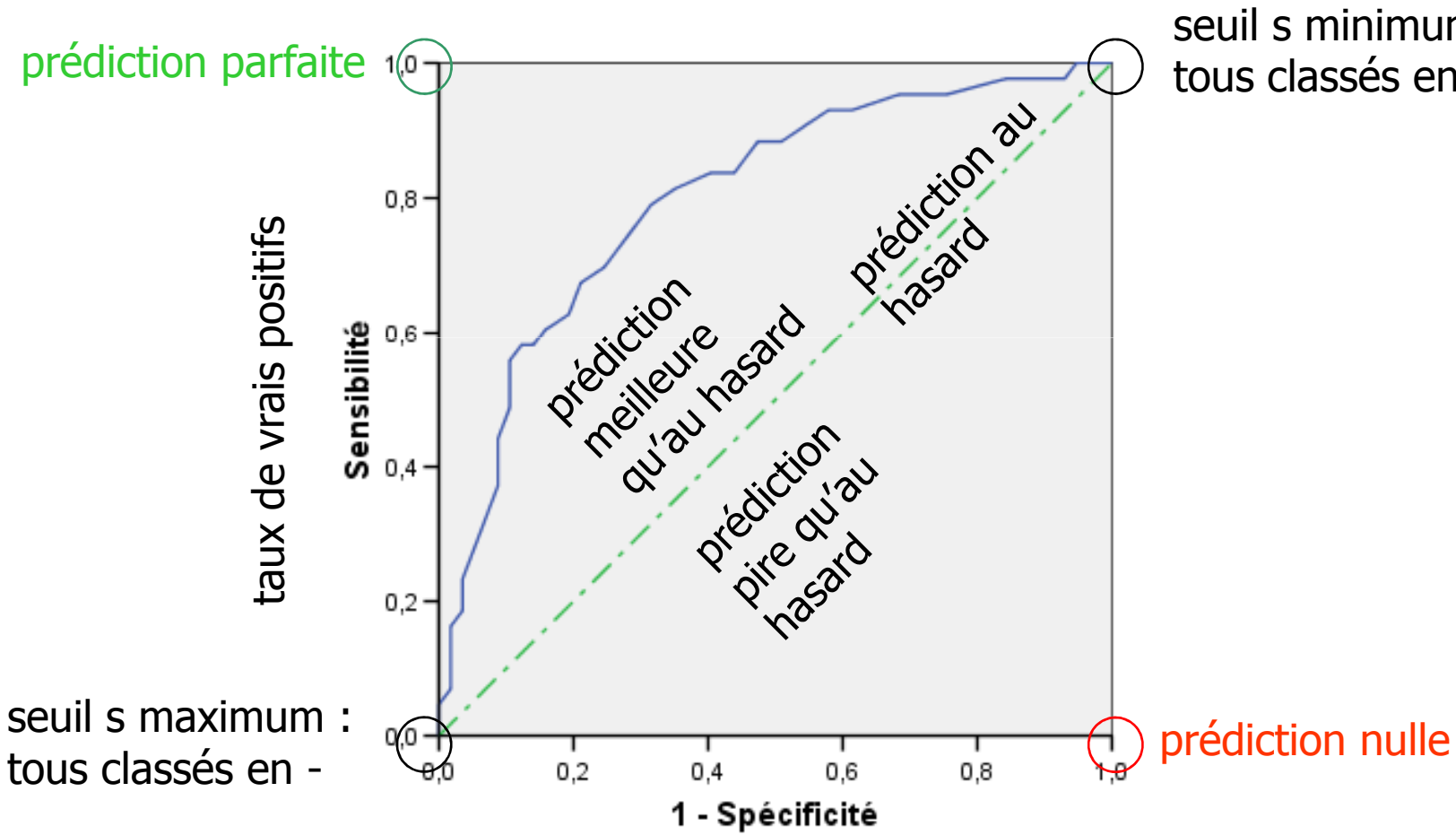
  (receiver operating characteristic)
  - sur l'axe Y : sensibilité = $\alpha(s)$
  - sur l'axe X : 1 - spécificité = 1 - $\beta(s)$
  - proportion y de vrais positifs en fonction de la proportion x de faux positifs, lorsque l'on fait varier le seuil s du score
- Exemple : si la courbe ROC passe par le point (0,3;0,9), ce point correspond à un seuil s qui est tel que : si on considère « risqués » tous les individus dont le score ≥ s, on a détecté :
  - 30% de faux risqués (30% des non-risqués ont un score ≥ s : ce sont les faux positifs)
  - 90 % de vrais risqués (90 % des risqués ont un score ≥ s : ce sont les vrais positifs)
  - NB : 0,3 ne correspond pas à 30 % de la population totale !

Exemple de courbe ROC

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



Interprétation de la courbe ROC



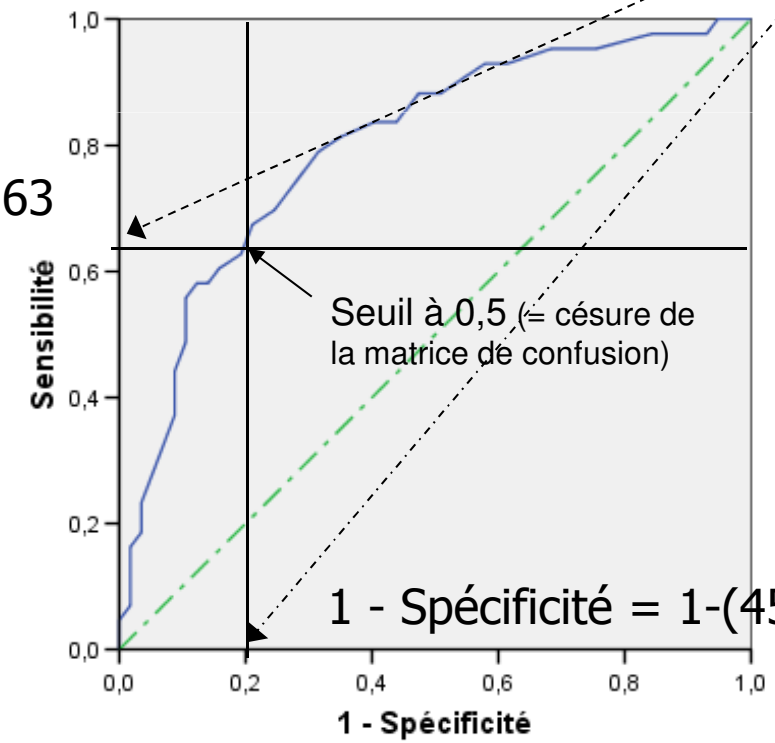
Matrice de confusion et courbe ROC

Tableau de classement[a]

Observé		Prévu		
		CHD		Pourcentage correct
		0	1	
CHD	0	45	12	78,9
	1	16	27	62,8
Pourcentage global				72,0

a. La valeur de césure est ,500

Sensibilité = 27/43 = 0,63



1 - Spécificité = 1-(45/57) = 0,21

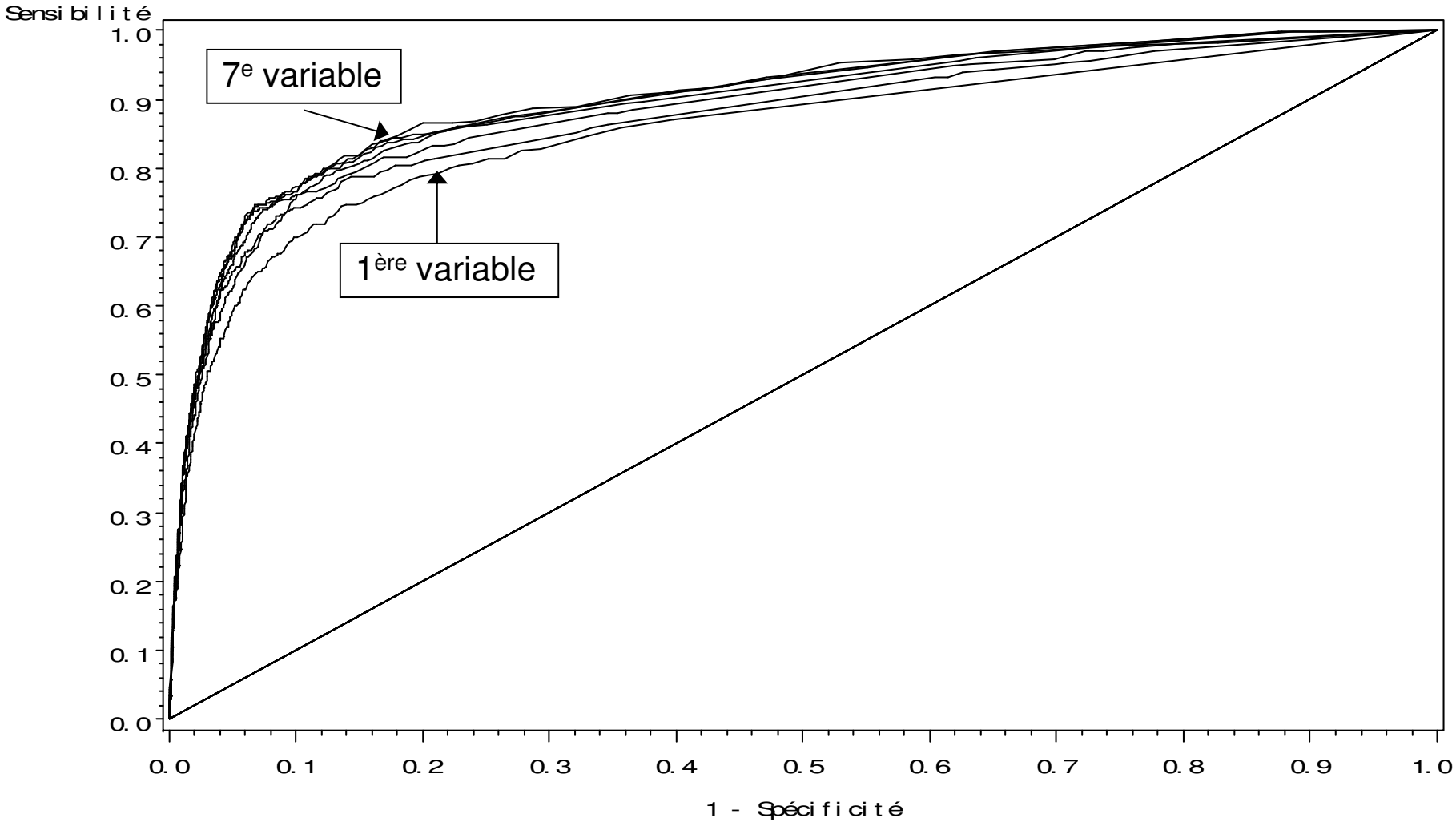
Matrice de confusion (avec SAS)

Table de classification									
Niveau de prob.	Correct		Incorrect		Pourcentages				
	Événement	Non-événement	Événement	Non-événement	Correct	Sensibilité	Spécificité	POS fausse	NEG fausse
0.000	57	0	43	0	57.0	100.0	0.0	43.0	.
0.100	57	1	42	0	58.0	100.0	2.3	42.4	0.0
0.200	55	7	36	2	62.0	96.5	16.3	39.6	22.2
0.300	51	19	24	6	70.0	89.5	44.2	32.0	24.0
0.400	50	25	18	7	75.0	87.7	58.1	26.5	21.9
0.500	45	27	16	12	**72.0**	**78.9**	**62.8**	**26.2**	**30.8**
0.600	41	32	11	16	73.0	71.9	74.4	21.2	33.3
0.700	32	36	7	25	68.0	56.1	83.7	17.9	41.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	63.0	42.1	90.7	14.3	45.8
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	48.0	10.5	97.7	14.3	54.8
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	43.0	0.0	100.0	.	57.0

prédit ➡ Observé ⬇	0	1	total
0	45	12	57
1	16	27	43
total	61	39	100

Correct = (45 + 27) / 100 = 72 %
Sensibilité = 45 / 57 = 78,9 %
Spécificité = 27 / 43 = 62,8 %
POS fausse = 16 / 61 = 26,2 %
NEG fausse = 12 / 39 = 30,8 %

Courbes ROC avec entrée progressive des variables du modèle



AUC : Aire sous la courbe ROC

- Aire AUC sous la courbe ROC = probabilité que score(x) > score(y), si x est tiré au hasard dans le groupe A (à prédire) et y dans le groupe B
- 1ère méthode d'estimation : par la méthode des trapèzes
- 2e méthode d'estimation : par les paires concordantes
  - soit n_1 (resp. n_2) le nb d'observations dans A (resp. B)
  - on s'intéresse aux $n_1 n_2$ paires formées d'un x dans A et d'un y dans B
  - parmi ces t paires : on a concordance si score(x) > score(y) ; discordance si score(x) < score(y)
  - soient nc = nb de paires concordantes ; nd = nb de paires discordantes ; $n_1 n_2$ - nc - nd = nb d'ex aequo
  - aire sous la courbe ROC $\approx$ (nc + 0,5[t - nc - nd]) / $n_1 n_2$
- 3e méthode équivalente : par le test de Mann-Whitney
  - U = $n_1 n_2 (1 - \text{AUC})$ ou $n_1 n_2 \text{AUC}$

AUC : calcul avec SAS

```
ODS OUTPUT WilcoxonScores = wilcoxon;

PROC NPAR1WAY WILCOXON DATA=&data
CORRECT=no;
CLASS &cible;
VAR &score;
RUN;

DATA auc;
SET wilcoxon;
n2 = N; R2 = SumOfScores ;
n1 = LAG(N); R1 = LAG(SumOfScores) ;
u1 = (n1*n2) + (n1*(n1+1)/2) - R1 ;
u2 = (n1*n2) + (n2*(n2+1)/2) - R2 ;
u = MIN(u1,u2);
AUC = ROUND(1- (u/(n1*n2)),0.001);
RUN;

PROC PRINT DATA=auc (KEEP = AUC) ;
TITLE "Aire sous la courbe ROC de &data";
WHERE AUC > .;
RUN;
```

U est la statistique de Mann-Whitney, qui se déduit des effectifs n_i et de la somme des rangs R_i fournis par la proc NPAR1WAY de SAS

$$U = \min\left\{ n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 , n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \right\}$$

nb de fois où un score du groupe 1 > un score du groupe 2

Obs	Class	N	SumOfScores	n2	R2	n1	R1	U1	U2	U	AUC
1	1	711	1038858.0	711	1038858	.	.	.	.	.	.
2	0	1490	1384443.0	1490	1384443	711	1038858	273648	785742	273648	0.74169

Utilisation de l'AUC

- Le modèle est d'autant meilleur que l'AUC est plus proche de 1
- Si l'AUC = 0,5 : modèle pas meilleur qu'une notation aléatoire. Il existe un intervalle de confiance sur l'AUC et un test associé :

Variable(s) de résultats tests	Zone	Erreur Std.[a]	Signif. asymptotique[b]	Intervalle de confiance 95% asymptotique	
				Borne inférieure	Borne supérieure
arbre de décision	,887	,008	,000000	,872	,902
régression logistique	,906	,007	,000000	,892	,921
analyse discriminante	,889	,008	,000000	,873	,904

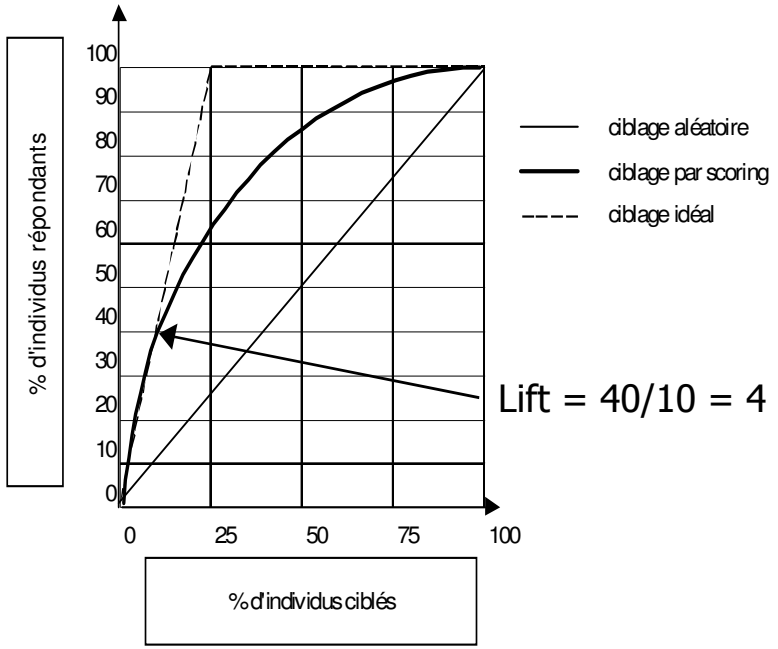
a. Dans l'hypothèse non-paramétrique

b. Hypothèse nulle /: zone vraie = 0.5

- Permet de comparer des modèles de types différents
  - sur tout échantillon

Courbe de lift

- La courbe de lift :
  - sur l'axe Y : on a la sensibilité = $\alpha(s)$ = Proba(score ≥ s / A)
  - sur l'axe X : on a Proba(score ≥ s)
  - proportion y de vrais positifs en fonction des individus sélectionnés, lorsque l'on fait varier le seuil s du score
  - même ordonnée que la courbe ROC, mais une abscisse généralement plus grande
  - > la courbe de lift est généralement sous la courbe ROC
- Très utilisée en marketing



Lien entre courbe de lift et ROC

- Relation entre l'aire AUL sous la courbe de lift et l'aire AUC :
  - AUC – AUL = p(AUC – 0,5) ⇔ AUL = p/2 + (1 – p)AUC
  - où p = Proba(A) = probabilité *a priori* de l'événement dans la population
- Cas particuliers :
  - AUC = 1 ⇒ AUL = p/2 + (1 – p) = 1 – p/2
  - AUC = 0,5 ⇒ AUL = p/2 + 1/2 – p/2 = 0,5
  - p petit ⇒ AUC et AUL sont proches
  - $AUC_1 > AUC_2 \Leftrightarrow AUL_1 > AUL_2$
- Ces indicateurs sont des critères universels de comparaison de modèles

Technique de prédiction :
La régression linéaire

Cadre du modèle linéaire

- Dans le modèle simple : X et Y deux variables continues
- Les valeurs x_i de X sont contrôlées et sans erreur de mesure
- On observe les valeurs correspondantes y_1, ..., y_n de Y
- Exemples :
  - X peut être le temps et Y une grandeur mesurée à différentes dates
  - Y peut être la différence de potentiel mesurée aux bornes d'une résistance pour différentes valeurs de l'intensité X du courant

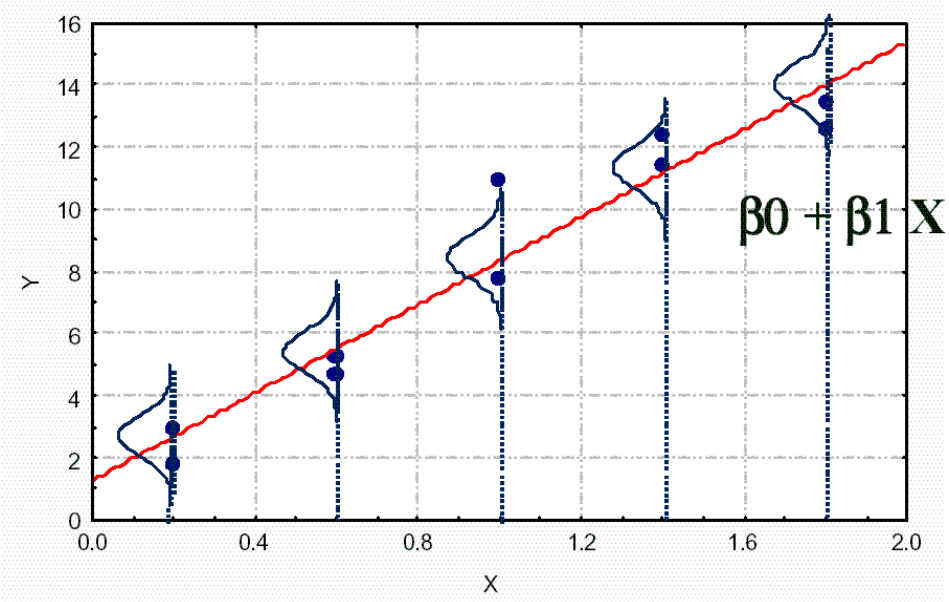
Hypothèse fondamentale du modèle linéaire

- X et Y ne sont pas indépendantes et la connaissance de X permet d'améliorer la connaissance de Y
- Savoir que X = x permet rarement de connaître exactement la valeur de Y, mais on suppose que cela de connaître la valeur moyenne E(Y|X=x), l'espérance conditionnelle de Y sachant que X = x
- On suppose plus précisément que E(Y|X=x) est une fonction linéaire de x, ce qui permet d'écrire
  - $E(y_i) = \alpha + \beta x_i$ pour tout i = 1, …, n

  $\Leftrightarrow y_i = \alpha + \beta x_i + \varepsilon_i$, avec $E(\varepsilon_i) = 0$ pour tout i = 1, …, n
  - n = nb d'observations et ε_i = « résidu » de l'observation i
- Régression linéaire multiple :
  - $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$
  - important : on suppose l'indépendance linéaire des X_i

Autres hypothèses du modèle linéaire

- La variance des résidus est la même pour toutes les valeurs de X (homoscédasticité)
  - $V(\varepsilon_i) = s^2$
- Les résidus sont linéairement indépendants
  - $cov(\varepsilon_i,\varepsilon_j) = 0 \quad \forall\ i \neq j$
- Les résidus sont normalement distribués
  - $\varepsilon_i \sim N(0,s^2)$



La composante stochastique

- L'existence de la composante stochastique (ε_i) correspond au fait que :
  - des individus avec même valeur x_i peuvent avoir des réponses Y différentes (variation synchronique)
  - OU un même individu mesuré à plusieurs reprises avec la même valeur x_i peut avoir des réponses Y différentes (variation diachronique)
- On a équivalence de $\varepsilon_i \sim N(0,s^2)$ et $Y/X=x_i \sim N(\alpha + \beta x_i, s^2)$
- Cette hypothèse de normalité classe la régression linéaire dans la famille des modèles linéaires généraux (GLM)
- Dans les modèles linéaires généralisés, la loi de $Y/X=x_i$ n'est plus nécessairement normale

Que signifie la variance des estimateurs ?

- Après avoir postulé l'existence d'une relation $E(Y) = \alpha + \beta X$, on recherche des estimateurs a et b de α et β
- On n'atteint jamais les véritables coefficients α et β car :
  - le modèle linéaire n'est le plus souvent qu'une approximation de la réalité
  - on ne travaille que sur des échantillons et non la population entière
  - on commet des erreurs de mesure
- Des modèles sur des échantillons différents donneront des estimateurs a' et b' différents
- D'où une variance des estimateurs a et b

Méthode des moindres carrés ordinaires (MCO)

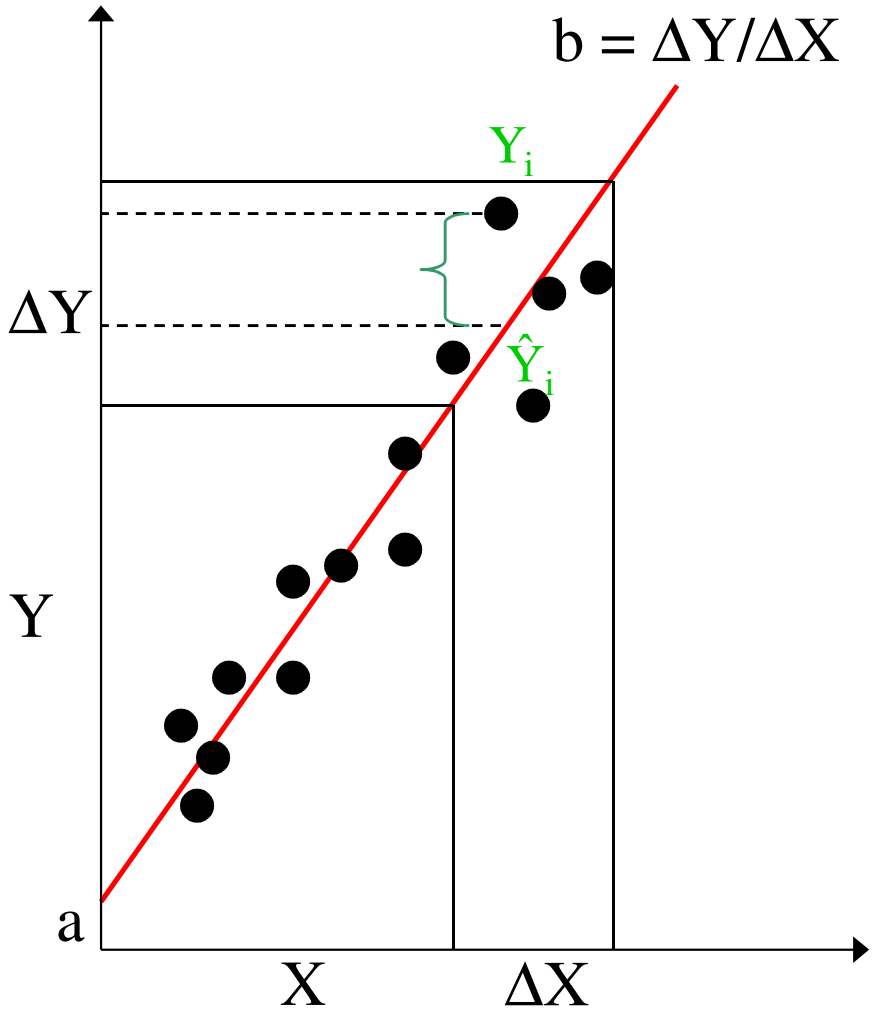
- On recherche des estimateurs a et b de α et β qui minimisent les résidus $\varepsilon_i^2 = (Y_i - \hat{Y}_i)^2$, où $\hat{Y}_i$ est prédit par la droite $\hat{Y} = a + bX$
- L'estimateur b de la pente est :

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(X,Y)}{\sigma^2_X}$$

- L'estimateur a de la constante vaut :

$$a = \bar{y} - b.\ \bar{x}$$

- La droite $\hat{Y} = a + b.X$ ajuste le nuage de points



Propriétés des estimateurs MCO

- Les estimateurs MCO des coefficients ont :
  - une moyenne : E(a) et E(b)
  - une variance :
    - constante : $\sigma_a^2 = s^2 [1/n + \bar{x}^2 / \Sigma (x_i - \bar{x})^2]$
    - avec : s^2 = variance des résidus
    - > IC au niveau $100(1-\alpha)\% = a \pm t_{\alpha/2,n-p-1} . \sigma_a$
    - pente : $\sigma_b^2 = s^2 [1/ \Sigma (x_i - \bar{x})^2]$
    - > IC au niveau $100(1-\alpha)\% = b \pm t_{\alpha/2,n-p-1} . \sigma_b$
- La méthode MCO est optimale car :
  - les estimateurs sont sans biais : $E(a) = \alpha$ et $E(b) = \beta$
  - de variance minimale parmi tous les estimateurs linéaires
  - on dit qu'ils sont « BLUE » : best linear unbiased estimators
- Hypothèse de normalité $\varepsilon_i \sim N(0,s^2) \Rightarrow$ les estimateurs sont de variance minimale parmi tous les estimateurs

Conséquence des formules de variance

- Pour diminuer les variances :
  - diminuer la variance résiduelle s^2 de l'échantillon
  - augmenter la taille n de l'échantillon
  - augmenter l'étendue des valeurs observées de X
- Mais : on accepte parfois (régression ridge) des estimateurs légèrement biaisés pour diminuer leur variance

Coefficients de régression et tests

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Signification
1	(constante)	1467,643	62,422		23,512	,000
	TEMPERAT	-37,060	2,295	-,866	-16,147	,000
	ISOLATIO	-29,774	3,492	-,457	-8,526	,000

a. Variable dépendante : CONSOMMA

Valeur des coefficients

Écart-type des estimateurs

Coefficients comparables entre eux

Statistique t de Student

Une valeur t > 2 ou t < - 2 est significative à 95 % d'un coeff ≠ 0

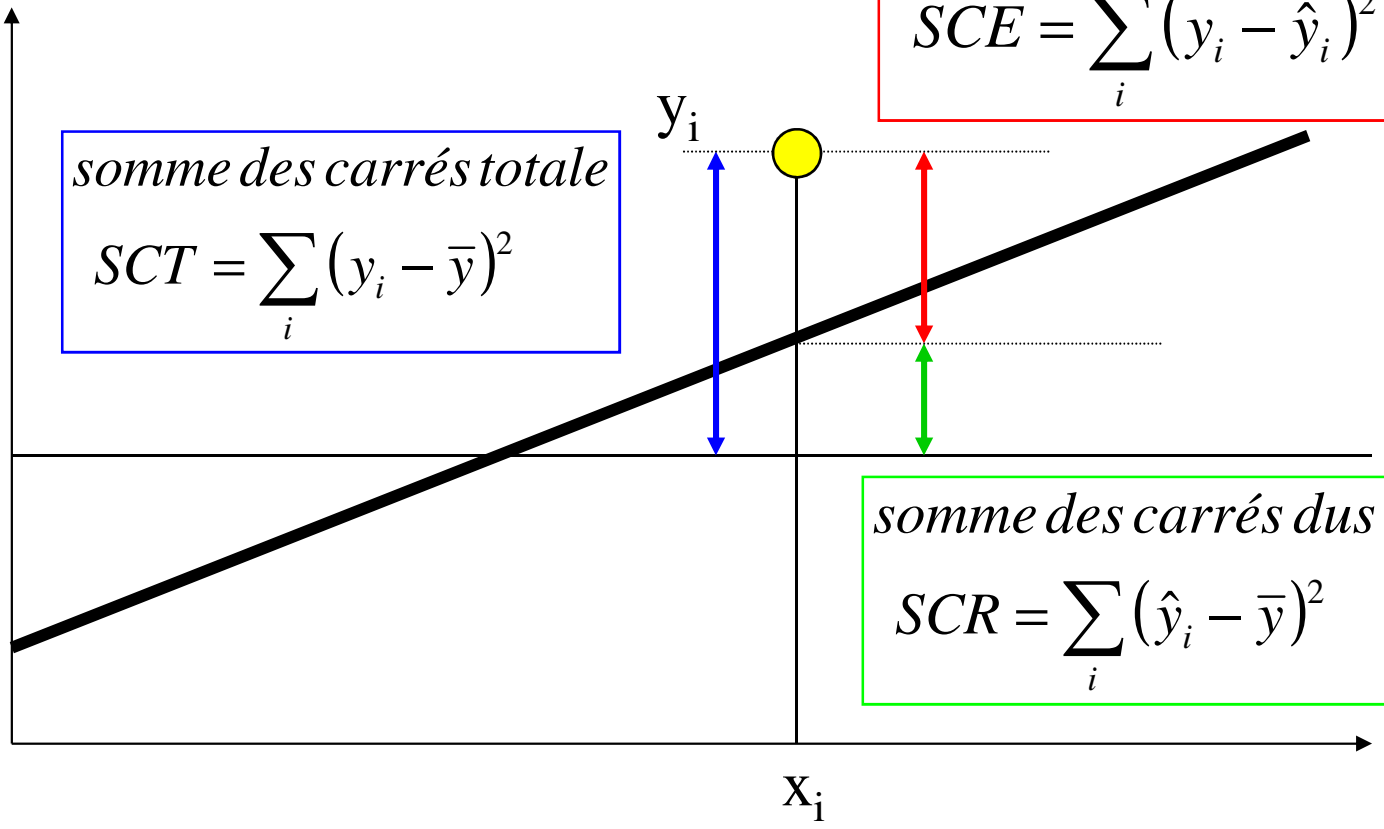
Sommes des carrés

somme des carrés résiduels

$$SCE = \sum_i (y_i - \hat{y}_i)^2$$

somme des carrés totale

$$SCT = \sum_i (y_i - \bar{y})^2$$



somme des carrés dus à la rég.

$$SCR = \sum_i (\hat{y}_i - \bar{y})^2$$

Test global du modèle

ANOVA[b]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	3267046,7	2	1633523,333	167,933	,000[a]
	Résidu	116727,068	12	9727,256		
	Total	3383773,7	14			

a. Valeurs prédites : (constantes), ISOLATIO, TEMPERAT

b. Variable dépendante : CONSOMMA

2 prédicteurs ⇒ régression linéaire multiple

$$F = \frac{\frac{SCR}{p}}{\frac{SCE}{n-p-1}}$$

SCR = somme des carrés « Régression »

SCE = somme des carrés « Erreurs »

p = nombre de variables

n = nombre d'observations

R^2 = SCR / SCT = 1 – (SCE / SCT)

suit une loi F de ddl (p,n-p-1)

sous l'hypothèse nulle (H_0) :
($b_1 = b_2 = 0$)

variance s^2 du terme d'erreur = $98{,}627^2$

Coefficient de détermination

- R^2 = SCR / SCT
- R^2 = proportion de variation de la variable cible expliquée par tous les prédicteurs (syn : régresseurs)
- Bon ajustement si R^2 proche de 1
- R^2 est biaisé (optimiste car croissant avec le nb de variables) et on lui substitue le R^2 ajusté :

$$R^2 ajusté = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}$$

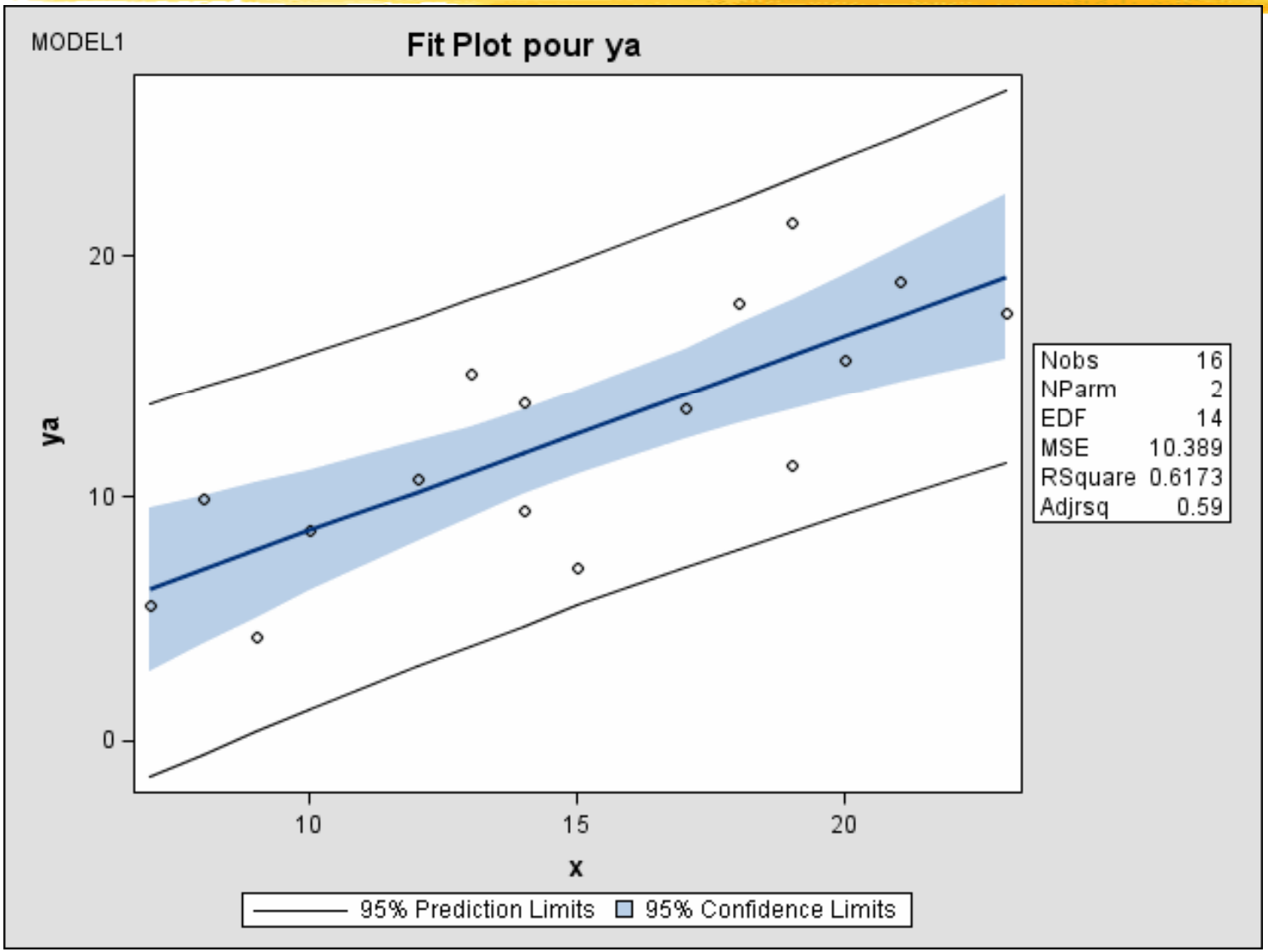
- R^2 ajusté est toujours < R^2 et peut être < 0

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,983[a]	,966	,960	98,627

Intervalles de confiance

- $\hat{y}_0 = a + bx_0$ est une prévision de Y et de la moyenne E(Y) en tout point x_0 de l'intervalle de mesure (car $E(\varepsilon_i) = 0$)
- D'après les formules sur les variances des estimateurs, les IC à (100-α) % de E(Y) et Y au point X_0 sont :
  - $\hat{y}_0 \pm t_{\alpha/2,n-p-1}. s [1/n + (x_0 - \overline{x})^2 / \Sigma (x_i - \overline{x})^2]^{1/2}$ pour E(Y)
  - $\hat{y}_0 \pm t_{\alpha/2,n-p-1}. s [1 + 1/n + (x_0 - \overline{x})^2 / \Sigma (x_i - \overline{x})^2]^{1/2}$ pour Y (on a ajouté la variance du terme d'erreur)
- Autrement dit, la variance de la valeur prédite pour une observation est :
  - $s^2 [1 + 1/n + (x_0 - \overline{x})^2 / \Sigma (x_i - \overline{x})^2]$
- > Plus difficile d'estimer une valeur possible de Y sachant $X=x_0$ que la moyenne des valeurs possibles sachant $X=x_0$
- > L'IC augmente quand x_0 s 'éloigne de $\overline{x}$

IC de la moyenne et des observations



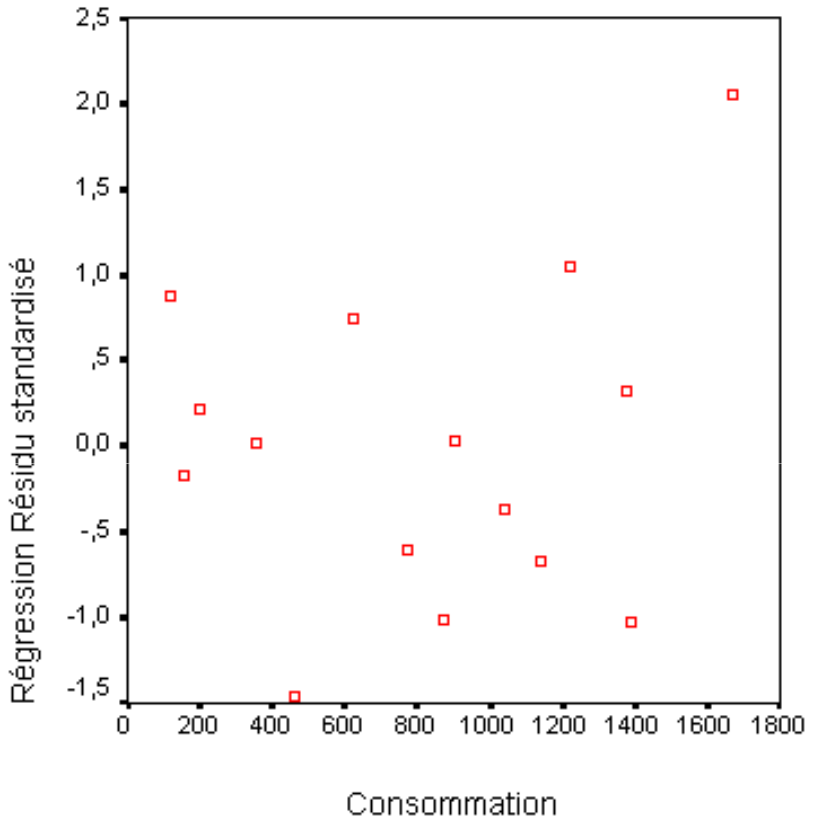
Précautions d'utilisation

- Le modèle n'est valide que sur l'étendue des observations, et surtout près de la moyenne de X
- Un petit échantillon (< 20) ne détecte que les relations fortes ; un grand détecte toutes les relations même faibles (rejet de H_0 malgré petit R^2)
- Minimum de 5 observations (mieux vaut en avoir > 15)
- Attention aux résidus standardisés (résidu / s) > 3
- Pour savoir si les extrêmes ont une influence : les enlever et voir les coeff. restent dans les IC des coeff. initiaux
- Attention aux distances de Cook > 1
  - la distance de Cook d'une observation i mesure l'écart des coefficients avec et sans cette observation
- Régression multiple : vérifier l'absence de multicolinéarité

Analyse des résidus

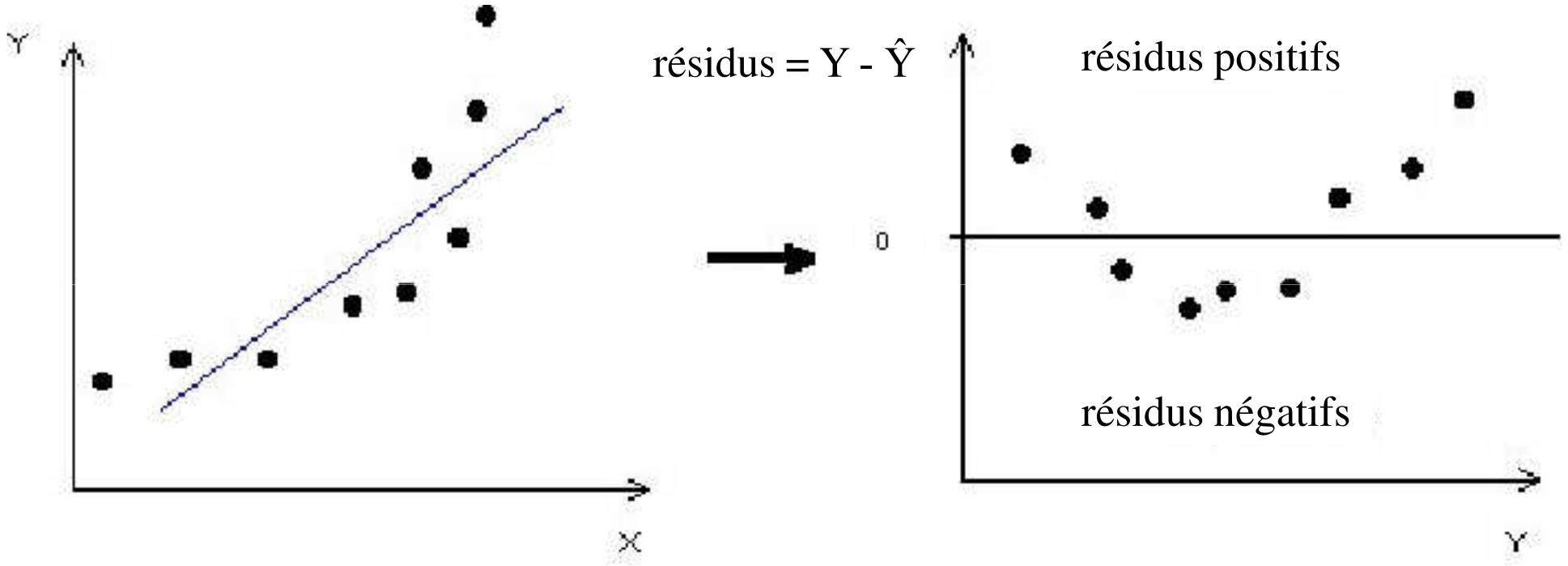
Vérification du respect des hypothèses de base

- Test d'**autocorrélation** (statistique de Durbin-Watson comprise entre 1,5 et 2,5)
- Test d'**homoscédasticité** (égalité de la variance en fonction de y)
- Test de **normalité** (test de Kolmogorov)
- Vérification d'absence de points extrêmes
- Un diagramme des résidus est souvent très parlant



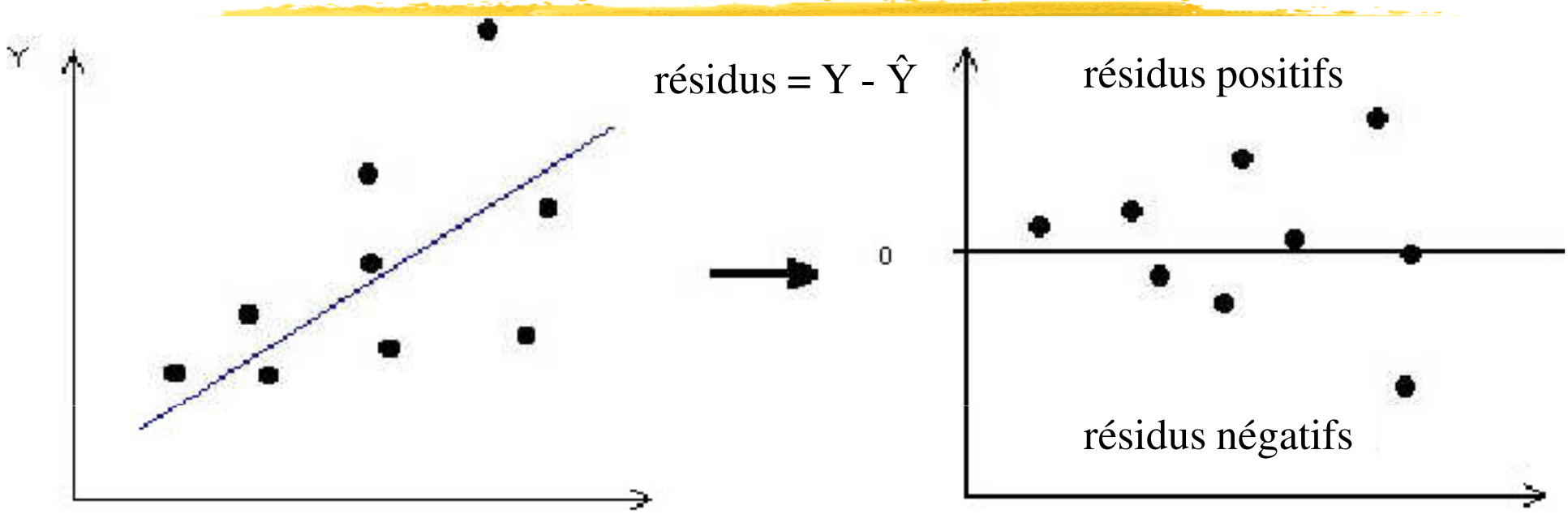
Les résidus standardisés doivent être répartis aléatoirement autour de 0 et rester dans les bornes [-3 ; +3]

Problème 1 : Autocorrélation des résidus



Corrélation entre ε_i et $\varepsilon_{i+1} \Rightarrow$ les valeurs moyennes de Y sont sur-estimées ; les autres sont sous-estimées

Problème 2 : Hétéroscédasticité des résidus



Appliquer le test de Levene en regroupant en classes les valeurs de Y

Estimation précise de Y en fonction de X lorsque Y est petit ; grande incertitude quand Y est grand

⇒remplacer Y par son log, son inverse ou sa racine carrée

(ou par le carré ou l'exponentielle quand la variance diminue)

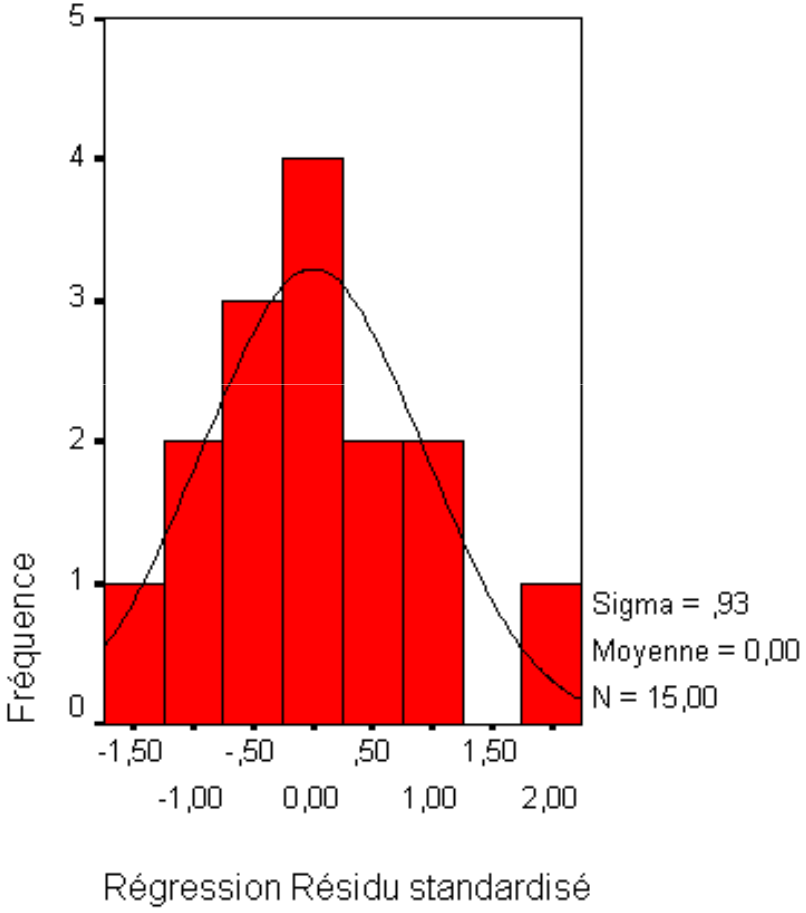
⇒ ou utiliser la méthode des moindres carrés pondérés

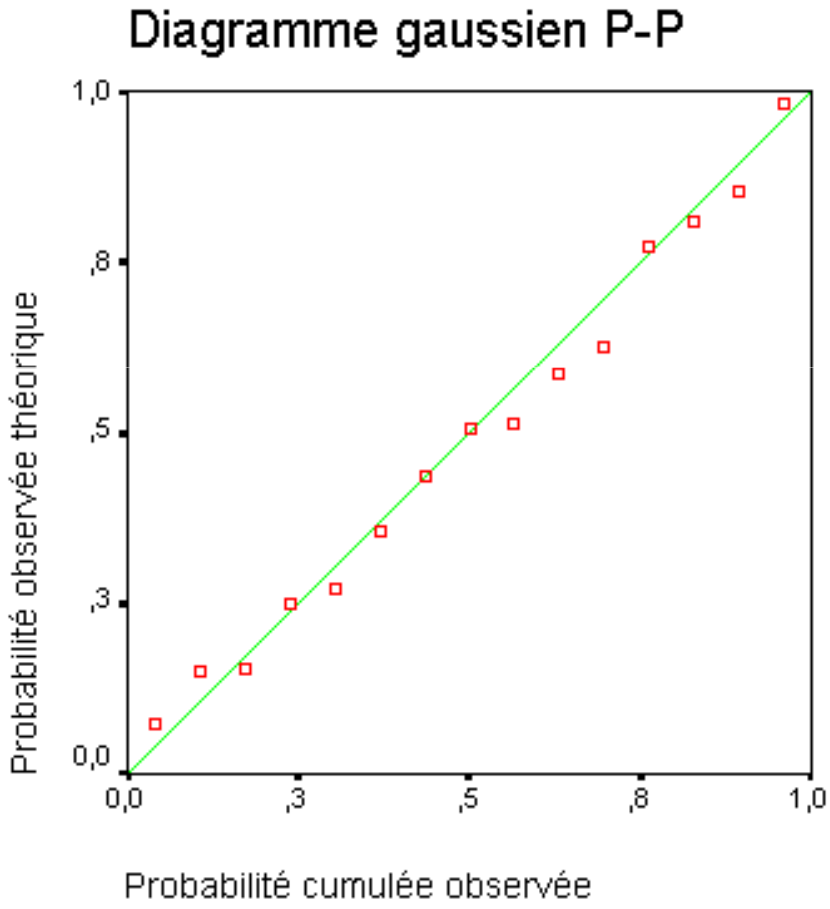
Homoscédasticité et autocorrélation des résidus

- Utiliser un diagramme des résidus pour vérifier l'homoscédasticité et l'absence d'autocorrélation
- Statistique de Durbin-Watson pour l'autocorrélation :
  - $= \Sigma (\varepsilon_i - \varepsilon_{i-1})^2 / \Sigma \varepsilon_i^2$
  - vaut entre 0 et 4
  - proche de 2 si pas d'autocorrélation (OK entre 1,5 et 2,5)
  - < 2 pour des corrélations positives
  - > 2 pour des corrélations négatives

	R	R-deux	R-deux ajusté	Durbin-Watson
	,983[a]	,966	,960	1,819

Normalité des résidus





Utilité des tests sur les résidus 1/3

- Exemple tiré de :
  - Tomassone, Lesquoy, Millier : *La Régression nouveaux regards sur une ancienne méthode statistique*, 1986
  - Anscombe F.J. : *Graphs in Statistical Analysis*, 1973

	x	ya	yb	yc	yd	xe	ye
1	7	5.535	0.113	7.399	3.864	13.715	5.654
2	8	9.942	3.77	8.546	4.942	13.715	7.072
3	9	4.249	7.426	8.468	7.504	13.715	8.491
4	10	8.656	8.792	9.616	8.581	13.715	9.909
5	12	10.737	12.688	10.685	12.221	13.715	9.909
6	13	15.144	12.889	10.607	8.842	13.715	9.909
7	14	13.939	14.253	10.529	9.919	13.715	11.327
8	14	9.45	16.545	11.754	15.86	13.715	11.327
9	15	7.124	15.62	11.676	13.967	13.715	12.746
10	17	13.693	17.206	12.745	19.092	13.715	12.746
11	18	18.1	16.281	13.893	17.198	13.715	12.746
12	19	11.285	17.647	12.59	12.334	13.715	14.164
13	19	21.365	14.211	15.04	19.761	13.715	15.582
14	20	15.692	15.577	13.737	16.382	13.715	15.582
15	21	18.977	14.652	14.884	18.945	13.715	17.001
16	23	17.69	13.947	29.431	12.187	33.281	27.435

Utilité des tests sur les résidus 2/3

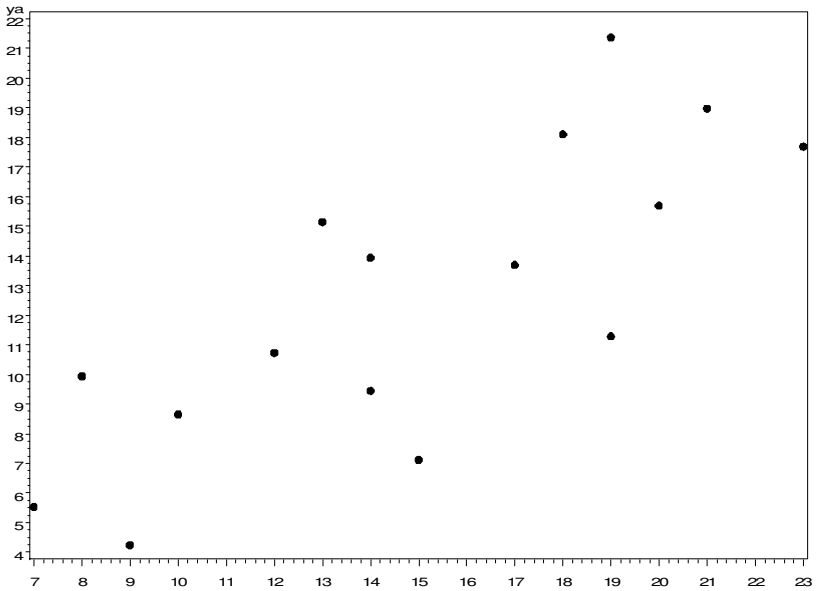
- Dans les 5 régressions : mêmes sommes de carrés, même variance résiduelle, même F-ratio, mêmes R^2, même droite de régression, mêmes écarts-types des coefficients…

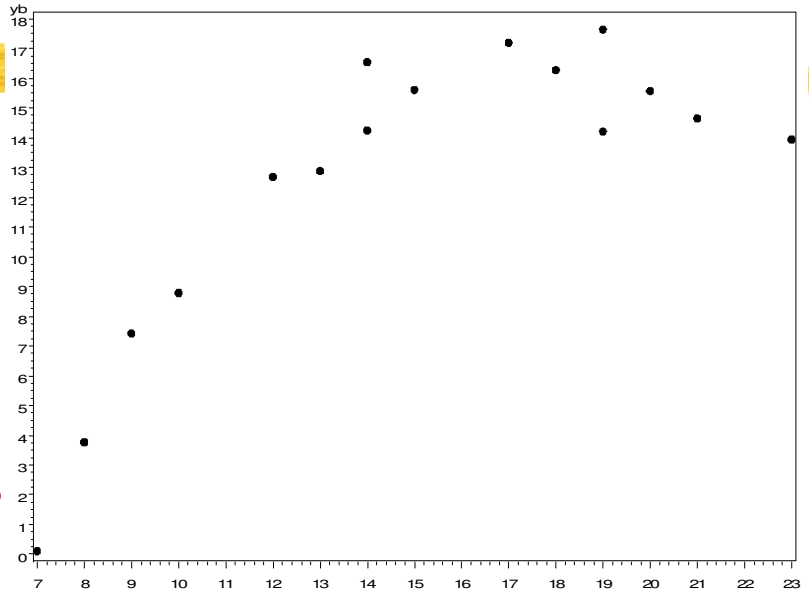
Analyse de variance					
Source	**DF**	**Somme des carrés**	**Carré moyen**	**Valeur F**	**Pr > F**
Model	1	234.6	234.6	22.6	0.0003
Error	14	145.4	10.4		
Corrected Total	15	380.1			

Root MSE	3.22	**R-Square**	0.62
Dependent Mean	12.60	**Adj R-Sq**	0.59
Coeff Var	25.60		

Résultats estimés des paramètres							
Variable	**DF**	**Résultat estimé des paramètres**	**Erreur std**	**Valeur du test t**	**Pr > |t|**	**Tolérance**	**Inflation de variance**
Intercept	**1**	0.52	2.67	0.20	0.8476	.	0
x	**1**	0.81	0.17	4.75	0.0003	1.00	1.00

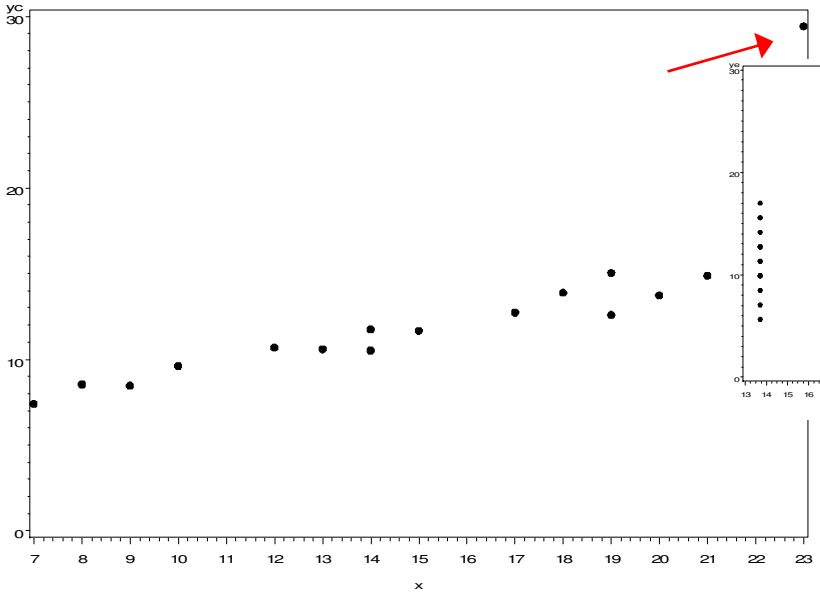
Utilité des tests sur les résidus 3/3

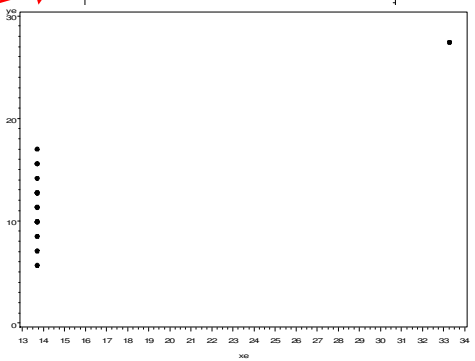


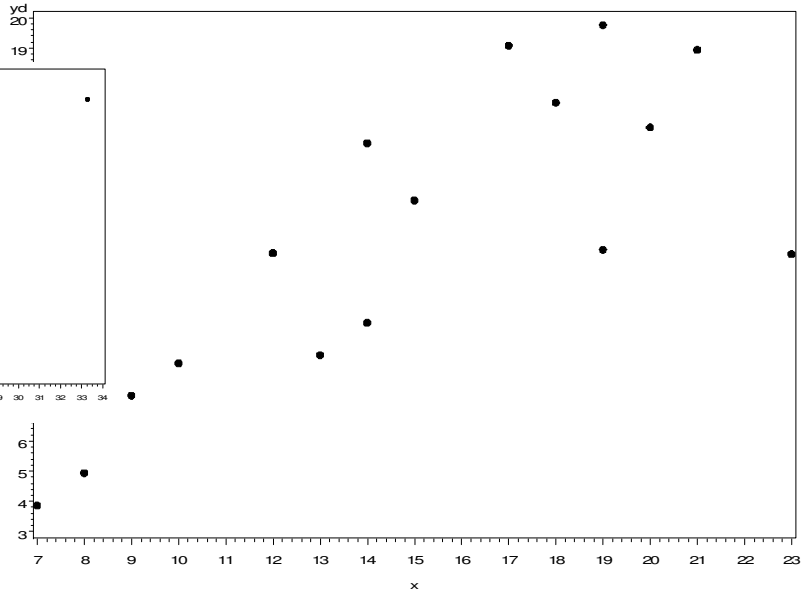


Et pourtant !

Les situations sont bien différentes

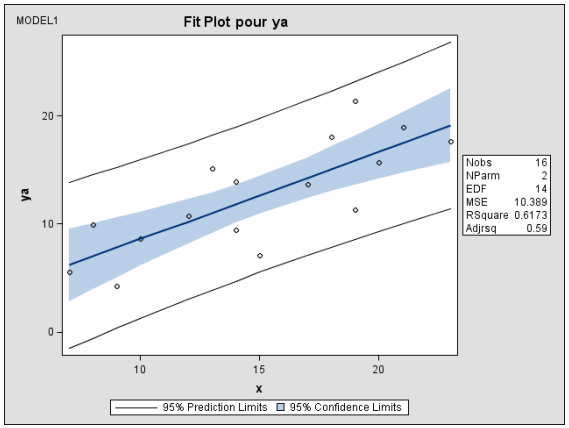


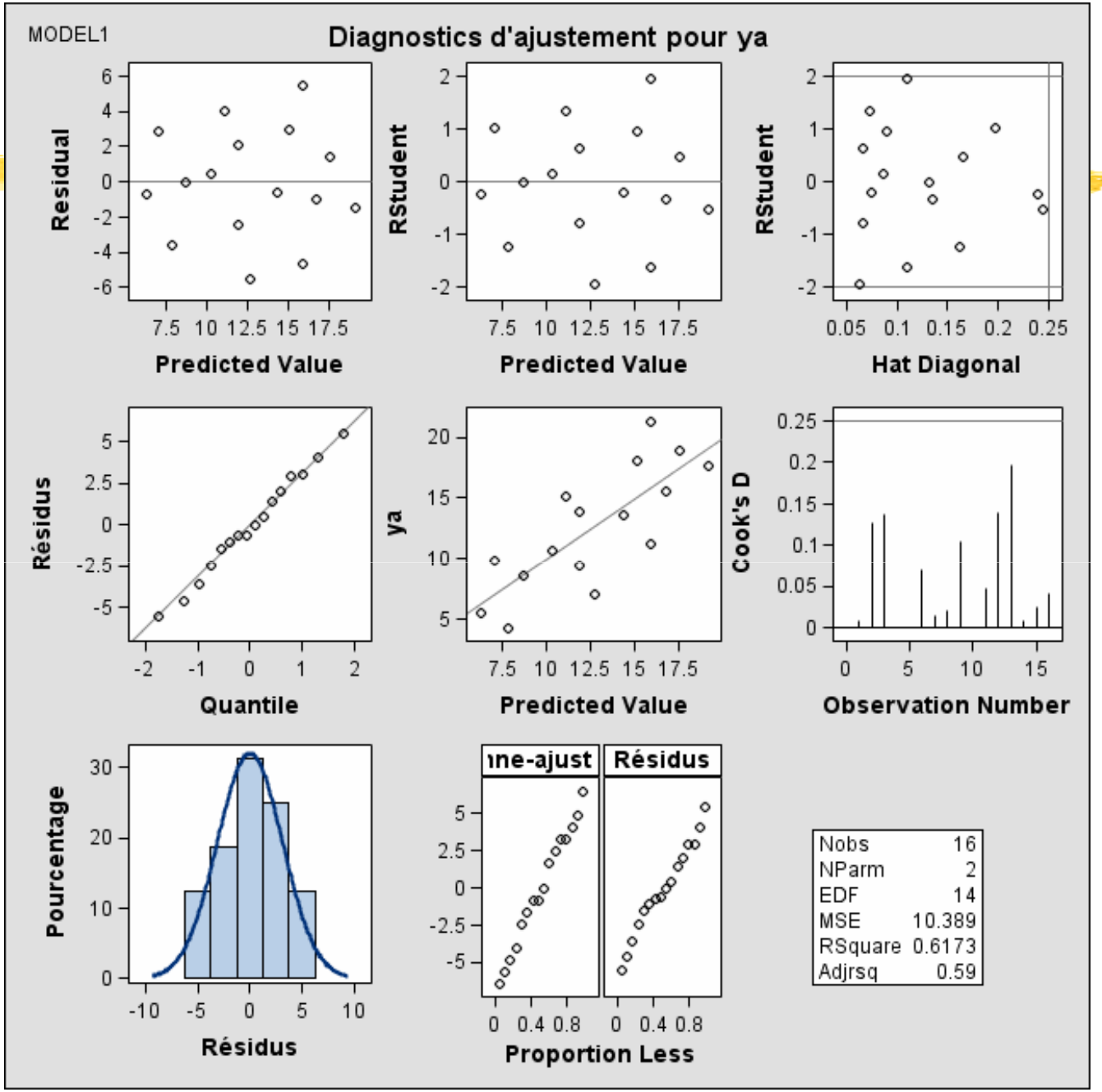




Régression 1 :

Durbin-Watson D	2.538
Number of Observations	16
1st Order Autocorrelation	-0.277

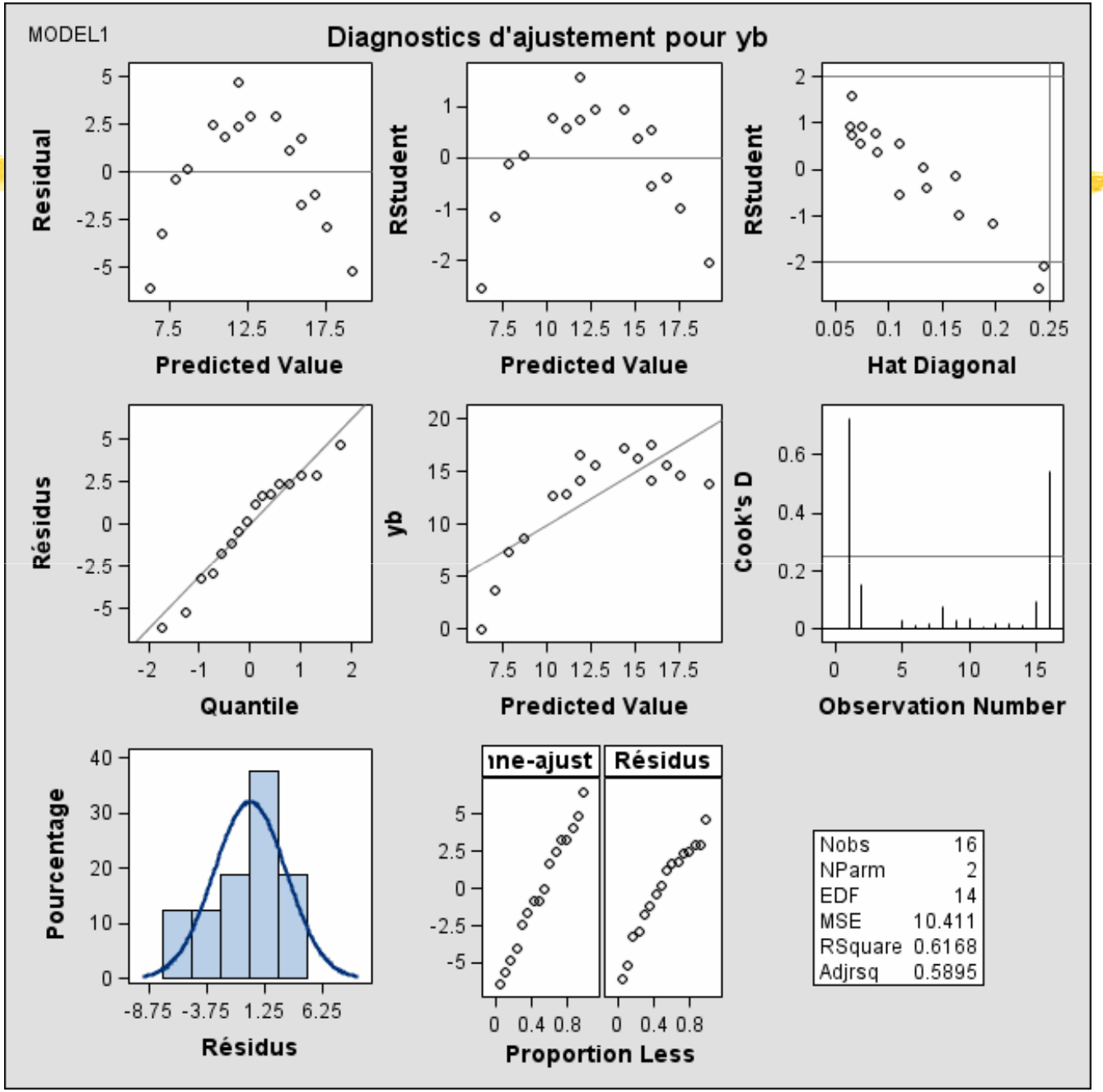


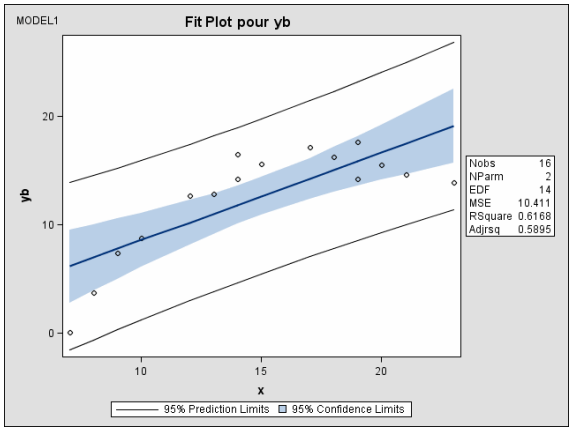


Régression 2 :

Forte auto-corrélation positive !

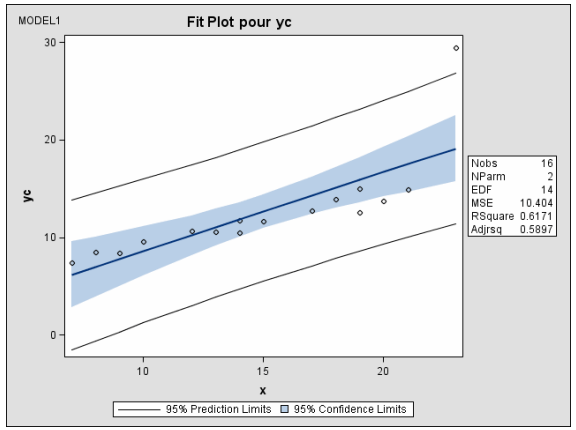
Durbin-Watson D	0.374
Number of Observations	16
1st Order Autocorrelation	0.595

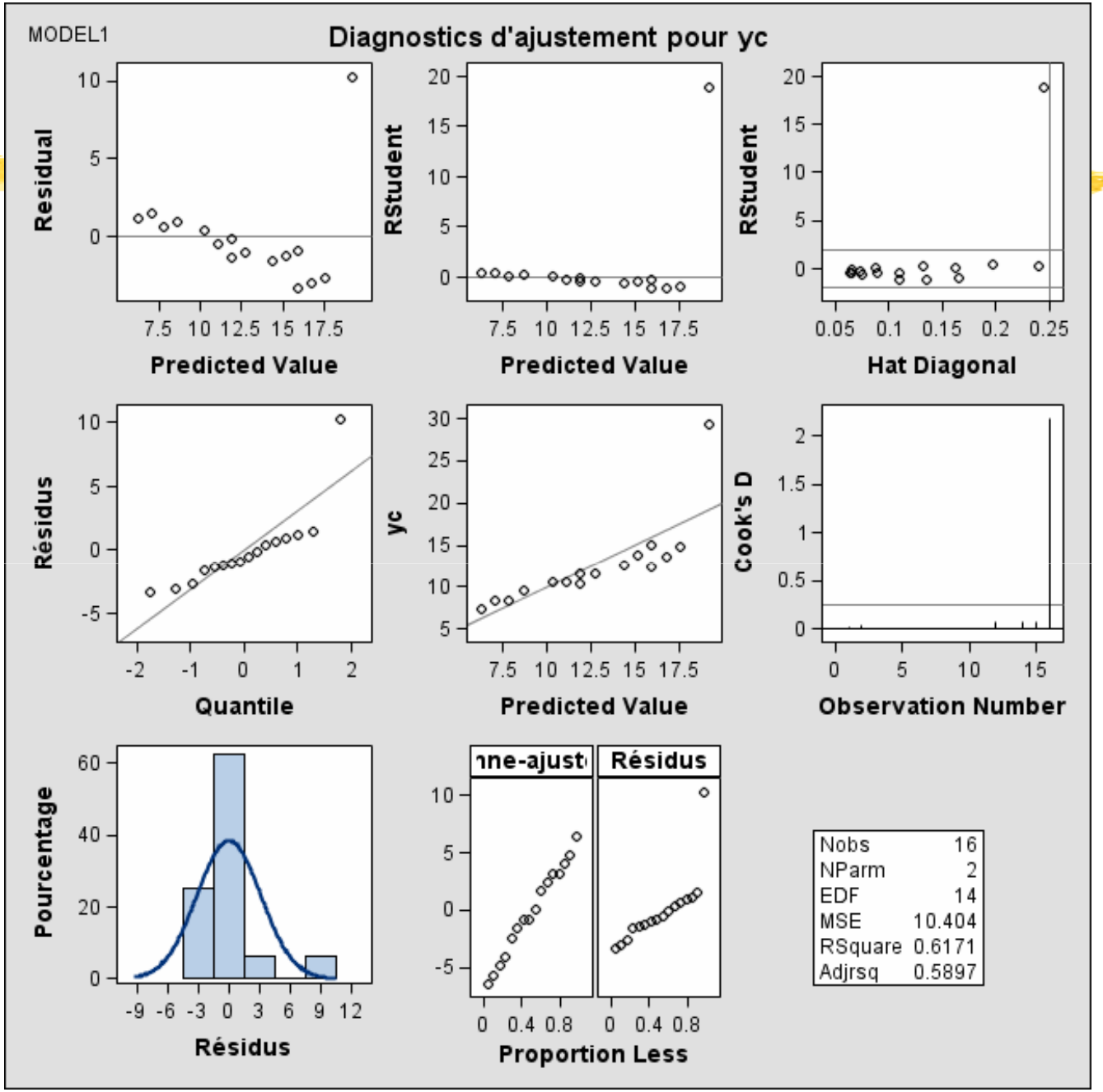




Régression 3 :

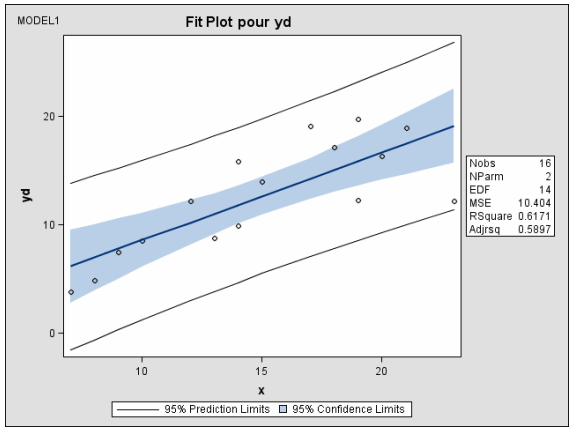
Durbin-Watson D	1.289
Number of Observations	16
1st Order Autocorrelation	-0.015

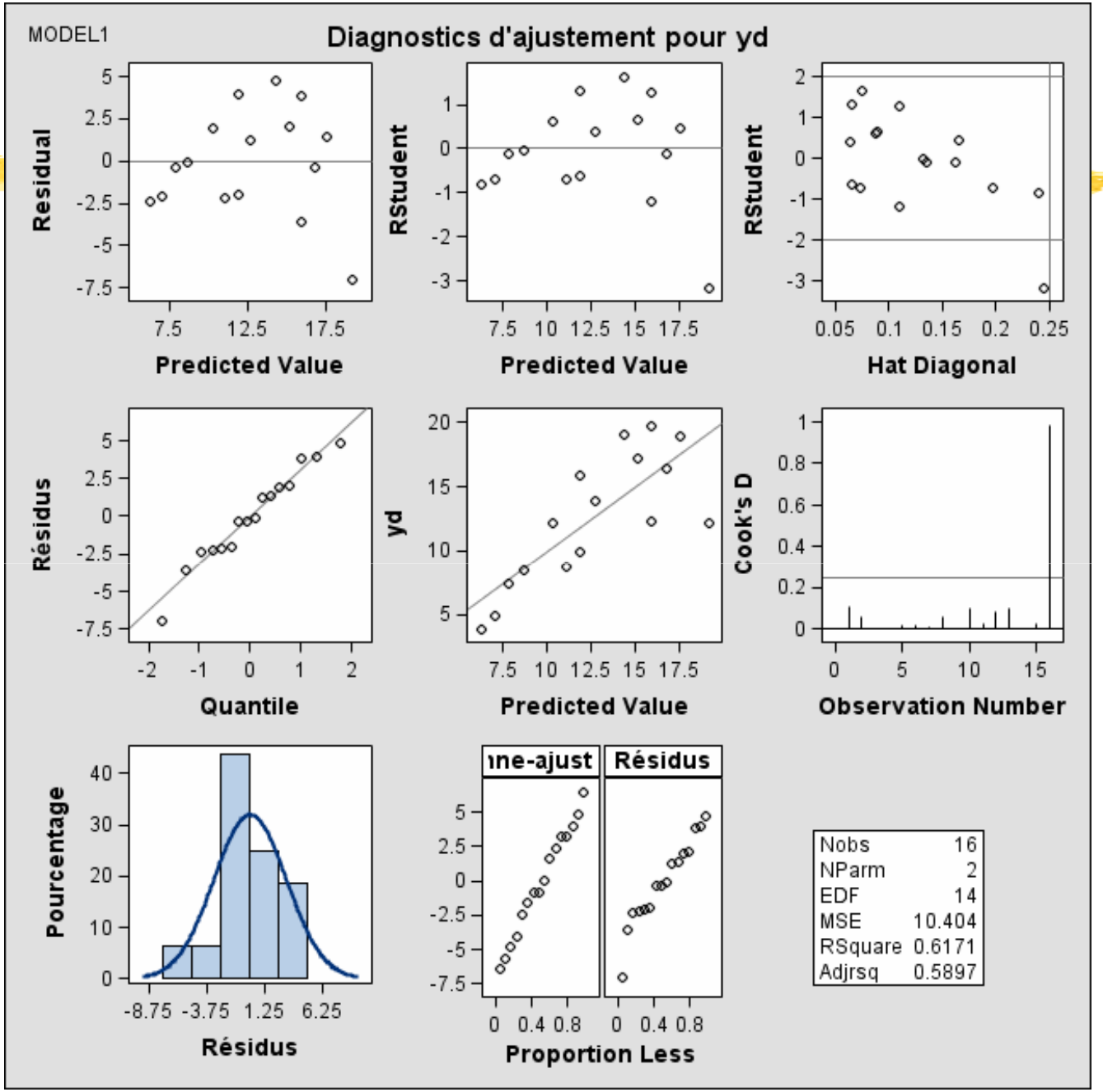




Régression 4 :

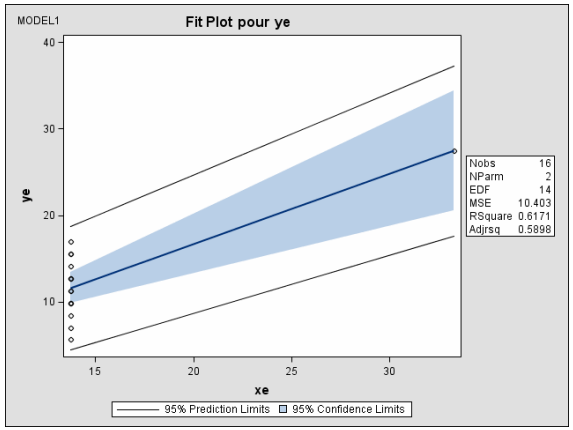
Durbin-Watson D	1.821
Number of Observations	16
1st Order Autocorrelation	-0.094

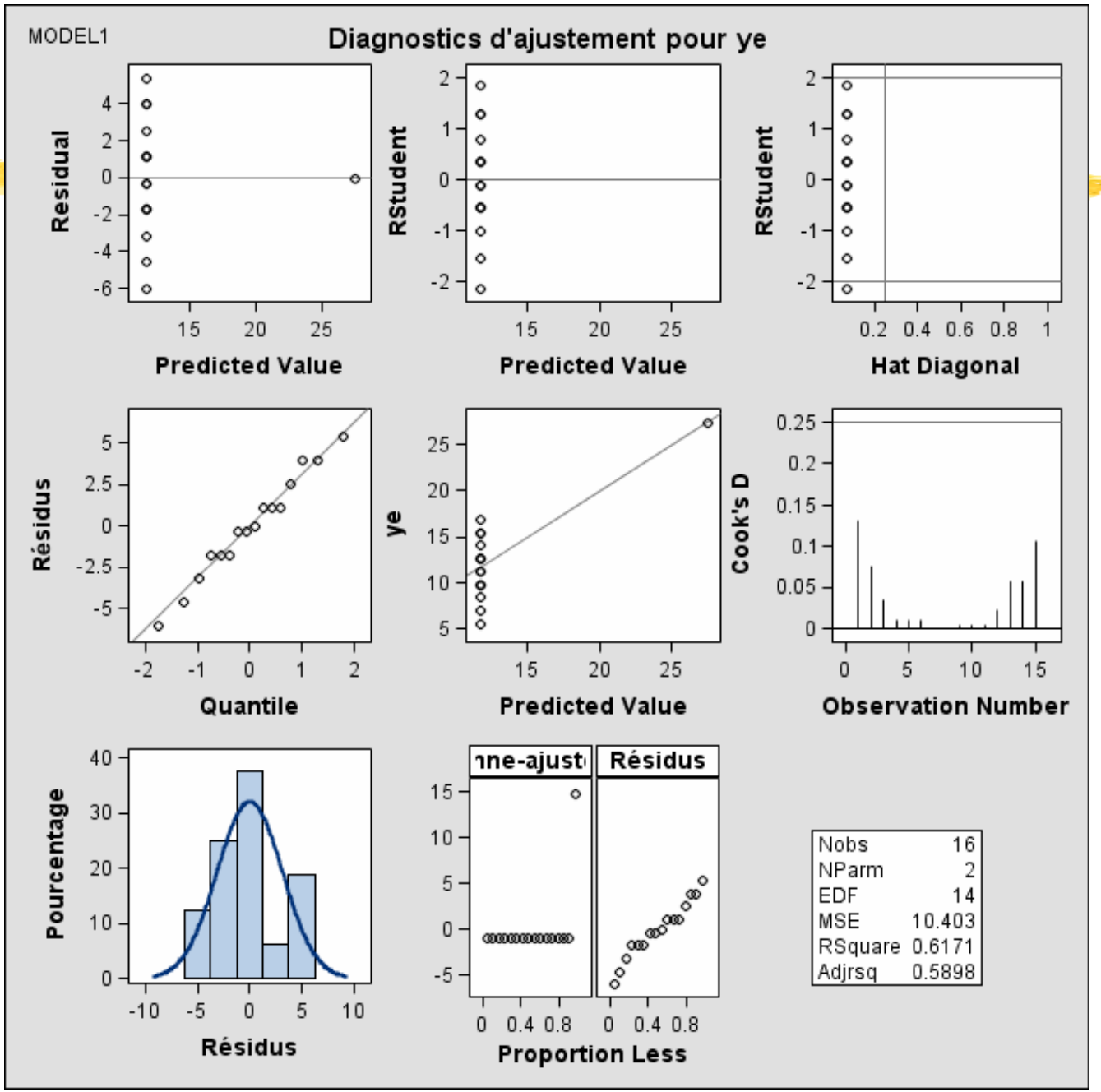




Régression 5 :

Durbin-Watson D	0.310
Number of Observations	16
1st Order Autocorrelation	0.723





Attention à la multicolinéarité

- Multicolinéarité = plusieurs variables explicatives (fortement) corrélées entre elles.
- Cela entraîne :
  - des coefficients de régression très sensibles aux fluctuations même faibles des données
  - des écarts-types élevés pour les coefficients de régression
  - une dégradation de la précision des prévisions
- Mesurée par :
  - tolérance X_i = 1 - (coefficient de détermination de la régression de X_i sur les autres variables)
    - doit être > 0,2
  - VIF = 1 / tolérance
    - doit être < 5

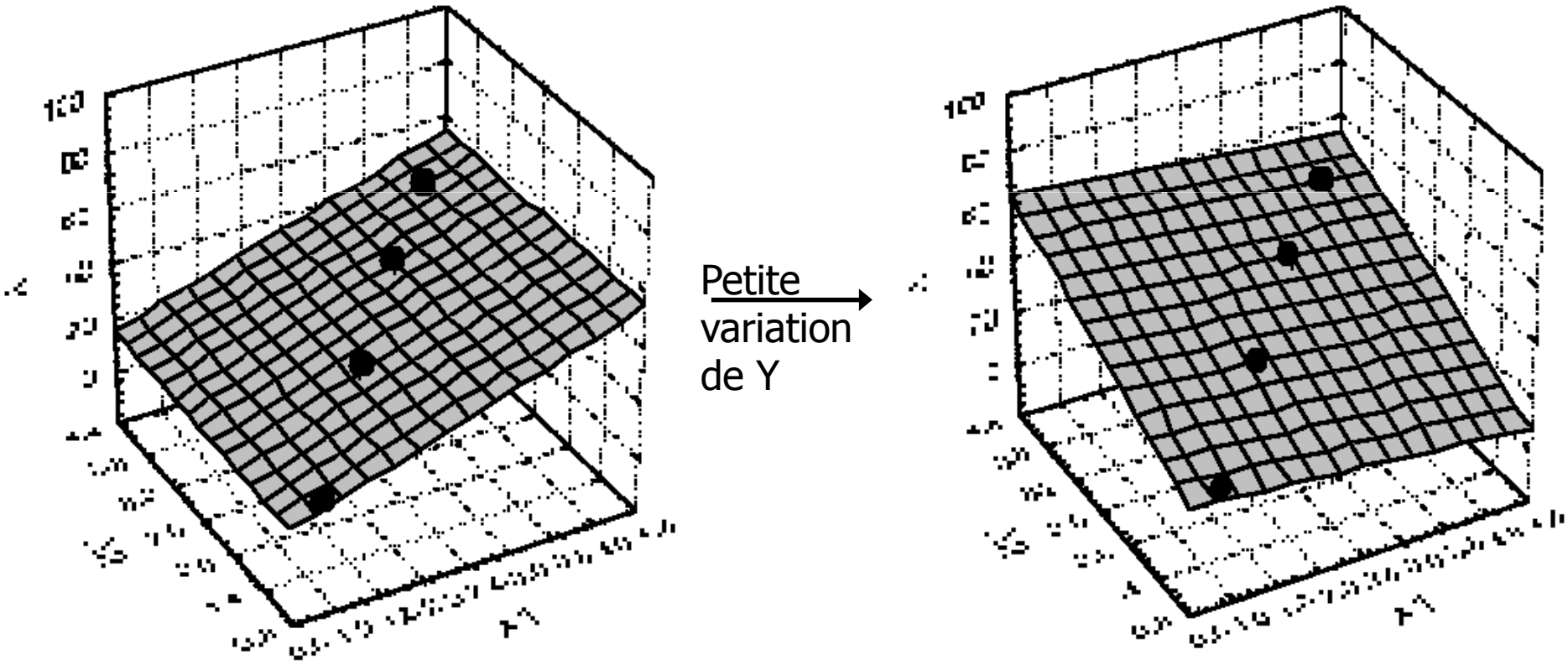
Attention à la multicolinéarité

- Autre mesure possible : les indices de conditionnement de la matrice des corrélations
  - on a multicolinéarité modérée (resp. forte) si présence d'indices η_k > 10 (resp. 30)
  - on regarde si on peut relier la valeur propre correspondante à une forte contribution (> 50 %) de la composante à la variance de 2 ou plusieurs variables

Modèle	Dimension	Valeur propre	Indice de conditionnement	Proportions de la variance		
				(constante)	TEMPERAT	ISOLATIO
1	1	2,145	1,000	,03	,07	,03
	2	,766	1,673	,02	,92	,02
	3	,089	4,915	,95	,01	,95

Effets de la multicolinéarité

- X_1 et X_2 presque colinéaires => coefficients de la régression très sensibles à de petites variations de Y



Solutions à la multicolinéarité

- Suppression des variables concernées
  - accepter de baisser un peu R^2 pour baisser la multicolinéarité
- Transformation (logarithme…) des variables concernées
- Régression biaisée (ridge)
  - l'erreur quadratique de l'estimation de la pente β de la régression = variance_estimateur + (biais_estimateur)², d'où une « erreur quadratique avec biais » < « erreur sans biais » si le biais est compensé par une faible variance
- Régression sur composantes principales
  - passer ensuite des coefficients de régression des composantes principales à ceux des variables initiales
- Régression PLS (Partial Least Squares)
  - utilisable même si : nb observations << nb variables
  - on démontre (De Jong, 1993) que la régression PLS sur k composantes est toujours plus prédictive que la régression sur les k premières composantes principales

Technique de prédiction :
La régression PLS

La méthode Partial Least Squares

- C'est une méthode qui se juxtapose à d'autres méthodes de régression (linéaire, logistique, analyse discriminante)
- Utile en présence d'un grand nombre de variables présentant de la colinéarité ou des valeurs manquantes
- Algorithme simple (suite de régressions simples, sans inversion ni diagonalisation de matrices) $\Rightarrow$ efficace sur de grands volumes de données
- Utilisation en chimie, industrie pétrolifère, cosmétique, biologie, médecine, agroalimentaire
  - en cosmétique : conserver tous les ingrédients d'un produit $\Rightarrow$ très nombreuses variables explicatives
  - en agroalimentaire (analyse sensorielle) : expliquer le classement d'un produit par plusieurs dégustateurs (variable Y), en fonction de ses propriétés (jusqu'à plusieurs centaines) physico-chimiques et de saveur

Principe de la régression PLS

- Régression PLS inventée par Herman et Svante Wold (1983)
- On a Y variable à expliquer et X_i variables explicatives
- Le choix des variables transformées résulte d'un compromis entre :
  - maximisation de la variance des X_i (ACP)
  - maximisation de la corrélation entre X_i et Y (régression)
  - donc : on cherche les combinaisons linéaires T_j des X_i maximisant $cov^2(T_j,Y) = r^2(T_j,Y).var(T_j).var(Y)$

Etape 1 de la régression PLS

- On cherche une combinaison $T_1 = \Sigma_i \lambda_{1i} X_i$ des X_i qui maximise la variance de T_1 et la corrélation entre T_1 et Y
  - $\Leftrightarrow$ maximiser $cov^2(T_1,Y) = r^2(T_1,Y).var(T_1).var(Y)$
- La solution est $\lambda_{1i} = cov(Y,X_i)$
  - en normant $||(\lambda_{11},..., \lambda_{1p})|| = 1$
  - on a donc $T_1 = \Sigma_i cov(Y,X_i).X_i$
- La régression de Y sur T_1 donne un résidu Y_1 :
  - $Y = c_1 T_1 + Y_1$
- La régression de X_i sur T1 donne aussi des résidus X_{1i} :
  - $X_i = c_{1i} T_1 + X_{1i}$
- On réitère en remplaçant Y par Y_1 et les X_i par les $X_{1i} \Rightarrow$ étape 2

Etape 2 de la régression PLS

- On répète la même opération en remplaçant Y par son résidu Y_1 et les X_i par leurs résidus X_{1i}
- On obtient une combinaison $T_2 = \Sigma_i \lambda_{2i} X_i$ en normant $||(\lambda_{21},..., \lambda_{2p})|| = 1$
- Puis on régresse Y_1 sur T_2 et les X_{1i} sur T_2 : on obtient des résidus Y_2 et X_{2i}
  - $Y_1 = c_2 T_2 + Y_2$
  - $X_{1i} = c_{2i} T_2 + X_{2i}$
- On réitère jusqu'à ce que le nb de composantes T_k donne un résultat satisfaisant (vérifié par validation croisée)
- A la fin, on a :
  - $Y = c_1 T_1 + Y_1 = c_1 T_1 + c_2 T_2 + Y_2 = \Sigma_j c_j T_j$ + résidu
- Et on remplace cette expression par une expression de la régression de Y en fonction des X_i

Choix du nombre de composantes 1/2

- On procède généralement par validation croisée
- On se place à l'étape h et on veut décider de conserver ou non la composante h
- On calcule la somme des carrés résiduels (REsidual Sum of Squares), comme en régression linéaire :

$$RESS_h = \Sigma_k (y_{(h-1),k} - \hat{y}_{(h-1),k})^2$$

où $\hat{y}_{(h-1),k} = c_h t_{h,k}$ = prévision de $y_{(h-1),k}$ calculée pour chaque observation k

- Ensuite, les observations sont partagées en G groupes, et on réalise G fois l'étape courante de l'algorithme PLS sur Y_{h-1} et les $X_{h-1,i}$ en ôtant chaque fois un groupe

Choix du nombre de composantes 2/2

- Puis on calcule la somme *prédite* des carrés résiduels (Predicted REsidual Sum of Squares) $PRESS_h$
- Analogue à la précédente mais qui évite le sur-apprentissage en remplaçant la prévision $\hat{y}_{(h-1),k}$ par la prévision $\hat{y}_{(h-1),-k}$ déduite de l'analyse réalisée sans le groupe contenant l'observation k
- $PRESS_h = \Sigma_k(y_{(h-1),k} - \hat{y}_{(h-1),-k})^2$
- On retient la composante h si : $PRESS_h \leq \gamma.RESS_{h-1}$
  en posant $RESS_0 = \sum (y_i - \bar{y})^2$
  Souvent : on fixe $\gamma = 0{,}95$ si $n < 100$, et $\gamma = 1$ si $n \geq 100$

Nombre de composantes PLS

- Cette sélection par validation croisée permet de retenir un nombre de composantes :
  - suffisamment grand pour expliquer l'essentiel de la variance des X_i et de Y
  - suffisamment petit pour éviter le sur-apprentissage
- En pratique le nombre de composantes dépasse rarement 3 ou 4
- Notons également que la régression PLS sur k composantes est toujours plus prédictive que la régression sur les k premières composantes principales

Généralisations de la régression PLS

- Régression PLS2 développée pour prédire plusieurs Y_j simultanément
  - on peut avoir nb(Y_j) >> nb observations
- Régression logistique PLS développée par Michel Tenenhaus (2000)
  - algorithme analogue au précédent
- Et régression logistique sur composantes PLS, équivalente à la régression logistique PLS mais plus simple :
  - on commence par une régression PLS de l'indicatrice de Y sur les X_i (ou des indicatrices de Y, si Y a plus de 2 modalités)
  - on obtient k composantes PLS (éventuellement : k = 1)
  - puis on effectue une régression logistique de Y sur les composantes PLS

Technique de prédiction :
La régression robuste

Régression robuste

- Méthodes valides quand les résidus des observations ne suivent pas une loi normale
- Peu sensibles aux « outliers »
- De plus en plus répandues dans les logiciels statistiques
  - SAS, R, S-PLUS, STATA…

Algorithmes de régression robuste

- Moindres médianes de carrés
- Moindres carrés winsorisés (least winsored squares)
  - remplacement des x centiles extrêmes par Q_x
- Moindres carrés écrêtés (least trimmed squares)
  - suppression des x centiles extrêmes
- Moindres carrés pondérés
  - par l'inverse de la variance de la variable à expliquer, pour compenser l'hétéroscédasticité, en posant par ex. $p_i = s^2/s_i^2$ au voisinage d'un point x_i
- Moindres carrés localement pondérés sur les voisins (LOESS)
- Doubles moindres carrés

- Régression spline
- Méthode du noyau

Autres algorithmes de régression

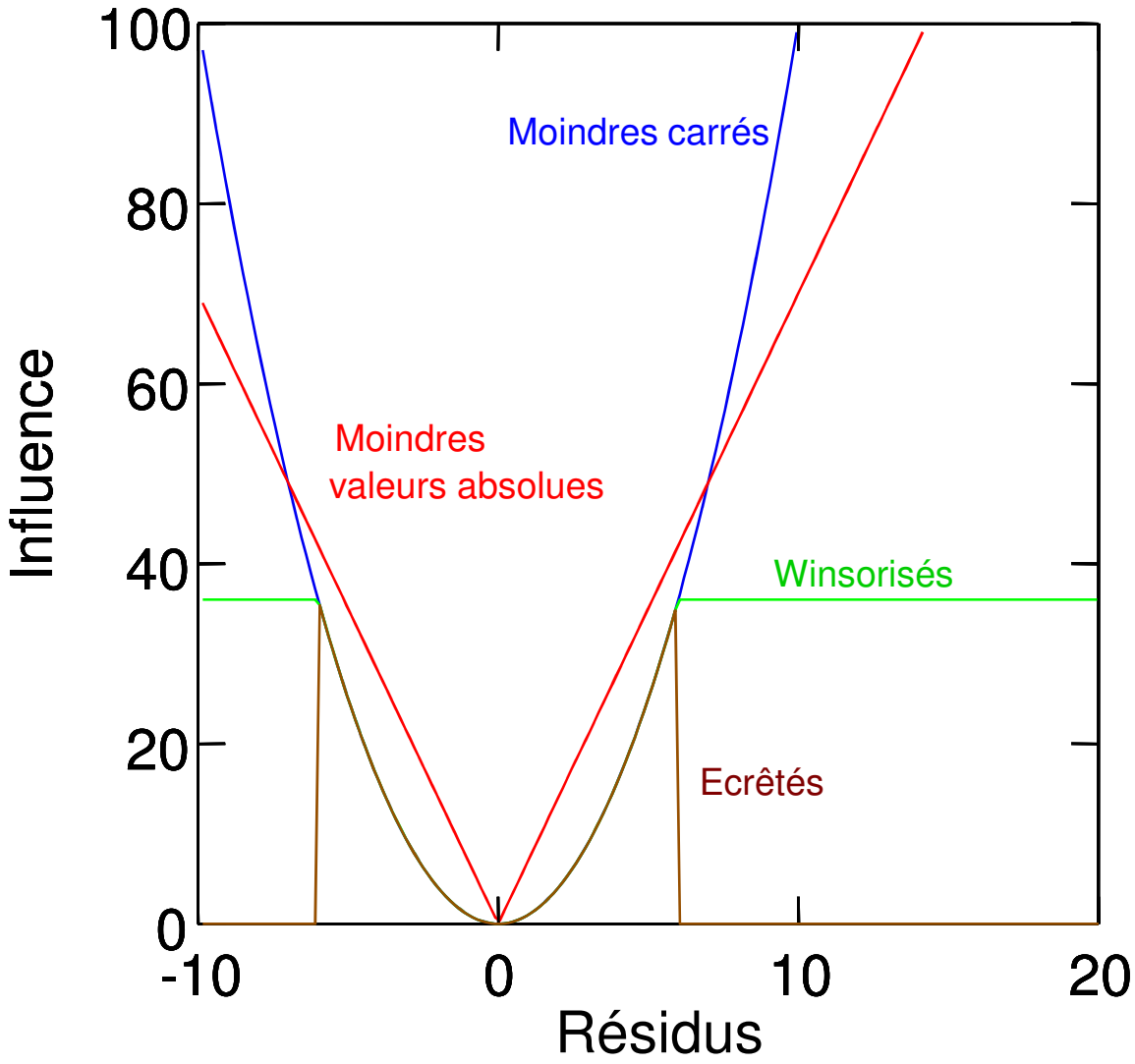
- Moindres valeurs absolues

$$\sum_i \left| x_i - \overline{x} \right|$$

- Régression polynomiale

- Régression sur variables qualitatives par codage optimal (moindres carrés alternés)

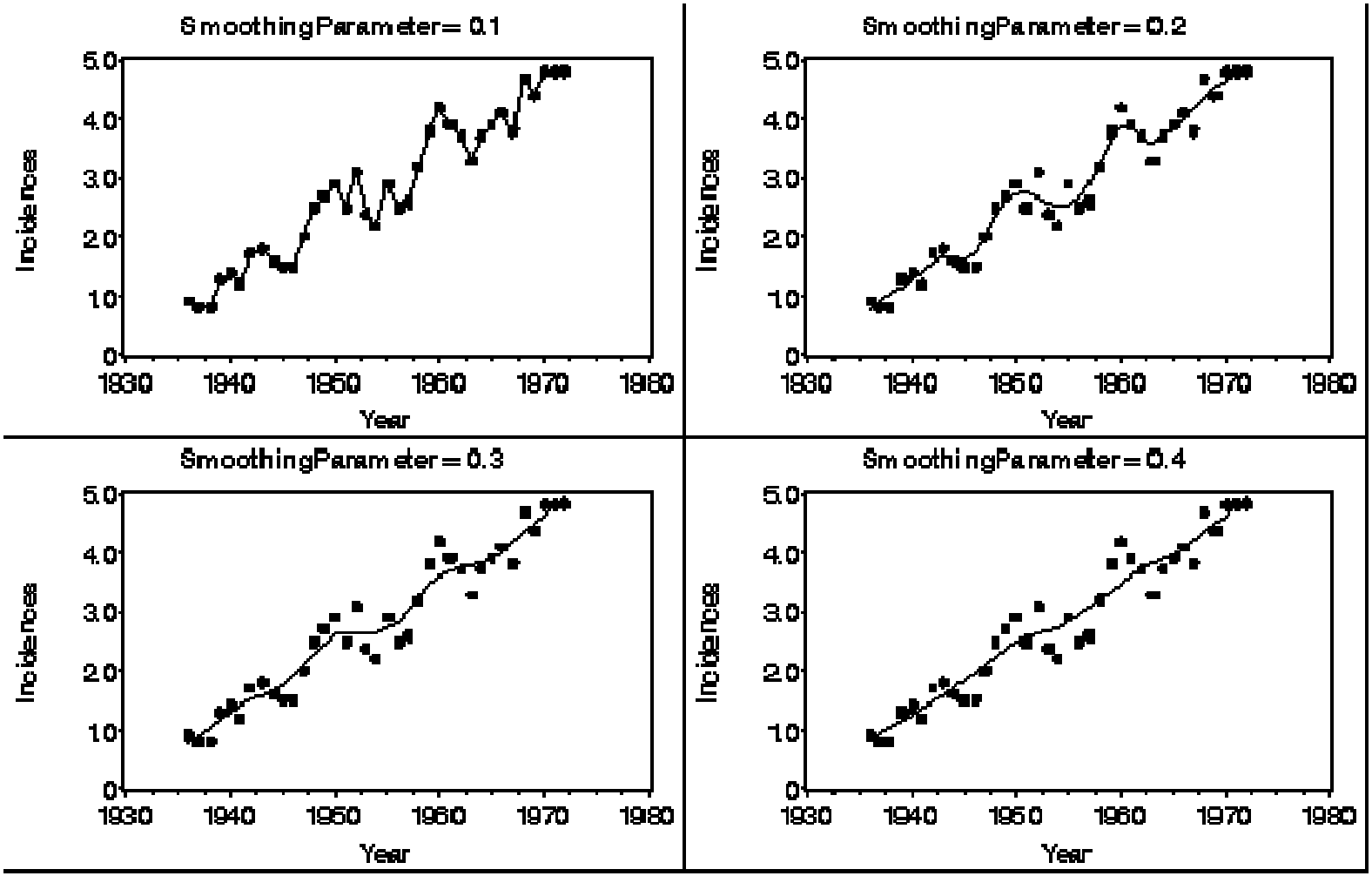
Influence des résidus



Principe de la régression LOESS

- Pour chaque point x : on prend les n voisins
  - le nombre n est choisi pour représenter un certain % de l'ensemble des points
  - ce % est appelé « paramètre de lissage » (« smoothing parameter »)
    - il existe des critères pour le choix de ce paramètre
- On pondère chacun de ces n points selon une fonction décroissante de leur distance à x
- On calcule la régression pondérée sur les n voisins pour prédire x
- LOESS utilisable avec plusieurs régresseurs
- Initiateur : Cleveland (1979)

Exemples de régressions LOESS



Technique de classement :
Analyse discriminante

Deux problématiques

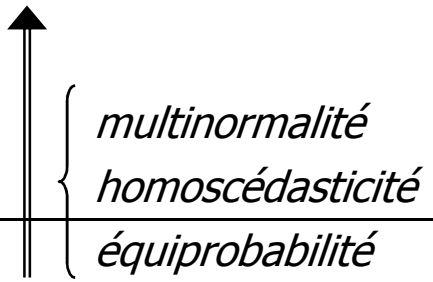
- Situation : on a un ensemble d'individus appartenant chacun à un groupe, le nb de groupes étant fini et > 1
- Analyse discriminante descriptive : trouver une représentation des individus qui sépare le mieux les groupes
- Analyse discriminante prédictive : trouver des règles d'affectation des individus à leur groupe
- L'analyse discriminante offre une solution à ces deux problématiques

Autre formulation

- Situation : on a un ensemble d'individus caractérisés par une variable à expliquer Y qualitative et des variables explicatives X_i quantitatives
- Analyse discriminante descriptive : trouver une représentation des liaisons entre Y et les X_i
- Analyse discriminante prédictive : trouver des règles de prédiction des modalités de Y à partir des valeurs des X_i
- Cette formulation est équivalente à la précédente

Les différentes formes d'analyse discriminante

	Méthode descriptive (représenter les groupes)	**Méthode prédictive** (prédire l'appartenance à un groupe)
Approche géométrique	Oui analyse factorielle discriminante	Oui analyse discriminante linéaire
Approche probabiliste (bayésienne)	Non	Oui analyse discriminante linéaire a. d. quadratique a. d. non paramétrique régression logistique



Technique de classement :

Analyse discriminante géométrique

L'analyse discriminante géométrique

- Y variable cible qualitative à k modalités
  - correspondant à k groupes G_i
- X_j p variables explicatives continues
- Principe de l'analyse factorielle discriminante : remplacer les X_j par des axes discriminants : combinaisons linéaires des X_j prenant les valeurs les + différentes possibles pour des individus différant sur la variable cible
- Remarquer l'analogie avec l'ACP
- On a k-1 axes (si nb individus n > p > k)
- Exemple historique : les iris de Fisher (3 espèces – 4 variables, longueur et largeur des pétales et des sépales)

Exemple historique : les iris de Fisher

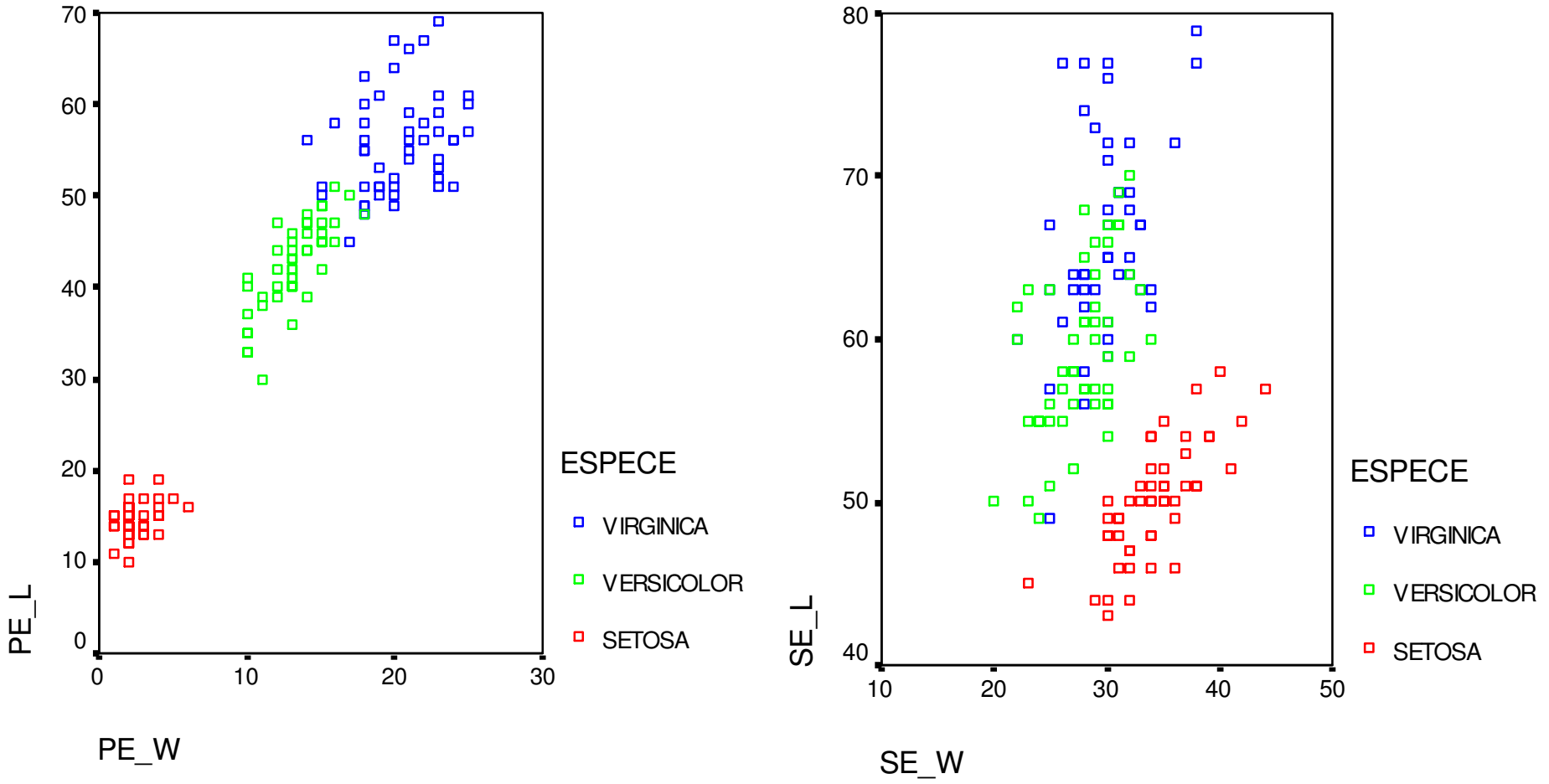
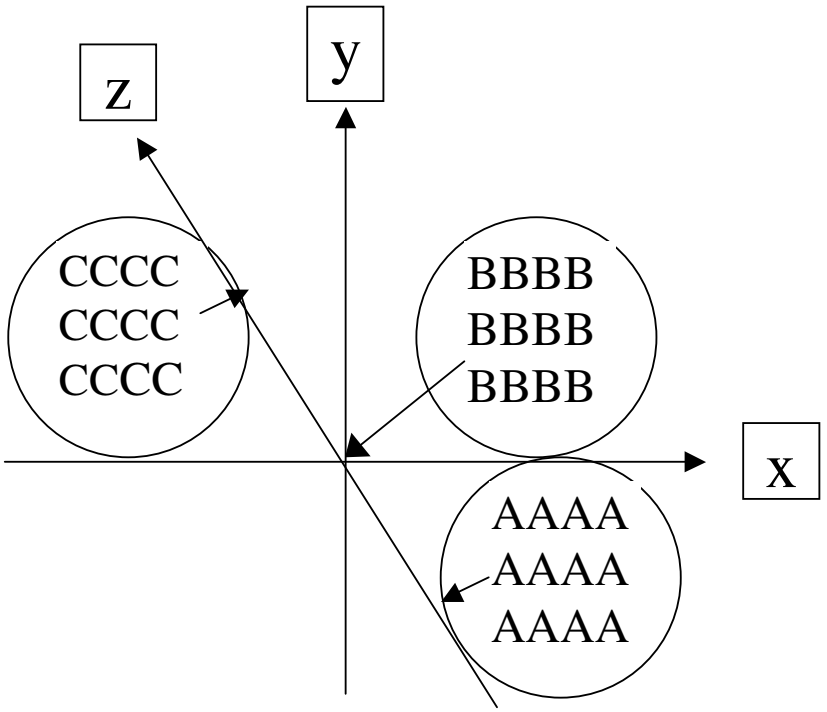


Illustration de la problématique descriptive

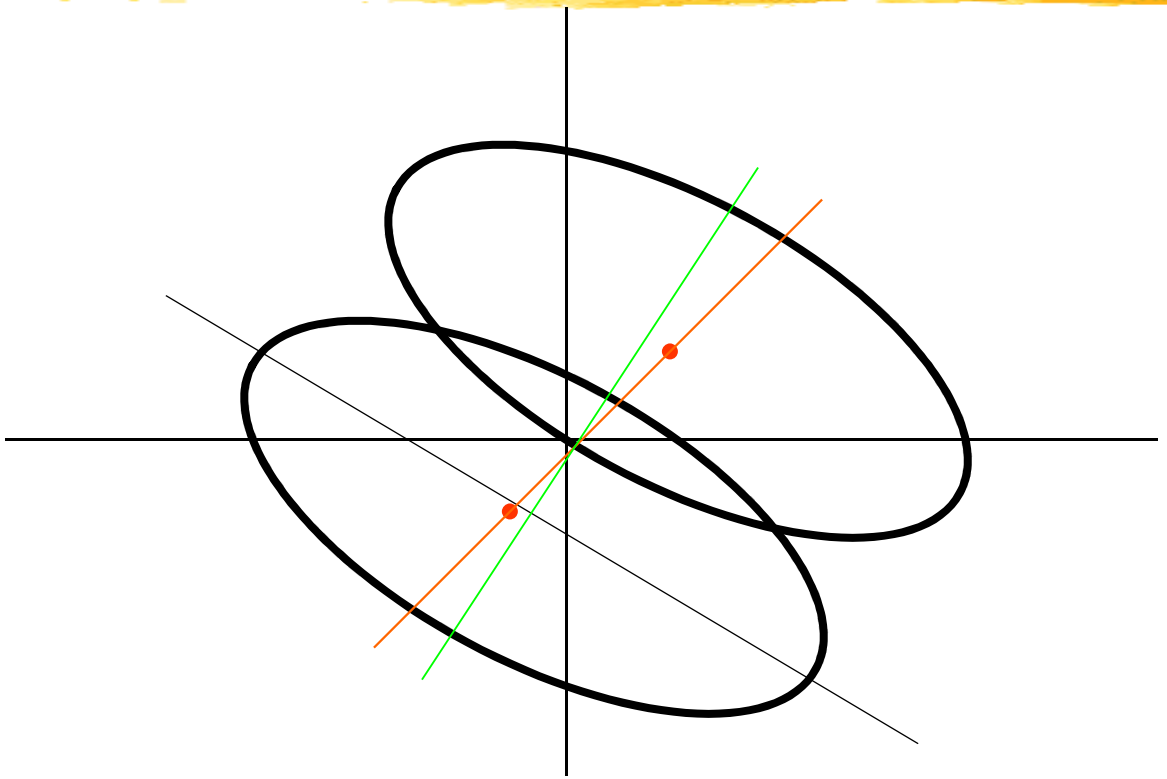
- Dans l'exemple suivant :
  - l'axe « x » différencie bien les groupes « B » et « C » mais non les groupes « A » et « B »
  - l'axe « y » différencie bien les groupes « A » et « B » mais non les groupes « B » et « C »
  - *en revanche l'axe « z » différencie bien les trois groupes.*
- La droite :
  - z = + 1 sépare les « B » et « C »
  - z = - 1 sépare les « A » et « B »
- z est une fonction de score



Double objectif de l'analyse factorielle discriminante

- Les n individus forment un nuage de n points dans R^p, formé des k sous-nuages G_i à différencier
- Variance interclasse (« between ») = variance des barycentres g_i (« centroïdes ») des classes G_i
  - $B = 1/n \Sigma n_i(g_i - g)(g_i - g)'$ = matrice de covariance « between »
- Variance intraclasse (« within ») = moyenne des variances des classes G_i
  - $W = 1/n \Sigma n_i V_i$ = matrice de covariance « within »
- Théorème de Huygens : B + W = variance totale V
- Impossible de trouver un axe u qui simultanément :
  - maximise la variance interclasse sur u : max u'Bu
  - minimise la variance intraclasse sur u : min u'Wu

Visualisation du double objectif



Maximum de dispersion interclasse : u parallèle au segment joignant les centroïdes

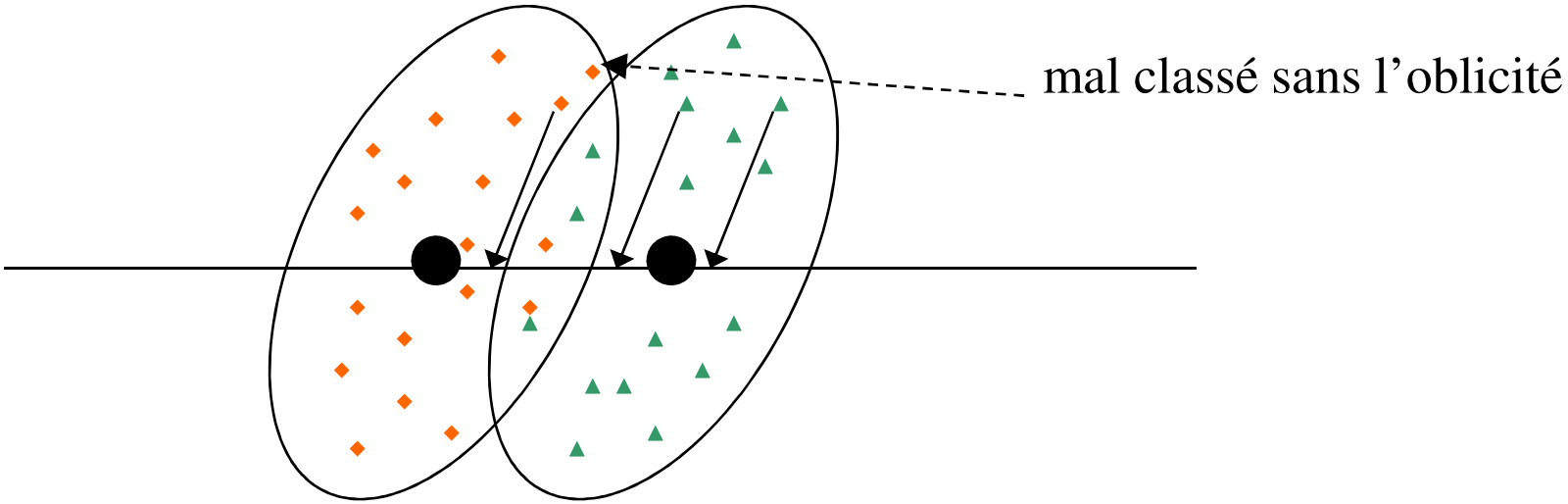
Minimum de dispersion intraclasse : u perpendiculaire à l'axe principal des ellipses (on suppose l'homoscédasticité)

Compromis entre les 2 objectifs

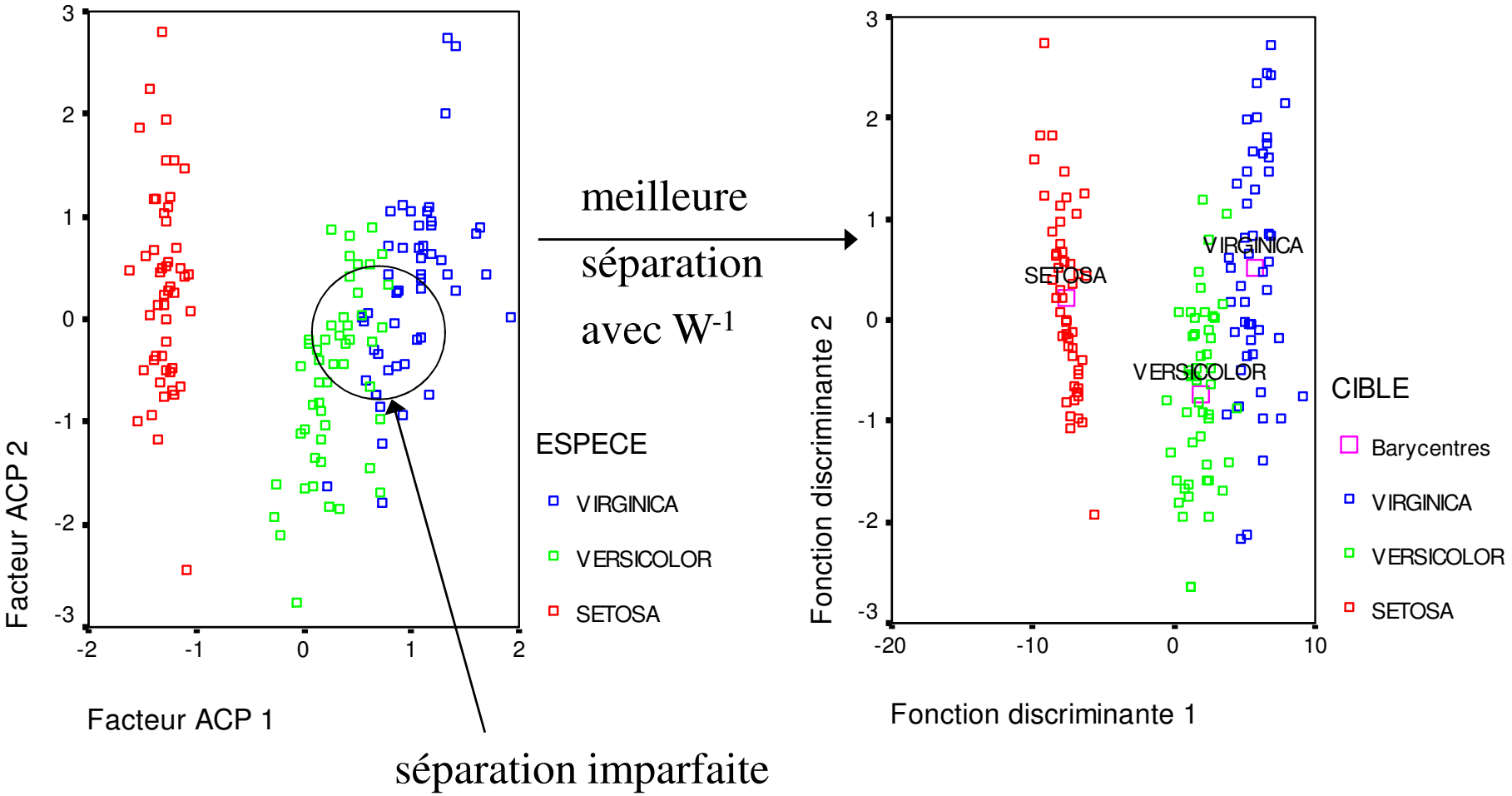
- On reformule l'objectif : au lieu de maximiser u'Bu ou minimiser u'Wu, on maximise u'Bu/u'Wu
- ⇔ maximiser u'Bu/u'Vu (Huygens)
- On montre que :
  - la solution u est le vecteur propre de $V^{-1}B$ associé à λ la plus grande valeur propre de $V^{-1}B$
  - u vecteur propre de $V^{-1}B$ ⇔ u vecteur propre de $W^{-1}B$, de valeur propre $\lambda/1-\lambda$
- On dit que les métriques V^{-1} et W^{-1} sont équivalentes
  - la métrique W^{-1} (de Mahalanobis) est plus utilisée par les Anglo-saxons et les éditeurs de logiciels
- Distance d de 2 points x et y : $d^2(x,y) = (x-y)' W^{-1}(x-y)$

Autre formulation de la solution

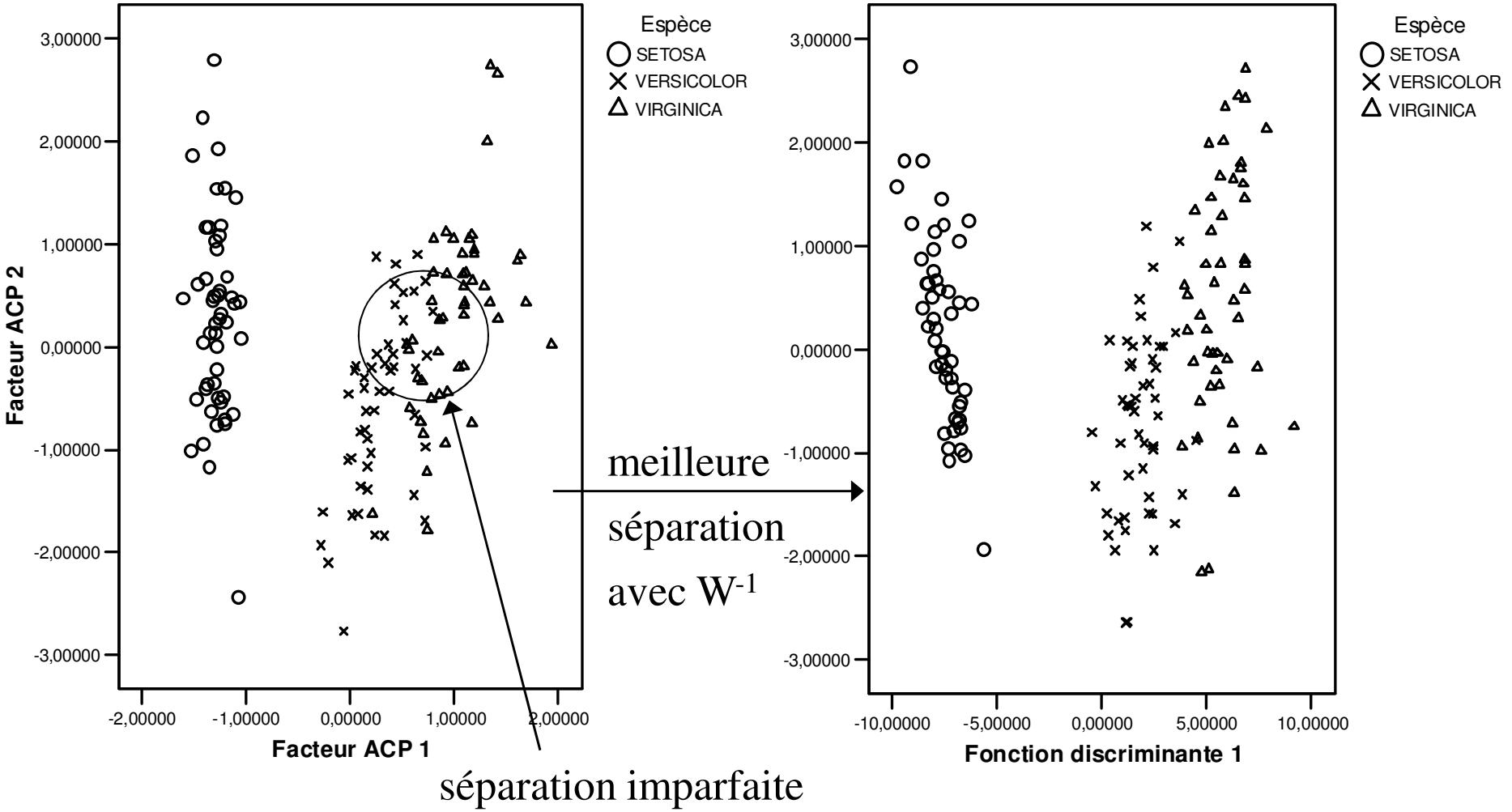
- ACP du nuage des centroïdes g_i avec :
  - métrique V^{-1}
  - ou métrique W^{-1} équivalente
- Ces métriques correspondent à une projection oblique
- Sans cette oblicité, il s'agirait d'une simple ACP mais les groupes seraient mal séparés



ACP avec métrique usuelle et avec W^{-1}



ACP avec métrique usuelle et avec W^{-1}



Analyse discriminante prédictive et fonctions de Fisher

- On classe x dans le groupe G_i pour lequel la distance au centre g_i est minimale :
- $d^2(x,g_i) = (x-g_i)'W^{-1}(x-g_i) = x' W^{-1}x - 2g_i' W^{-1}x + g_i' W^{-1}g_i$

- Minimiser $d^2(x,g_i) \Leftrightarrow$ maximiser $(2g_i' W^{-1}x - g_i' W^{-1}g_i)$
- $g_i' W^{-1}g_i = \alpha_i$ est une constante ne dépendant pas de x

- Pour chacun des k groupes G_i, on a une fonction discriminante de Fisher :
  - $\alpha_i + \beta_{i,1}X_1 + \beta_{i,2}X_2 + \ldots \beta_{i,p}X_p$
- et on classe x dans le groupe pour lequel la fonction est maximale

Exemple des iris de Fisher

Coefficients des fonctions de classement

	CIBLE		
	SETOSA	VERSICOLOR	VIRGINICA
SE_L	2,35442	1,56982	1,24458
SE_W	2,35879	,70725	,36853
PE_L	-1,64306	,52115	1,27665
PE_W	-1,73984	,64342	2,10791
(Constante)	-86,30847	-72,85261	-104,36832

Fonctions discriminantes linéaires de Fisher

			Classe d'affectation pour analyse 1			
			SETOSA	VERSICOLOR	VIRGINICA	Total
CIBLE	SETOSA	Effectif	50	0	0	50
	VERSICOLOR	Effectif	0	48	2	50
	VIRGINICA	Effectif	0	1	49	50
Total		Effectif	50	49	51	150

Règle d'affectation dans le cas de 2 groupes

- L'aspect descriptif est simple (l'axe discriminant joint les 2 centroïdes) et on s'intéresse à l'aspect prédictif

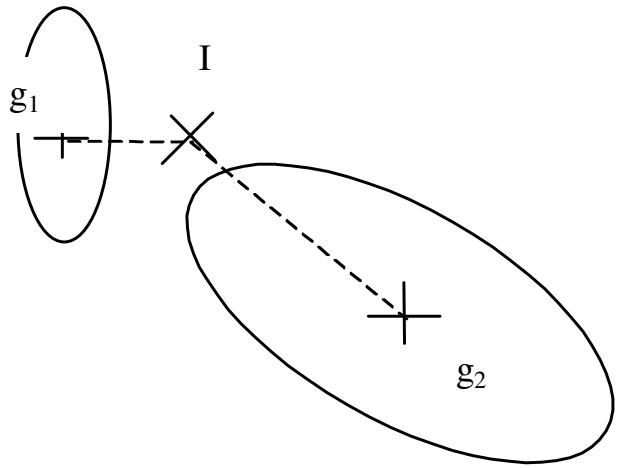
- On classe x dans le groupe G_1 si :
- $2g_1' W^{-1}x - g_1' W^{-1}g_1 > 2g_2' W^{-1}x - g_2' W^{-1}g_2$
- $\Leftrightarrow \underbrace{(g_1-g_2)' W^{-1}x - \frac{1}{2}(g_1' W^{-1}g_1 - g_2' W^{-1}g_2)}_{f(x)} > 0$

- f(x) : fonction de score de Fisher

- D^2 de Mahalanobis : $d^2(g_1,g_2) = (g_1-g_2)' W^{-1}(g_1-g_2)$
- $W^{-1}(g_1-g_2)$ = axe discriminant proportionnel à $V^{-1}(g_1-g_2)$

Limite de la règle géométrique d'affectation

- Règle géométrique : affecter chaque individu au groupe dont il est le + proche (distance de l'individu au centroïde du groupe)
  - ce n'est pas trivial car il faut prendre la métrique W^{-1} (faire une projection oblique de x sur l'axe discriminant)
- A éviter si les 2 groupes ont des probabilités *a priori* ou des variances différentes



- Dans ce cas : analyse discriminante quadratique (voir plus loin)

Technique de classement :

Analyse discriminante probabiliste

L'approche probabiliste (bayésienne)

- Pour tout i ≤ k, soient :
  - $P(G_i/x)$ = proba *a posteriori* d'appartenance à G_i sachant x (connaissant les caractéristiques de x, son « dossier »)
  - $p_i = P(G_i)$ = proba *a priori* d'appartenance à G_i (proportion de G_i dans la population)
  - $f_i(x) = P(x/G_i)$ = densité conditionnelle de la loi de x connaissant son groupe G_i
- D'après le théorème de Bayes :

$$P(G_i / x) = \frac{P(G_i)P(x / G_i)}{\sum_j P(G_j)P(x / G_j)}$$

- Règle de classement bayésienne :
  - on classe x dans le groupe G_i où $P(G_i/x)$ est maximum

3 possibilités pour estimer $P(G_i/x)$

- En commençant par calculer $P(x/G_i)$ selon une méthode paramétrique (on suppose la multinormalité de $P(x/G_i)$ avec éventuellement égalité des Σ_i, donc le nb de paramètres du problème est fini : ADL ou ADQ)
- En commençant par estimer $P(x/G_i)$ selon une méthode non paramétrique (pas d'hypothèse sur la densité $P(x/G_i)$: méthode du noyau ou des plus proches voisins)
- Directement par une approche semi-paramétrique (régression logistique) où on écrit $P(G_i/x)$ sous la forme :

$$P(G_i / x) = \frac{e^{\alpha' x + \beta}}{1 + e^{\alpha' x + \beta}}$$

1e possibilité : Hypothèse de multinormalité

- La densité d'une loi multinormale $N(\mu_i,\Sigma_i)$ est :

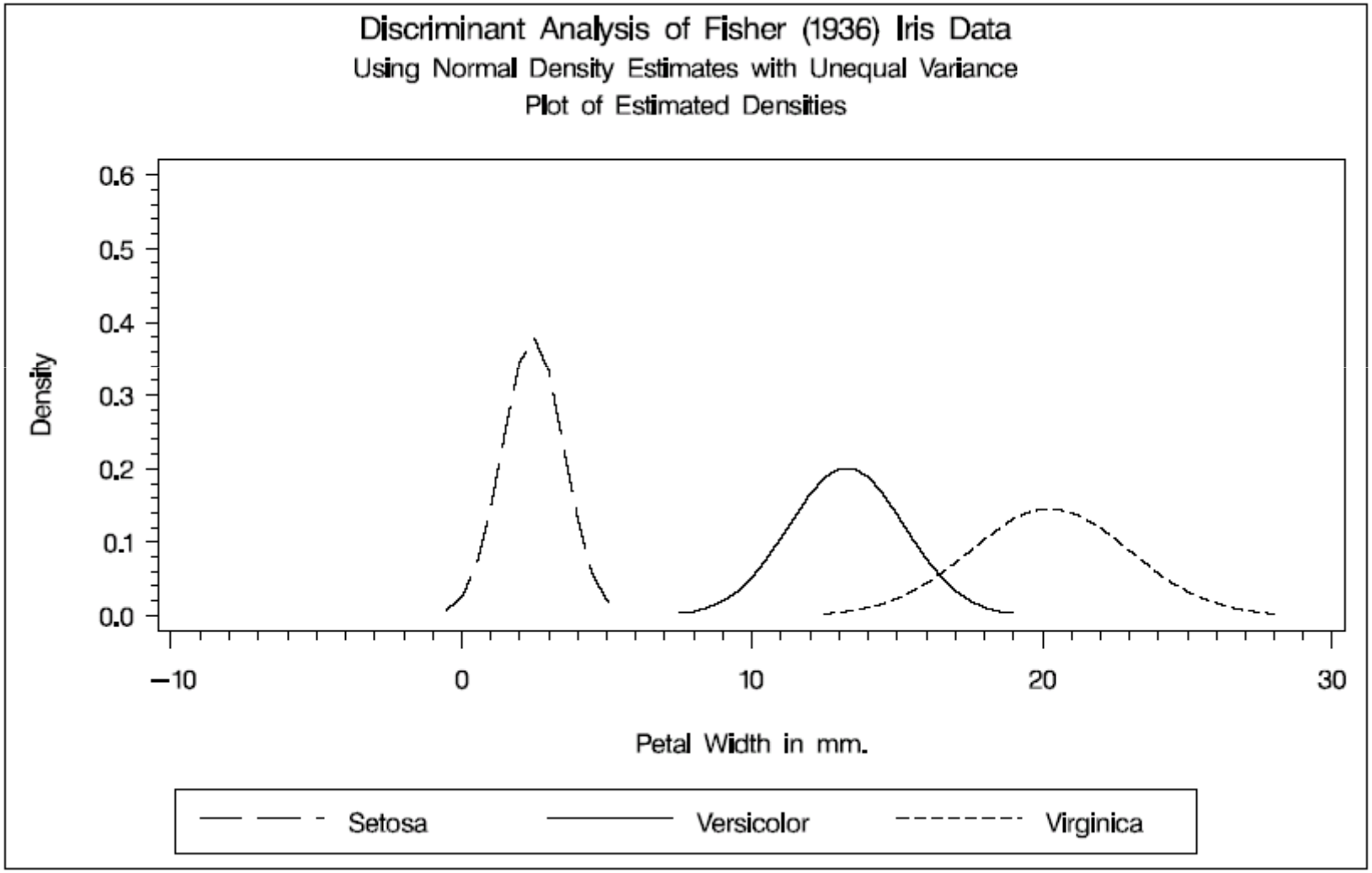
$$f_i(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma_i)}} \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right]$$

- D'après Bayes, maximiser $P(G_i/x)$ ⇔ maximiser $p_i f_i(x)$:

$$\underset{i}{\text{maximum}}\left[Log(p_i) - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log(\det(\Sigma_i))\right]$$

> On obtient une règle quadratique en x

Multinormalité



Hypothèse d'homoscédasticité

- Sous cette hypothèse, on a : $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k = \Sigma$
- On classe x dans le groupe G_i pour avoir :

$$\underset{i}{\text{maximum}} \left[Log(p_i) - \frac{1}{2} x'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + x'\Sigma^{-1}\mu_i \right]$$

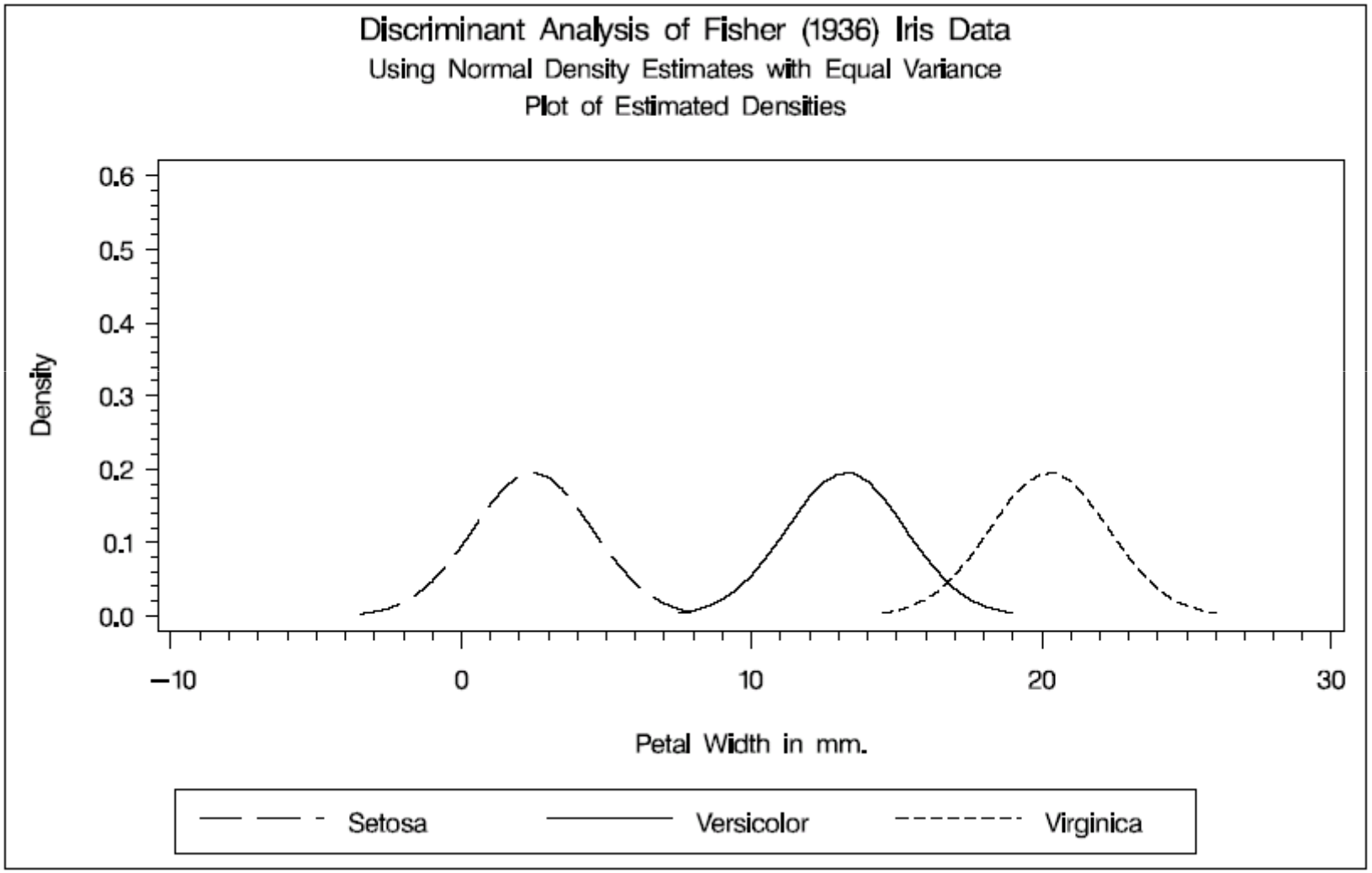
- Soit, puisque $x'\Sigma^{-1}x$ est indépendant de i :

$$\underset{i}{\text{maximum}} \left[Log(p_i) - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + x'\Sigma^{-1}\mu_i \right]$$

Les probabilités *a priori* ne changent qu'une constante additive

- Homoscédasticité (+ multinormalité) => on passe d'une fonction quadratique à une fonction linéaire
- Avec en + l'équiprobabilité => on a équivalence des règles géométrique (maximiser la fct de Fisher) et bayésienne

Homoscédasticité



Cas de 2 groupes
(hypothèses de multinormalité et homoscédasticité)

- Probabilité d'appartenance au groupe 1 :

$$P(G_1 / x) = \frac{p_1 \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right]}{p_1 \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right] + p_2 \exp\left[-\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right]}$$

$$\frac{1}{P(G_1 / x)} = 1 + \frac{p_2}{p_1} \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right]$$

- On peut écrire $1/P(G_1/x) = 1 + (p_2/p_1)e^{-f(x)}$
  - avec $f(x) = \frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)$
- On classe x dans G_1 si $P(G_1/x) > 0,5$
- $\Leftrightarrow (p_2/p_1)e^{-f(x)} < 1 \Leftrightarrow f(x) > \log(p_2/p_1)$

Cas de 2 groupes (suite)

- Développons la fonction f(x) :
  - $f(x) = (\mu_1 - \mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2)$
- On reconnaît la fonction de score de Fisher

> La règle bayésienne précédente équivaut à la règle :

  - fonction de Fisher > $\log(p_2/p_1)$

- qui généralise la règle géométrique f(x) > 0 lorsque les probabilités *a priori* p_1 et p_2 sont différentes
- De plus, la probabilité *a posteriori* $P(G_1/x)$ s'écrit :

$$P(G_1 / x) = \frac{1}{1 + \left(\frac{p_2}{p_1}\right) e^{-f(x)}} = \frac{e^{f(x)}}{\left(\frac{p_2}{p_1}\right) + e^{f(x)}}$$

- Généralisation de la fonction logistique !

En résumé :

- Avec l'hypothèse de multinormalité :
  - La règle bayésienne est quadratique
- Avec les hypothèses de multinormalité et d'homoscédasticité :
  - La règle bayésienne est linéaire
  - Dans le cas de 2 groupes, elle s'écrit $f(x) > \log(p_2/p_1)$, où $f(x)$ est la fonction de Fisher obtenue par un raisonnement géométrique
- Avec les hypothèses de multinormalité, d'homoscédasticité et d'équiprobabilité :
  - La règle bayésienne est linéaire et équivalente à la règle géométrique
  - Dans le cas de 2 groupes, elle s'écrit $f(x) > 0$ et la probabilité *a posteriori* $P(G_1/x)$ s'écrit sous la forme logistique $P(G_1/x) = 1 / (1 + e^{-f(x)})$

Coûts de mauvais classement

- On peut introduire des coûts d'erreurs
  - C(i/j) = coût de classement dans G_i au lieu de G_j
  - C(i/i) = 0

- Coût moyen de classement en $G_i = \Sigma_j C(i/j) P(G_j/x)$
- On classe x dans le G_i qui minimise le coût

- Cas de 2 groupes :
  - Coût moyen d'un classement en G_1 : $C(1/2) P(G_2/x)$
  - Coût moyen d'un classement en G_2 : $C(2/1) P(G_1/x)$
  - On classe x en G_1 si $C(1/2) P(G_2/x) < C(2/1) P(G_1/x)$

An. Discriminante non paramétrique

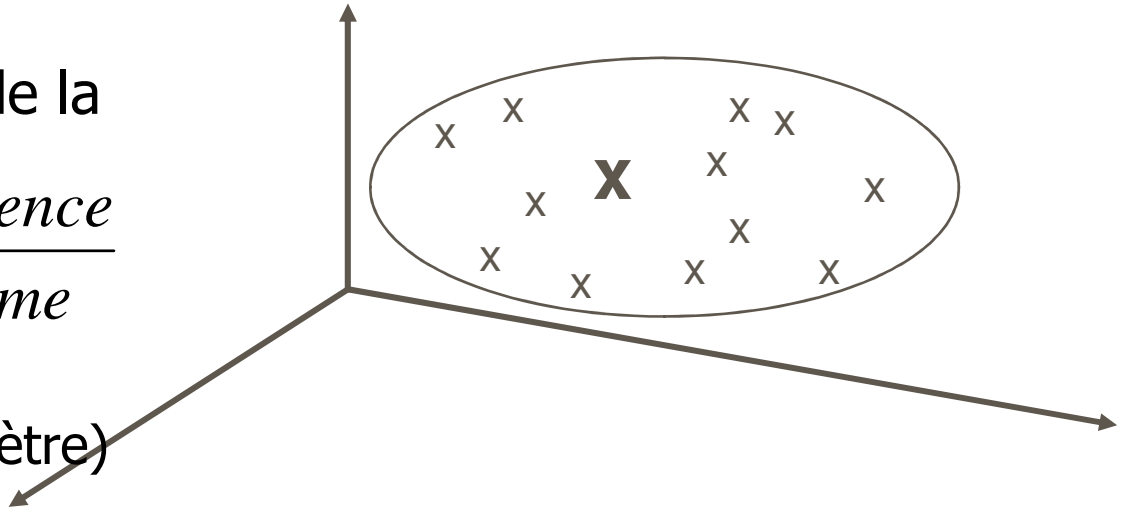
- D'après Bayes :

$$P(G_i / x) = \frac{p_i f_i(x)}{\sum_j p_j f_j(x)}$$

- Problème d'estimation de la densité :

$$f_i(x) = \frac{fréquence}{volume}$$

- Méthodes :
  - noyau (on fixe le diamètre)
  - k-plus proches voisins (on fixe le nb de voisins)
- Condition : avoir un échantillon de grande taille



D^2 de Mahalanobis

- Définition : $D^2 = d^2(g_1,g_2) = (g_1 - g_2)' W^{-1}(g_1 - g_2)$
- Le carré D^2 de la distance de Mahalanobis fournit une mesure de la distance entre les deux groupes à discriminer, et donc de la qualité de la discrimination
- Analogue au R^2 d'une régression
- Plus D^2 est grand, mieux c'est
- On peut faire un test de Fisher sur l'hypothèse nulle que tous les centroïdes sont égaux
- Il peut servir de critère dans une régression pas à pas

R^2

- Corrélation canonique = coefficient de corrélation entre la fonction de score et la moyenne par classe (pour chaque individu : on prend la moyenne de la fonction discriminante dans sa classe)

- Carré de la corrélation canonique R = coefficient de détermination R^2 = proportion de la variance de la fonction discriminante expliquée par l'appartenance à l'une ou l'autre classe à discriminer

- Autrement dit R^2 = variance interclasse / variance totale
  - Le but de l'analyse discriminante est de maximiser ce rapport

Lambda de Wilks

- Lambda de Wilks = variance intraclasse / variance totale
  - varie entre 0 et 1 (var. totale = var. intra + var. inter)
  - $\lambda = 1$ => tous les centroïdes sont égaux
- Plus λ est bas, mieux c'est
- Test de Fisher sur le lambda de Wilks <=> Test de l'hypothèse nulle que tous les centroïdes sont égaux
- Il peut servir de critère dans une régression pas à pas

	Lambda de Wilks	F	ddl1	ddl2	Signification
SE_L	,381	119,265	2	147	,000
SE_W	,599	49,160	2	147	,000
PE_L	,059	1180,161	2	147	,000
PE_W	,071	960,007	2	147	,000

Les groupes diffèrent beaucoup sur la longueur des pétales

Matrice de confusion
Validation croisée

Matrice de confusion[b,c]

		CIBLE	Classe(s) d'affectation prévue(s) SETOSA	VERSICOLOR	VIRGINICA	Total
Original	Effectif	SETOSA	50	0	0	50
		VERSICOLOR	0	48	2	50
		VIRGINICA	0	1	49	50
	%	SETOSA	100,0	,0	,0	100,0
		VERSICOLOR	,0	96,0	4,0	100,0
		VIRGINICA	,0	2,0	98,0	100,0
Validé-croisé[a]	Effectif	SETOSA	50	0	0	50
		VERSICOLOR	0	48	2	50
		VIRGINICA	0	1	49	50
	%	SETOSA	100,0	,0	,0	100,0
		VERSICOLOR	,0	96,0	4,0	100,0
		VIRGINICA	,0	2,0	98,0	100,0

a. Dans la validation croisée, chaque observation est classée par les fonctions dérivées de toutes les autres observations.

b. 98,0% des observations originales classées correctement.

c. 98,0% des observations validées-croisées classées correctement.

Résumé des critères statistiques

- D^2 de Mahalanobis : test de Fisher
- Lambda de Wilks = 1 - R^2 : test de Fisher

Pas	Nombre de variables	Lambda	F exact			
			Statistique	ddl1	ddl2	Signification
1	1	,059	1180,161	2	147,000	,000
2	2	,037	307,105	4	292,000	,000
3	3	,025	257,503	6	290,000	,000
4	4	,023	199,145	8	288,000	,000

- Matrice de confusion : test Q de Press
- Coefficients discriminants standardisés (sur var. centrées réduites)
  - pour comparer l'importance des variables explicatives

Syntaxe SAS de l'analyse discriminante

```
ods rtf file="c:\fisher_sas.doc";
proc stepdisc data=matable.ascorer;
class cible;
var var1 var2 … vari; run;
proc discrim data=matable.ascorer method=normal pool=yes
  crossvalidate all canonical out=matable.scoree
  outstat=matable.destat;
class cible;
priors proportional;
var var1 var2 … vari; run;
proc discrim data=matable.destat testdata=matable.test
   testout=tout;
class cible;
var var1 var2 … vari; run;
ods rtf close ;
```

Fichier en sortie OUTSTAT

Obs	cible	_TYPE_	_NAME_	nbproduits	abonnement1	nbenfants	abonnement2	evolconsom	nbsorties
1	.	N		6385.00	6385.00	6385.00	6385.00	6385.00	6385.00
2	0	N		5306.00	5306.00	5306.00	5306.00	5306.00	5306.00
3	1	N		1079.00	1079.00	1079.00	1079.00	1079.00	1079.00
4	.	MEAN		8.94	371.28	1.34	23.11	1.16	6.48
5	0	MEAN		8.47	281.68	1.38	19.62	1.14	5.96
6	1	MEAN		11.23	811.86	1.15	40.28	1.25	9.05
…	…	…	…	…	…	…	…	…	…
119	0	LINEAR	_LINEAR_	0.38	-0.00	1.12	-0.00	8.42	0.05
120	0	LINEAR	_CONST_	-7.50	-7.50	-7.50	-7.50	-7.50	-7.50
121	1	LINEAR	_LINEAR_	0.48	0.00	0.83	0.01	9.14	0.09
122	1	LINEAR	_CONST_	-11.27	-11.27	-11.27	-11.27	-11.27	-11.27

Avantages de l'analyse discriminante

- Problème à solution analytique directe (inverser W)
- Optimale quand les hypothèses de non colinéarité, homoscédasticité et multinormalité sont vérifiées
- Les coefficients des combinaisons linéaires constituent un résultat relativement explicite
- Modélise très bien les phénomènes linéaires
- Aptitude à détecter les phénomènes globaux
- Ne nécessite pas un gros ensemble d'apprentissage
- Rapidité de calcul du modèle
- Possibilité de sélection pas à pas
- Facilité d'intégrer des coûts d'erreur de classement
- Technique implémentée dans de nombreux logiciels

Inconvénients de l'analyse discriminante

- Ne détecte que les phénomènes linéaires
- Ne s'applique pas à tout type de données (données numériques sans valeurs manquantes)
  - mais possibilité d'utiliser une ACM (méthode DISQUAL)
- Hypothèses contraignantes, et pour s'en rapprocher :
  - normaliser les variables
  - sélectionner soigneusement les variables les + discriminantes
  - éliminer les variables colinéaires
  - éliminer les individus hors norme
  - s'il reste de l'hétéroscédasticité, mieux vaut avoir des classes de tailles comparables
  - travailler sur des populations homogènes
    - il vaut donc mieux préalablement segmenter

Technique de classement :
La régression logistique

La régression logistique binaire

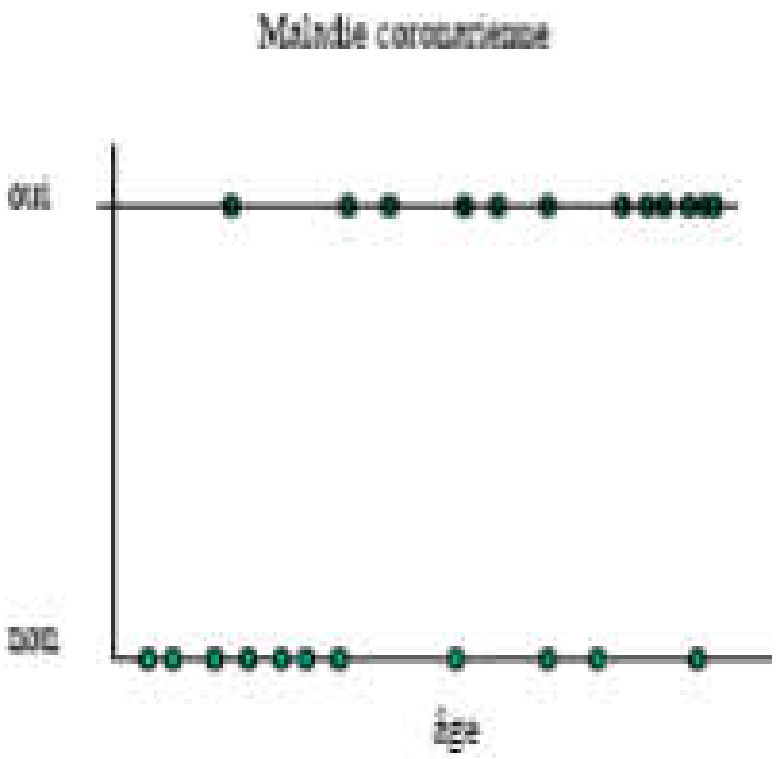
- Y variable cible binaire Y = 0 / 1
- X_j p variables explicatives continues, binaires ou qualitatives
  - p = 1 régression logistique simple
  - p > 1 régression logistique multiple
- Généralisation : régression logistique polytomique
  - la variable cible Y est qualitative à k modalités
  - cas particulier : Y ordinale (régression logistique ordinale)
- Pb de régression : modéliser l'espérance conditionnelle

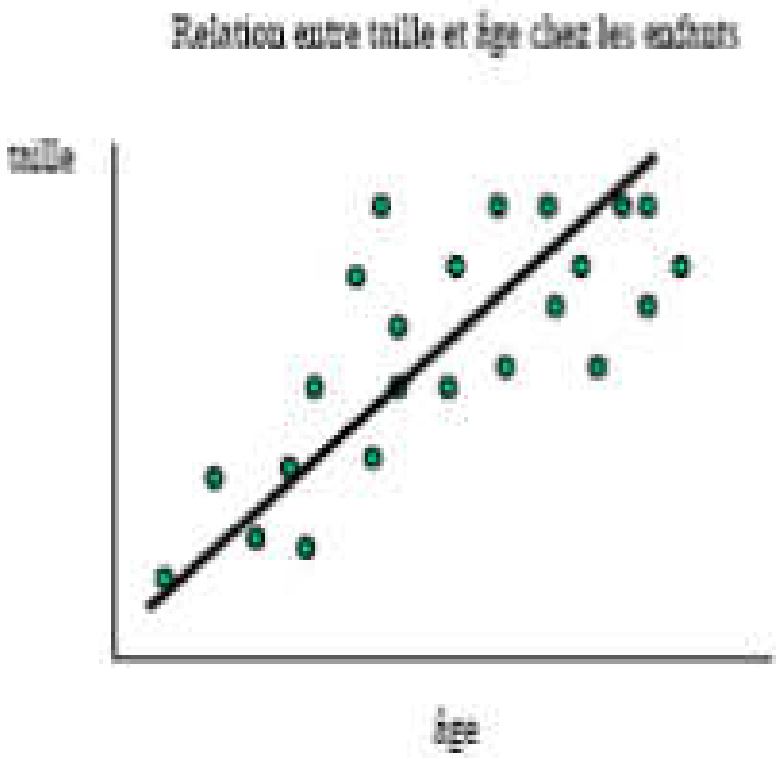
$$E(Y/X=x) = Prob(Y=1/X=x)$$

sous la forme $E(Y/X=x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

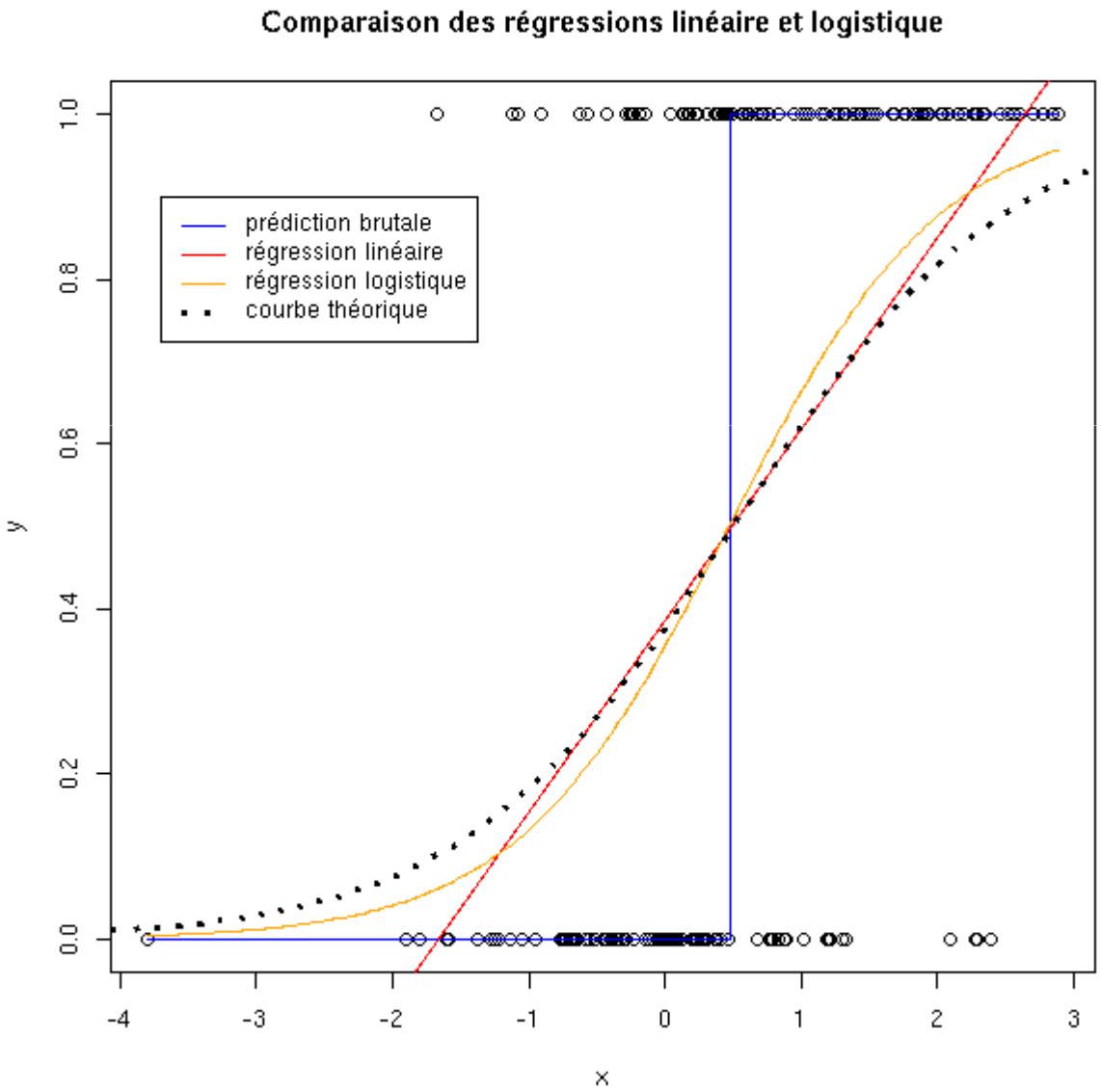
- Difficulté ! X_i continues => terme de droite non borné alors que $Prob(Y=1/X=x) \in [0,1]$ => il faut le transformer !
  - en régression linéaire : E(Y/X=x) n'est pas bornée

Variable à expliquer : discrète ou continue





Prédiction d'une variable binaire



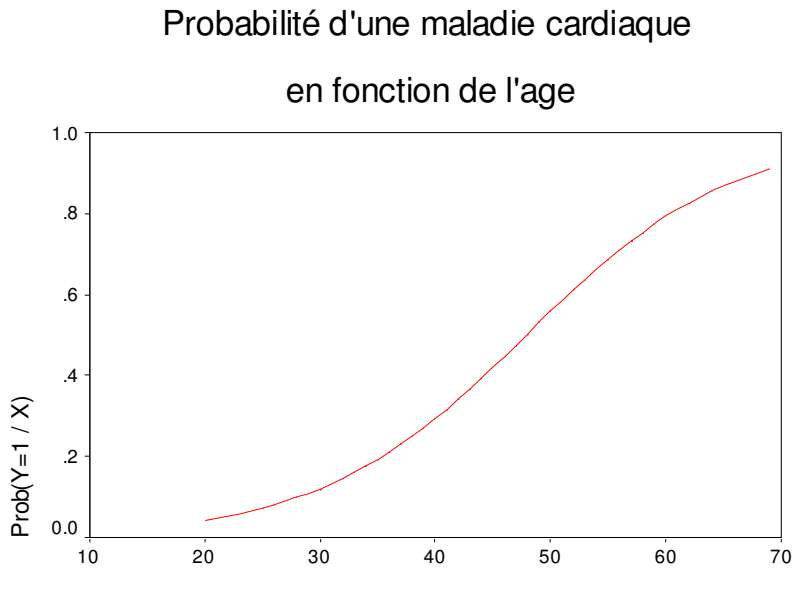
Cas d'une variable x multinormale : $x \approx N(0,1)$ sur l'ensemble des Y=0 et $x \approx N(1,1)$ sur l'ensemble des Y=1.

La courbe théorique E(Y/X=x) est donnée par $f_{N(1,1)}(x)/(f_{N(1,1)}(x)+f_{N(0,1)}(x))$

où $f_{N(\mu,\sigma)}$ est la fonction de densité de la loi $N(\mu,\sigma)$.

La régression logistique binaire

- Visiblement la régression linéaire ne convient pas (distribution des résidus !)
- La figure fait pressentir que ce n'est pas une fonction linéaire de $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ qu'il faut appliquer, mais une courbe en S
- Les courbes en S sont courantes en biologie et en épidémiologie



Age and Coronary Heart Disease (CHD)

(source : Hosmer & Lemeshow - chapitre 1)

CHD = maladie coronarienne (rétrécissement des artères du muscle cardiaque)

ID	AGRP	AGE	CHD
1	1	20	0
2	1	23	0
3	1	24	0
4	1	25	0
5	1	25	1
⋮	⋮	⋮	⋮
97	8	64	0
98	8	64	1
99	8	65	1
100	8	69	1

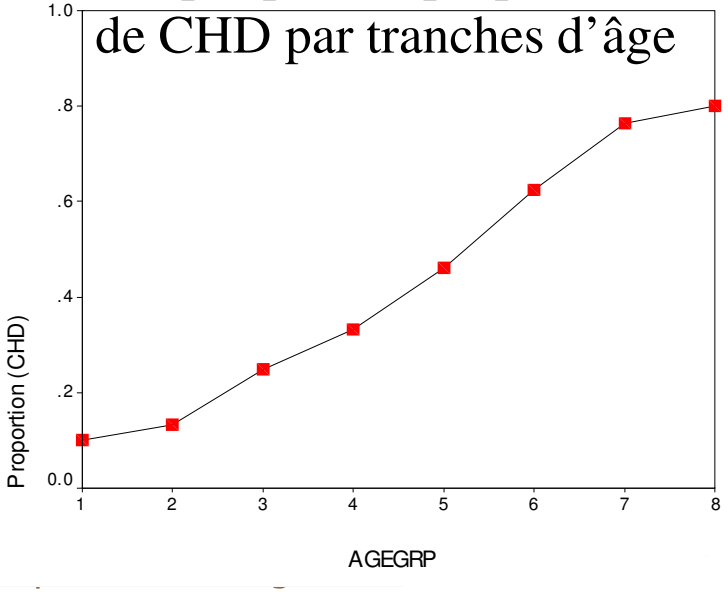
La régression logistique binaire

- Ici, difficile de calculer $\pi(x) := Prob(Y=1/X=x)$ car trop peu de valeurs de Y pour une valeur x donnée
- On regroupe les valeurs de X par tranches :
  - proportion des Y = 1 sachant x : meilleur estimateur de la probabilité que Y = 1 sachant x
  - procédure de regroupement en classes : classique en scoring !

Tableau des effectifs
de CHD par tranches d'âge

Age Group	n	CHD absent	CHD present	Mean (Proportion)
20 – 29	10	9	1	0.10
30 – 34	15	13	2	0.13
35 – 39	12	9	3	0.25
40 – 44	15	10	5	0.33
45 – 49	13	7	6	0.46
50 – 54	8	3	5	0.63
55 - 59	17	4	13	0.76
60 - 69	10	2	8	0.80
Total	100	57	43	0.43

Graphique des proportions
de CHD par tranches d'âge



Fonction de lien

- On écrit donc $\pi(x)$ = Prob(Y=1/X=x) sous la forme :

$$\pi(x) = \frac{e^{\beta_0 + \sum_j \beta_j x_j}}{1 + e^{\beta_0 + \sum_j \beta_j x_j}}$$

- $\Leftrightarrow$ $$Log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

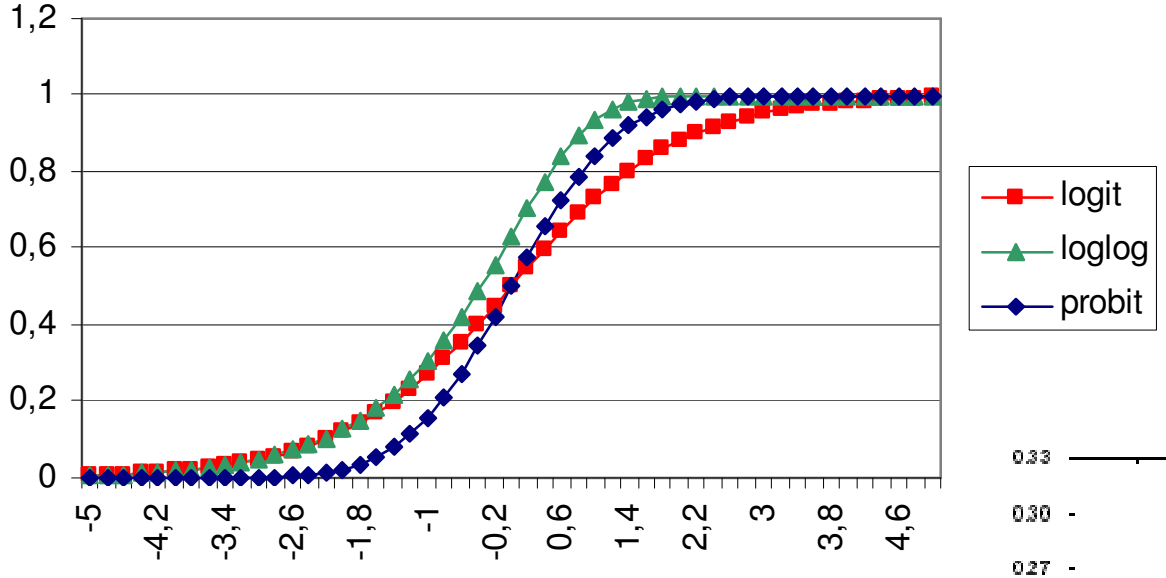
Fonction de lien : Logit($\pi(x)$)

- Cohérent avec la règle bayésienne de l'analyse discriminante et le calcul de la probabilité *a posteriori* dans le cas gaussien homoscédastique

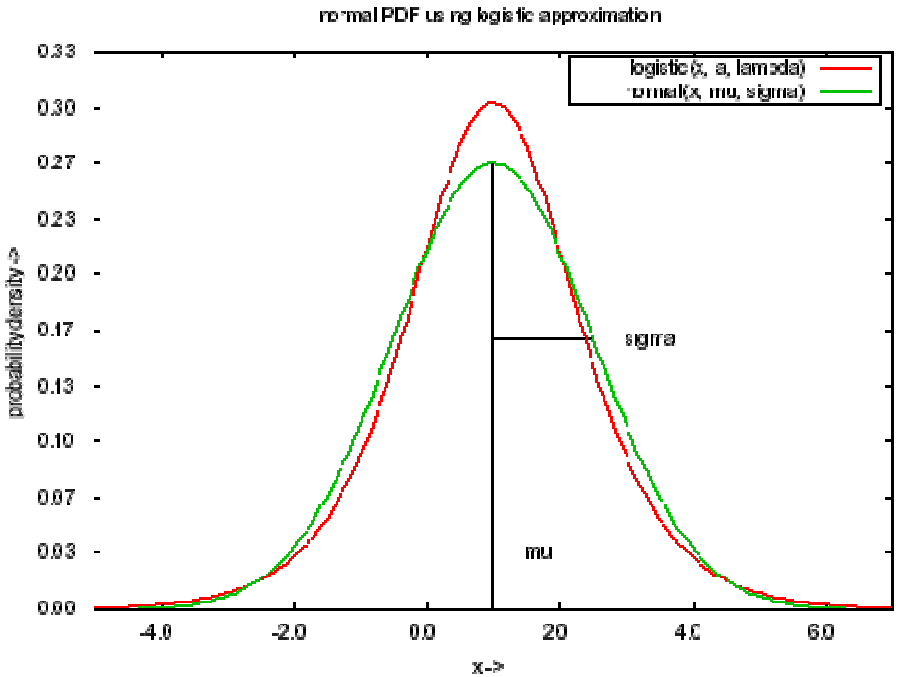
Les différentes fonctions de lien

Modèle	Fonction de lien	Fonction de transfert
Logit	Log (μ/ [1 – μ])	$\frac{\exp(t)}{1+\exp(t)} = \int_{-\infty}^{t} \frac{\exp(z)}{(1+\exp(z))^2} dz$
Probit (normit)	fonction inverse de la fonction de répartition d'une loi normale centrée réduite	$s(t) = \int_{-\infty}^{t} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz$
Log-log	Log [– Log(1–μ)]	1 – exp[– exp(t)]

Similarité des fonctions de transfert



$$\text{coeff (logit)} \approx \frac{\pi}{\sqrt{3}} \text{ coeff (probit)}$$



Logit : odds-ratio d'un régresseur X_i

- Mesure l'évolution du rapport des probas d'apparition de l'événement Y=1 contre Y=0 (odds = « cote » des parieurs) lorsque X_i passe de x à x+1. Dans ce cas, logit(π(x)) augmente du coefficient β_i de $X_i \Rightarrow$ la cote $\pi(x)/[1 - \pi(x)]$ est multipliée par exp(β_i)
- Formule générale :

$$OR = \frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = e^{\beta_i}$$

- Si X_i est binaire 0/1, la formule devient :

$$OR = \frac{P(Y=1/X_i=1)/P(Y=0/X_i=1)}{P(Y=1/X_i=0)/P(Y=0/X_i=0)} = e^{\beta_i}$$

Interprétation du odds-ratio OR

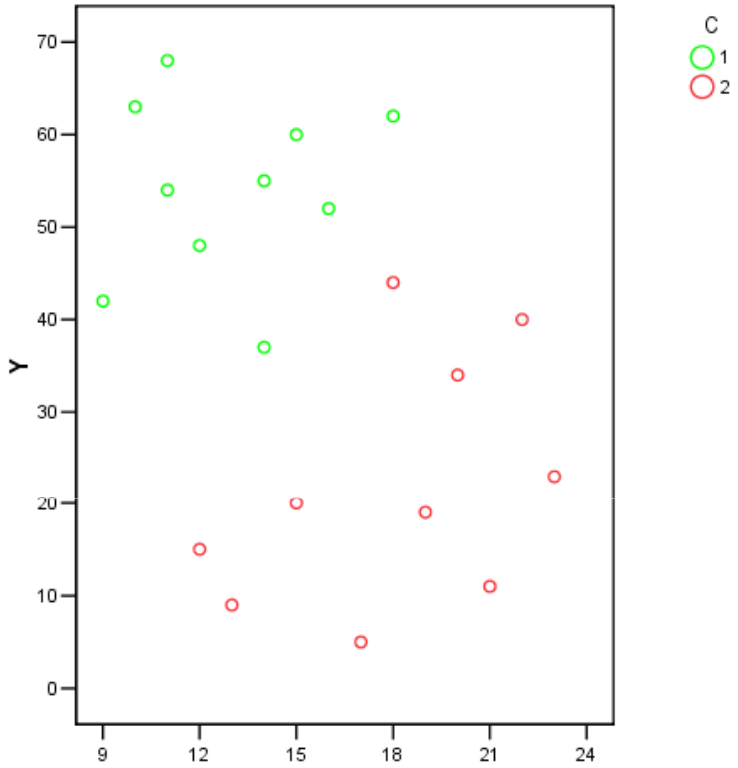
- Attention : odds-ratio ≠ du risque relatif $\pi(x+1)/\pi(x)$
  - sauf si $\pi(x)$ est petit (détection de phénomène rare)
- Un seul OR pour X binaire
  - ex : comparer les hommes (x=1) et les femmes (x=0)
- Un seul OR est plus douteux pour X continue
  - ex : comparer l'âge 61 et 60, 60 et 59… avec le même OR ? Risque de manque de robustesse par manque de données (voir CHD ci-dessus). Non détection de la non-linéarité.
- OR à n'utiliser sur des variables qualitatives qu'après dichotomisation (nb indicatrices = nb modalités - 1, en prenant une modalité comme référence)
  - ex : comparer « petites villes » et « campagne » avec un OR1 et comparer « grandes villes » et « campagne » avec un OR2, car aucune raison d'avoir OR1 = OR2
  - indicatrices crées automatiquement par certains logiciels

Odds-ratio d'une variable qualitative

- Exemple : comparaison de la probabilité $\pi(x)$ d'apparition d'un événement dans les grandes villes, les petites villes et à la campagne
  - quand on passe de la modalité de référence (« campagne ») à la modalité « petite ville », la cote $\pi(x)/[1 - \pi(x)]$ est multipliée par l'exponentielle 0,573 de la différence des coefficients B associés à la modalité « petite ville » (B = -0,558) et à la modalité de référence (B = 0)
  - autrement dit, la cote $\pi(x)/[1 - \pi(x)]$ de l'événement (différent de sa probabilité $\pi(x)$!) est presque 2 fois plus faible dans une petite ville qu'à la campagne

	B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
							Inférieur	Supérieur
campagne			36,671	2	0,000			
petite ville	-0,55767728	0,136	16,784	1	0,000	0,573	0,438	0,748
grande ville	0,28802599	0,143	4,057	1	0,044	1,334	1,008	1,765
Constante	-1,25610388	0,236	28,363	1	0,000	0,285		

Séparation complète des groupes



Historique des itérations[a,b,c,d]

Itération		-2log-vrais emblance	Coefficients		
			Constante	X	Y
Etape 1	1	9,271	-,132	,182	-,071
	2	5,000	-,750	,344	-,119
	3	2,974	-2,082	,563	-,172
	4	1,747	-4,940	,908	-,237
	5	,816	-10,239	1,505	-,339
	6	,319	-16,448	2,252	-,478
	7	,121	-22,508	3,017	-,629
	8	,045	-28,505	3,789	-,785
	9	,017	-34,483	4,567	-,944
	10	,006	-40,456	5,349	-1,105
	11	,002	-46,429	6,131	-1,267
	12	,001	-52,401	6,914	-1,429
	13	,000	-58,374	7,698	-1,591
	14	,000	-64,346	8,481	-1,753
	15	,000	-70,319	9,265	-1,915
	16	,000	-76,292	10,049	-2,077
	17	,000	-82,265	10,833	-2,239
	18	,000	-88,238	11,617	-2,401
	19	,000	-94,211	12,400	-2,564
	20	,000	-100,184	13,184	-2,726

a. Méthode : Entrée

b. La constante est incluse dans le modèle.

c. -2log-vraisemblance initiale : 27,726

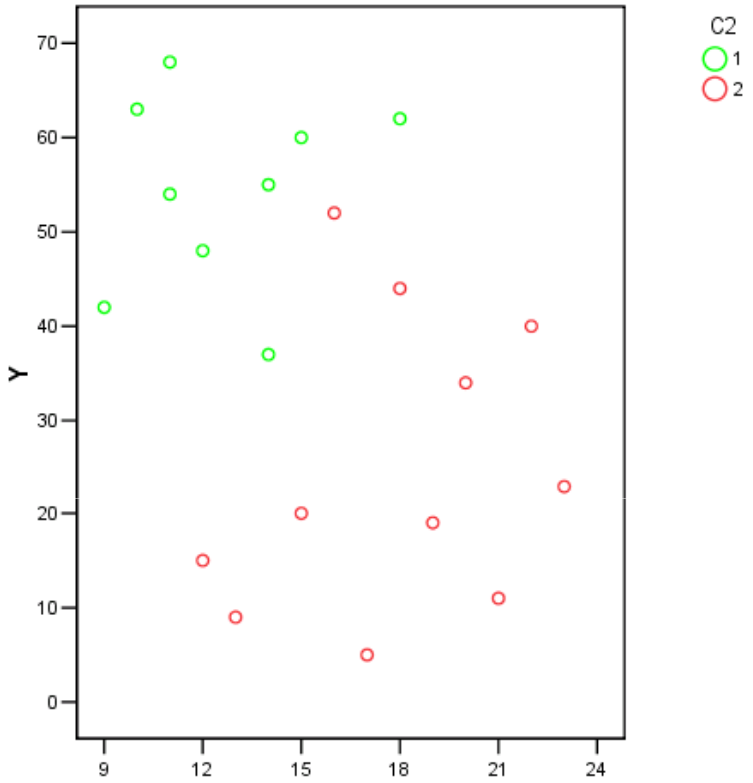
d. L'estimation a été interrompue au numéro d'itération 20 parce que le nombre maximal d'itérations a été atteint. Solution finale introuvable.

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	E
Etape 1[a]	X	13,184	2237,865	,000	1	,995	53
	Y	-2,726	441,662	,000	1	,995	
	Constante	-100,184	21856,781	,000	1	,996	

a. Variable(s) entrées à l'étape 1 : X, Y.

Séparation incomplète des groupes



Historique des itérations[a,b,c,d]

Itération		-2log-vrais emblance	Coefficients		
			Constante	X	Y
Etape 1	1	11,036	-,620	,204	-,062
	2	7,473	-1,523	,373	-,100
	3	5,973	-3,054	,583	-,136
	4	5,323	-5,345	,840	-,172
	5	5,079	-7,956	1,113	-,207
	6	5,020	-9,952	1,321	-,234
	7	5,014	-10,746	1,406	-,245
	8	5,014	-10,840	1,417	-,247
	9	5,014	-10,841	1,417	-,247
	10	5,014	-10,841	1,417	-,247

a. Méthode : Entrée

b. La constante est incluse dans le modèle.

c. -2log-vraisemblance initiale : 27,526

d. L'estimation a été interrompue au numéro d'itération 10 parce que les estimations de paramètres ont changé de moins de ,001.

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
								Inférieur	Supérieur
Etape 1[a]	X	1,417	1,379	1,056	1	,304	4,124	,276	61,535
	Y	-,247	,189	1,696	1	,193	,781	,539	1,133
	Constante	-10,841	13,949	,604	1	,437	,000		

a. Variable(s) entrées à l'étape 1 : X, Y.

Illustration du découpage en classes

- Un même modèle de score avec 4 variables explicatives :
  - continues
  - découpées en classes considérées comme var. ordinales
  - découpées en classes considérées comme var. nominales
- Comparaison des performances

Aire sous la courbe ROC

Variable(s) de résultats tests	Zone	Erreur Std.[a]	Signif. asymptotique[b]	Intervalle de confiance 95% asymptotique	
				Borne inférieure	Borne supérieure
Var explicatives en classes ordinales	,834	,008	,000	,818	,850
Var explicatives en classes nominales	,836	,008	,000	,820	,852
Var explicatives continues	,820	,010	,000	,801	,839

a. Dans l'hypothèse non-paramétrique

b. Hypothèse nulle /: zone vraie = 0.5

- Le découpage en classes nominales l'emporte

Estimation des coefficients

Les données

vecteur X	Y
x^1	y^1
⋮	⋮
x^i	y^i
⋮	⋮
x^n	y^n

$y^i = 0$ ou 1

Le modèle

$$\pi(x^i) = P(Y = 1 / X = x^i) = \frac{e^{\beta_0 + \sum_j \beta_j x^i{}_j}}{1 + e^{\beta_0 + \sum_j \beta_j x^i{}_j}}$$

Recherche du maximum de vraisemblance

- Vraisemblance = probabilité d'obtenir les données observées $[(x^1,y^1),(x^2,y^2),\ldots,(x^n,y^n)]$, exprimée en fonction des coefficients β_i

$$= \prod_{i=1}^{n} \text{Prob}(Y = y^i / X = x^i) \quad = \prod_{i=1}^{n} \pi(x^i)^{y^i} (1-\pi(x^i))^{1-y^i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\sum_j \beta_j x^i{}_j}}{1+e^{\beta_0+\sum_j \beta_j x^i{}_j}}\right)^{y^i} \left(1-\frac{e^{\beta_0+\sum_j \beta_j x^i{}_j}}{1+e^{\beta_0+\sum_j \beta_j x^i{}_j}}\right)^{1-y^i} \quad = L(\beta_0,\beta_1,\ldots,\beta_p)$$

- On cherche les coefficients β_i maximisant la vraisemblance et ajustant donc le mieux possible les données observées
- Pas de solution analytique $\Rightarrow$ utiliser une méthode numérique itérative (ex : Newton-Raphson)

Cas de la régression logistique simple

- On cherche 2 coefficients β_0 et β_1 maximisant la vraisemblance
  - $L(\beta_0,\beta_1) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x^i}}{1+e^{\beta_0+\beta_1 x^i}}\right)^{y^i} \left(1-\frac{e^{\beta_0+\beta_1 x^i}}{1+e^{\beta_0+\beta_1 x^i}}\right)^{1-y^i}$
- Pour ces coefficients, la matrice des covariances

$$V(\beta) = \begin{bmatrix} V(\beta_0) & Cov(\beta_0,\beta_1) \\ Cov(\beta_0,\beta_1) & V(\beta_1) \end{bmatrix}$$

- est estimée par la matrice

$$\left[-\frac{\partial^2 Log\ L(\beta)}{\partial \beta^2}\right]^{-1}_{\beta=(\beta_0,\beta_1)}$$

intervient dans la statistique de Wald (voir + loin)

- Il faut inverser la matrice hessienne $H = \partial^2 LogL(\beta)/\partial\beta^2$
  - impossible en cas de séparation complète des groupes

Vraisemblance et déviance d'un modèle

- Soit $L(\beta_0)$ = vraisemblance du modèle réduit à la constante
- Soit $L(\beta_n)$ = vraisemblance du modèle saturé (avec toutes les variables explicatives et toutes les interactions pour en avoir autant que d'observations distinctes) = vraisemblance maximale
- Soit $L(\beta_k)$ = vraisemblance du modèle avec k variables

- On définit la déviance :

  $D(\beta_k) = -2\,[\text{Log } L(\beta_k) - \text{Log } L(\beta_n)] = \text{Log } [L(\beta_n)/\, L(\beta_k)]^2$

  $= -2 \text{ Log } L(\beta_k)$ puisque $L(\beta_n) = 1$ pour une cible 0/1
- But de la régression logistique : maximiser la vraisemblance $L(\beta_k) \Leftrightarrow$ minimiser la déviance $D(\beta_k)$
  - $L(\beta_k)$ petit $\in [0,1] \Rightarrow$ -2 Log $L(\beta_k) \in [0,+\infty[$ avec un terme « 2 » pour avoir l'analogie entre déviance et Σ(erreurs)²

Comparaison de modèles

- Pour savoir s'il convient d'ajouter q variables explicatives à un modèle qui en contient déjà k
- On calcule la différence des déviances
  - $D(\beta_k) - D(\beta_{k+q}) = -2 [\text{Log } L(\beta_k) - \text{Log } L(\beta_{k+q})]$
- Sous l'hypothèse H_0 de la nullité des l derniers coefficients, $D(\beta_k) - D(\beta_{k+q})$ suit un χ^2 à q d° de liberté
  - Sous le seuil critique de la valeur du χ^2 ($\Leftrightarrow$ si la probabilité dépasse 0,05) : on rejette les q nouvelles variables

- Méthode la plus utilisée en régression pas à pas

Autres indicateurs

- Cas particulier
  - $D(\beta_0) - D(\beta_k) = -2\,[\text{Log } L(\beta_0) - \text{Log } L(\beta_k)]$
- suit une loi du χ^2 à *k* degrés de liberté sous l'hypothèse H_0 de la nullité de tous les coefficients $\beta_1, \ldots, \beta_k$. Rejet de H_0 si cette différence dépasse le seuil critique du χ^2.
- Critère d'Akaike $AIC = -2 \text{ Log } L(\beta_k) + 2(k+1)$
  - k = nb de ddl = nb de paramètres à estimer
- Critère de Schwartz $BIC = -2 \text{ Log } L(\beta_k) + (k+1).\log n$
  - n = nb total d'individus
  - pénalise les modèles complexes
- Ces 2 critères permettent de comparer 2 modèles
  - ils doivent être le plus bas possible

Le χ^2 de Wald

- Statistique de Wald = (β_i / écart-type(β_i))²
- suit un χ^2 à 1 degré de liberté sous l'hypothèse nulle H_0 : le coefficient $\beta_i = 0$

> teste la significativité de chaque coefficient β_i

  - en comparant le sous-modèle excluant X_i avec le modèle incluant toutes les variables
  - on doit avoir Wald > 4 (plus précisément 3,84 = 1,96² venant du test de Student)

- Méthode utilisée en régression pas à pas
- NB : Éviter le χ^2 de Wald si peu d'observations ou si les coefficients β_i sont grands
- NB : Pour les variables qualitatives à plus de 2 modalités, la significativité du résultat de ce test dépend du choix de la modalité de référence

Le χ^2 de Wald (suite)

- Wald > 3,84 = 1,96² ⇔ Intervalle de confiance de l'odds-ratio ne contient pas 1

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
								Inférieur	Supérieur
Etape 1[a]	AGE	,111	,024	21,254	1	,000	1,117	1,066	1,171
	Constante	-5,309	1,134	21,935	1	,000	,005		

a. Variable(s) entrées à l'étape 1 : AGE.

> 3,84

odds-ratio

1 ∉ IC

Influence du choix de la modalité de référence

Codages des variables nominales

		Fréquence	Codage des paramètres (1)	(2)	(3)
CLASS	0	885	1,000	,000	,000
	1	325	,000	1,000	,000
	2	285	,000	,000	1,000
	3	706	,000	,000	,000

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)
Etape 1[a]	CLASS			173,228	3	,000	
	CLASS(1)	-,068	,117	,336	1	,562	,934
	CLASS(2)	1,596	,144	123,520	1	,000	4,936
	CLASS(3)	,740	,148	24,920	1	,000	2,096
	Constante	-1,087	,087	157,383	1	,000	,337

a. Variable(s) entrées à l'étape 1 : CLASS.

- Le choix de la modalité de référence influe sur la significativité des coefficients !

Codages des variables nominales

		Fréquence	Codage des paramètres (1)	(2)	(3)
CLASS	0	885	1,000	,000	,000
	1	325	,000	,000	,000
	2	285	,000	1,000	,000
	3	706	,000	,000	1,000

Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp(B)
Etape 1[a]	CLASS			173,228	3	,000	
	CLASS(1)	-1,664	,139	143,335	1	,000	,189
	CLASS(2)	-,856	,166	26,593	1	,000	,425
	CLASS(3)	-1,596	,144	123,520	1	,000	,203
	Constante	,509	,115	19,757	1	,000	1,664

a. Variable(s) entrées à l'étape 1 : CLASS.

Test de Hosmer et Lemeshow

Test peu puissant : accepte facilement les modèles sur les petits effectifs

Tableau de contingence pour le test de Hosmer-Lemeshow

		CHD = 0		CHD = 1		
		Observé	Théorique	Observé	Théorique	Total
Etape 1	1	9	9,213	1	,787	10
	2	9	8,657	1	1,343	10
	3	8	8,095	2	1,905	10
	4	8	8,037	3	2,963	11
	5	7	6,947	4	4,053	11
	6	5	5,322	5	4,678	10
	7	5	4,200	5	5,800	10
	8	3	3,736	10	9,264	13
	9	2	2,134	8	7,866	10
	10	1	,661	4	4,339	5

Test de Hosmer-Lemeshow

Etape	Khi-deux	ddl	Signif.
1	,890	8	,999

très bon ajustement

On découpe les observations en g = 10 groupes, ordonnés par probabilité croissante (fournie par le modèle)

On calcule le χ^2 du tableau gx2 des fréquences pour l'événement modélisé (ici CHD = 1) et l'événement contraire, que l'on compare à la loi du χ^2 à (g - 2) degrés de libertés

Si le χ^2 est grand (la proba est faible), les fréquences observées et attendues sont significativement différentes et le modèle ne s'ajuste pas bien aux données

Autres tests (sur SPSS)

Récapitulatif du modèle

Etape	-2log-vrais emblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	107,353[a]	,254	,341

Déviance

$-2 \text{ Log } L(\beta_k)$

(est > 0 puisque $L(\beta_k) \in [0,1]$)

« Équivalent » du R^2 de la régression linéaire - ne peut atteindre 1

R^2 ajusté pour varier entre 0 et 1

Autres tests (sur SAS : proc logistic)

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	138.663	111.353
SC	141.268	116.563
-2 Log L	136.663	107.353

← déviance

R² de Cox & Snell →

R-Square	0.2541	Max-rescaled R-Square	0.3410

← R² de Nagelkerke

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	29.3099	1	<.0001
Score	26.3989	1	<.0001
Wald	21.2541	1	<.0001

Matrice de confusion (avec SAS)

Table de classification									
	Correct		Incorrect		Pourcentages				
Niveau de prob.	Événement	Non-événement	Événement	Non-événement	Correct	Sensibilité	Spécificité	POS fausse	NEG fausse
0.000	57	0	43	0	57.0	100.0	0.0	43.0	.
0.100	57	1	42	0	58.0	100.0	2.3	42.4	0.0
0.200	55	7	36	2	62.0	96.5	16.3	39.6	22.2
0.300	51	19	24	6	70.0	89.5	44.2	32.0	24.0
0.400	50	25	18	7	75.0	87.7	58.1	26.5	21.9
0.500	**45**	**27**	**16**	**12**	**72.0**	**78.9**	**62.8**	**26.2**	**30.8**
0.600	41	32	11	16	73.0	71.9	74.4	21.2	33.3
0.700	32	36	7	25	68.0	56.1	83.7	17.9	41.0
					63.0	42.1	90.7	14.3	45.8
					48.0	10.5	97.7	14.3	54.8
					43.0	0.0	100.0	.	57.0

prédit ➡ Observé ⬇	0	1	total
0	45	12	57
1	16	27	43
total	61	39	100

Correct = (45 + 27) / 100 = 72 %
Sensibilité = 45 / 57 = 78,9 %
Spécificité = 27 / 43 = 62,8 %
POS fausse = 16 / 61 = 26,2 %
NEG fausse = 12 / 39 = 30,8 %

Syntaxe SAS de la régression logistique

```
ods rtf file= « c:\logistic_sas.doc »;
proc logistic data=matable.ascorer outmodel=mon.modele;
   class
   var_quali_1 (ref='A1') … var_quali_i (ref='Ai') / param=ref;
   model cible (ref='0')=
   var_quali_1 … var_quali_i var_quanti_1 var_quanti_j
   / selection=forward sle=.05 maxiter=25 outroc=roc  rsquare lackfit
   ctable;
   output out=matable.scoree predicted=proba resdev=deviance;
run;
symbol1 i=join v=none c=blue;
proc gplot data=roc;
     where _step_ in (1 7);
     title 'Courbe ROC';
     plot _sensit_*_1mspec_=1 / vaxis=0 to 1 by .1 cframe=ligr;
run;
ods rtf close ;
proc logistic inmodel=mon.modele; score data= autretable.ascorer;run;
```

Hosmer-Lemeshow

R^2

niv. de signif. en entrée

enregistre la probabilité prédite pour l'événement

matrice de confusion

Tests de concordance

- Soit n_1 (resp. n_2) le nb d'observations où Y=0 (resp. Y=1)
- Soit $n = n_1 + n_2$ le nb total d'observations
- On s'intéresse aux $t = n_1 n_2$ paires formées d'une observation où Y = 1 et d'une observation où Y = 0
- Parmi ces t paires : on a concordance si la proba estimée que Y = 1 est + grande quand Y = 1 que quand Y = 0
- Soient nc = nb de paires concordantes ; nd = nb de paires discordantes ; t - nc - nd = nb d'ex-æquo (« tied »)
- D de Somers = (nc - nd) / t = indice Gini
- Gamma = (nc - nd) / (nc + nd)
- Tau-a = 2 (nc - nd) / n(n-1)
- c = (nc + 0,5[t - nc - nd]) / t = aire sous la courbe ROC
- Plus ces indices sont proches de 1, meilleur est le modèle

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.0	Somers' D	0.600
Percent Discordant	19.0	Gamma	0.612
Percent Tied	2.0	Tau-a	0.297
Pairs	2451	c	0.800

Effet de la multicolinéarité

- Régression logistique avec 2 variables VAR1 et VAR2 fortement corrélées :

		VAR1	VAR2
VAR1	Corrélation de Pearson	1	,975**
	N	36841	36300
VAR2	Corrélation de Pearson	,975**	1
	N	36300	36300

**. La corrélation est significative au niveau 0,01

- On constate une dégradation du pouvoir prédictif de VAR1 avec l'introduction de VAR2 :

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
								Inférieur	Supérieur
Etape 1[a]	VAR1	,098	,004	759,291	1	,000	1,103	1,096	1,111
	Constante	-4,898	,062	6290,898	1	,000	,007		

a. Variable(s) entrées à l'étape 1: VAR1.

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0%	
								Inférieur	Supérieur
Etape 2[a]	VAR1	,020	,014	2,125	1	,145	1,020	,993	1,048
	VAR2	,092	,015	39,280	1	,000	1,096	1,065	1,129
	Constante	-4,993	,065	5867,055	1	,000	,007		

a. Variable(s) entrées à l'étape 2: VAR2.

Résumé des tests

- Test du χ^2 sur indicateur de Wald (> 4)
- $1 \notin$ IC à 95 % de l'odds-ratio = $\exp(a_i \pm 1{,}96\sigma(a_i))$
- Test du χ^2 sur -2 [Log $L(\beta_0)$ – Log $L(\beta_k)$]
- (Test de Hosmer et Lemeshow sur comparaison des proportions observées et théoriques)
- R^2 de Cox-Snell et R^2 ajusté de Nagelkerke
- AIC et BIC
- Multicolinéarité (tolérance, VIF, indices de conditionnement)
- Matrice de confusion, tests de concordance, aire sous la courbe ROC
- Moins de 20 degrés de liberté (variables ou modalités) sont souvent retenus

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.0	Somers' D	0.600
Percent Discordant	19.0	Gamma	0.612
Percent Tied	2.0	Tau-a	0.297
Pairs	2451	c	0.800

Influence de l'échantillonnage 1/2

- La régression logistique consiste à écrire $\pi(x) := P(Y=1/X=x)$ sous la forme

$$Log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

  - avec des coefficients maximisant la vraisemblance
- Si l'on effectue un échantillonnage E <u>indépendant</u> de X, alors la probabilité $\pi_E(x) := P(Y=1/X=x, X \in E)$ vérifie

$$Log\left(\frac{\pi_E(x)}{1-\pi_E(x)}\right) = \beta'_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

  - avec $\beta'_0 = \beta_0 + \text{constante} \; (= \log(p_{1,E}/p_{0,E}) + \log(p_0/p_1))$
  - p_i = proportion de cas Y=i dans la population totale
  - $P_{i,E}$ = proportion de cas Y=i dans l'échantillon E
- Ceci est vrai de logit mais non de probit !

Influence de l'échantillonnage 2/2

- Si E est indépendant de X, la même fonction de score permet de décider si Y=1 (en changeant seulement le seuil de décision)
  - cas particulier : $p_{1,E}/p_{0,E} = p_1/p_0 \Rightarrow \beta'_0 = \beta_0$
- Un score calculé sur une sous-population E peut s'appliquer à une sous-population E', si la distribution des variables explicatives est la même dans E et E', même si l'événement à prédire est plus rare dans E'
  - en appliquant le calcul de $P(Y=1/X=x,X \in E)$ aux $X \in E'$ et en fixant le même seuil d'acceptation $P(Y=1/X=x,X \in E) > s_o$, on aura le même % d'acceptés dans E' (puisque les var. explicatives ont mêmes distributions dans E et E'), mais la fréquence de l'événement sera plus faible dans les acceptés de E', puisque leur proba $P(Y=1/X=x,X \in E') < P(Y=1/X=x,X \in E)$

Avantages de la régression logistique

- Permet de traiter les variables explicatives discrètes, qualitatives ou continues
- Permet de traiter une variable cible ordinale ou nominale
- Hypothèses + générales que l'analyse discriminante (pas de multinormalité ni d'homoscédasticité)
- Permet de traiter les réponses non monotones
- Odds-ratios facilement interprétables (pour modèle *logit*)
- Peut prendre en compte les interactions entre variables
- Modélise directement une probabilité
- Fournit des intervalles de confiance sur les résultats
- Nombreux tests statistiques disponibles
- Possibilité de sélection pas à pas des variables

Limites de la régression logistique

- Suppose la non-colinéarité des variables explicatives
- Approximation numérique :
  - calcul itératif moins rapide que le calcul direct de l'analyse discriminante
  - moindre précision que l'analyse discriminante quand les hypothèses de cette dernière sont satisfaites
  - ne converge pas toujours vers une solution optimale
  - inopérant dans le cas de la séparation complète des groupes ! puisque la log-vraisemblance s'approche de 0 (iris de Fisher et séparation des Setosa !)
- Ne traite pas les valeurs manquantes de variables continues (sauf découpage en classes)
- Sensible aux valeurs hors norme de variables continues (sauf découpage en classes)

La régression logistique ordinale 1/2

- La variable cible Y est ordinale
- Fonctions de lien :
  - logit
  - probit
  - log-log : Log [– Log(1–μ)]
    - utilisé quand les valeurs élevées de la cible sont plus probables
    - ex : valeurs 3 à 5 / 5 dans une enquête de satisfaction
  - Cauchit : tg[π(μ -0,5)]
    - utilisé quand les valeurs extrêmes de la cible sont plus probables
    - ex : valeur 5 / 5 dans une enquête de satisfaction

La régression logistique ordinale 2/2

- Y prend m valeurs ordonnées, notées 1, 2, …, m
- Dans le modèle à pentes égales : on suppose que le logit des probabilités cumulatives s'écrit sous la forme

$$\text{logit}\big(\text{Prob}(Y \le r / X = x)\big) = \alpha_r + \sum_i \beta_i x_i, \; pour \; 1 \le r < m$$

> Seule la constante dépend de *r*

- On parle de « proportional odds model » car :

$$\frac{\text{Prob}(Y \le r / X = x) / \text{Prob}(Y > r / X = x)}{\text{Prob}(Y \le r / X = x') / \text{Prob}(Y > r / X = x')} = \frac{\exp(\alpha_r + \sum_i \beta_i x_i)}{\exp(\alpha_r + \sum_i \beta_i x'_i)} = \exp\left(\sum_i \beta_i (x_i - x'_i)\right)$$

> Les odds-ratios pour un *r* fixé sont tous proportionnels entre eux et le rapport ne dépend pas de *r*

- Le modèle à pentes différentes : vite très complexe

La régression logistique multinomiale

- Y prend m valeurs non ordonnées, notées 1, 2, …, m
- On choisit une modalité de référence, par exemple m
- On écrit les probabilités sous la forme :

$$\text{Prob}(Y = j / X = x) = \frac{\exp\left(\alpha_j + \sum_k \beta_{jk} x_k\right)}{1 + \sum_{i=1}^{m-1} \exp\left(\alpha_i + \sum_k \beta_{ik} x_k\right)}, \; j = 1, ..., m-1$$

$$\text{Prob}(Y = m / X = x) = \frac{1}{1 + \sum_{i=1}^{m-1} \exp\left(\alpha_i + \sum_k \beta_{ik} x_k\right)}$$

- C'est un modèle plus complexe que le modèle ordinal à pentes égales, car les coefficients β_{ij} dépendent de *j*

Techniques de classement :

Le modèle linéaire général

Le modèle linéaire généralisé

Le modèle additif généralisé

Terminologie

- Covariables = variables explicatives continues (quantitatives)
- Facteurs = variables explicatives catégorielles (qualitatives)
  - niveaux d'un facteur = ses modalités

Effets fixes et aléatoires 1/2

- Effets fixes des facteurs et covariables
  - contrôlés par l'expérimentateur
  - en prenant toutes les valeurs
  - dont on veut quantifier l'effet sur la variable cible
  - similaire à une analyse de régression – but prédictif
- Effets aléatoires des facteurs et covariables
  - en prenant un échantillon de valeurs
  - on veut quantifier la proportion de la variance de la variable cible qu'ils expliquent
  - similaire à une analyse de corrélation
  - à but descriptif et non prédictif
- Effets mixtes
  - Présence d'effets fixes et aléatoires

Effets fixes et aléatoires 2/2

- Ex 1 : comparaison de 2 traitements sur plusieurs patients dans plusieurs hôpitaux – mettre la variable « hôpital » en effet aléatoire
  - permet d'éviter le biais dû au lieu où est administré le traitement
  - ne permet pas de prédire le résultat dans un nouvel hôpital
- Ex 2 : comparaison de 2 conditionnements d'un produit sur les achats de plusieurs consommateurs dans plusieurs magasins – mettre la variable « magasin » en effet aléatoire
  - permet d'éviter le biais dû au lieu d'achat
  - ne permet pas de prédire les achats dans un nouveau magasin

Modèle à mesures répétées 1/2

- Les mesures y_1, y_2, … y_k de Y à prédire sur plusieurs individus sont corrélées (données longitudinales) car
  - il s'agit d'un même individu observé k fois (par ex : avant et après un traitement médical)
  - ou de k individus partageant une caractéristique commune (même famille, même segment)
- On sort des hypothèses de la régression linéaire et de la régression logistique qui supposent l'absence de corrélation des mesures sur plusieurs individus
- Y peut être continue ou discrète
- Un modèle à mesures répétées peut traiter à la fois des effets fixes et aléatoires

Modèle à mesures répétées 2/2

- Dans un modèle à mesures répétées, on a des effets :
  - intra-individus (« within-subject effects »)
    - influence du temps, du traitement (comparaison du patient avant et après traitement)
    - généralise la comparaison de moyennes sur 2 échantillons appariés
  - inter-individus (« between-subject effects »)
    - influence des caractéristiques du patient, telles que l'âge, le sexe, la formulation sanguine… (comparaison du patient par rapport aux autres)
  - interactions intra-inter (« within-subject-by-between-subject effects »)
    - interactions du traitement et des caractéristiques du patient

Application aux données de survie 1/2

- Pour chaque individu, les observations sont répétées dans le temps à des instants t_1, t_2,..., t_N
- On s'intéresse à la survenue d'un événement (par ex : décès, départ) à un instant t_i, modélisée par la var cible :
  - $y_k = 0$ si $k < i$, $y_i = 1$, pas d'observation si $k > i$: on connaît le délai de survenue de l'événement
  - on a $y_k = 0$ pour tout $k \leq N$ si l'événement ne survient pas (et si l'individu est observé jusqu'au bout) : on ne connaît que la limite inférieure du délai de survenue de l'événement (cette donnée est « censurée »)
  - la donnée est aussi censurée si l'individu est perdu de vue avant la fin et avant la survenance de l'événement
- On cherche à expliquer la variable « délai de survie » pour mettre en évidence les facteurs favorables

Application aux données de survie 2/2

- Un modèle courant (de Cox)

Modèle de survie de Kaplan-Meier

- Modélise la durée avant l'apparition d'un événement (décès, départ…)
- Certaines données sont censurées (encore vivant), mais on doit en tenir compte (les durées de vie + longues étant + censurées, par définition)
- On cherche des modèles intégrant à la fois les données censurées et non censurées
- Le modèle de Kaplan-Meier permet de calculer une estimation non paramétrique de la fonction de survie : S(t) = Prob(durée de vie > t)
- Il permet de comparer les fonctions de survie (et les courbes de survie) de plusieurs échantillons (« strates »)
  - correspondant par ex. à plusieurs traitements médicaux différents
  - et d'effectuer des tests

Modèle de survie de Cox 1/3

- Même champ d'application que le modèle de Kaplan-Meier
- Le modèle de régression de Cox à hasards proportionnels (1972) permet d'ajouter p variables explicatives et d'estimer leurs coefficients dans la fonction de survie, donc leur impact sur la durée de vie
  - ex : sexe / nb cigarettes fumées par jour
- C'est un modèle semi-paramétrique (forme paramétrique (Σ) pour les effets des var. explicatives, et forme non paramétrique de la fonction de survie)
- Pour tout individu i de var. explicatives x_{ij}, la fonction de survie s'exprime sous la forme :

$$S(t, x_i) = S_0(t)^{\exp(\sum_{j=0}^{p} \beta_j x_{ij})}$$

  - où $x_{i0} = 1 \ \forall i$ et $S_o(t)$ est la fonction de survie de base (« hasard de base »),
  - et où l'on recherche le vecteur β supposé indépendant de i.

Modèle de survie de Cox 2/3

- On trouve le vecteur β des coefficients de régression par maximisation d'une fonction de vraisemblance (comme la régression logistique)
  - plusieurs méthodes de sélection des var. explicatives existent (ascendante, descendante, pas à pas)
  - interprétation des « odds-ratios »
- Les données censurées :
  - n'interviennent pas dans le calcul de β
  - interviennent dans le calcul de $S_o(t)$
- Le terme de « hasards proportionnels » vient de ce que le rapport $h_i(t) / h_k(t)$ ne dépend pas de t
  - sauf si les x_{ij} en dépendent

Modèle de survie de Cox 3/3
(fonctionnalités supplémentaires)

- Les variables explicatives x_{ij} peuvent dépendre ou non de *t*
  - soit en étant une fonction de t, soit en prenant une valeur différente par valeur de t
- On peut faire des analyses stratifiées (sur des échantillons différents), en supposant que le vecteur β des coefficients de régression est indépendant de l'individu i et de la strate
  - en revanche, le hasard de base $S_o(t)$ dépend de la strate
  - d'où l'utilisation des analyses stratifiées sur une strate X_j quand une variable explicative X_j ne satisfait pas l'hypothèse des hasards proportionnels
    - X_j n'intervient plus dans le terme exp() mais intervient dans $S_o(t)$

Modèle linéaire général (GLM)

- Généralise la régression linéaire multiple de plusieurs façons
- Les variables explicatives peuvent non seulement être continues, mais :
  - qualitatives (ANOVA)
  - continues et qualitatives (ANCOVA)
- Il peut y avoir plusieurs variables continues à expliquer
  - MANOVA, MANCOVA
- Prise en compte des modèles à effet fixes, aléatoires ou mixtes
- Prise en compte des modèles à mesures répétées

Modèle linéaire généralisé (GLZ)

- Généralise le modèle linéaire général quand Y à prédire n'est plus forcément continue
- On écrit $g(E(Y/X=x)) = \beta_0 + \Sigma_i \beta_i x_i$
- g = fonction de lien monotone différentiable (g^{-1} = fonction de transfert)
- La distribution de Y/X=x peut être :
  - normale (continue : régression) $g(\mu) = \mu$
  - gamma (continue positive) $g(\mu) = -1/\mu$
  - Bernoulli (discrète : oui/non) $g(\mu)=\log(\mu/1-\mu)$... (logit, probit, log-log)
  - de Poisson (discrète : comptage) $g(\mu) = \log(\mu)$
    - Y = nb de sinistres (assurance) ou effectif d'un tableau de contingence (modèle log-linéaire)
  - multinomiale, etc.

Modèle linéaire généralisé (GLZ)

- Double généralisation du modèle linéaire général : loi de Y/X=x non normale et $g \neq 1$
- Estimation du modèle : par la méthode du maximum de vraisemblance (analogue des moindres carrés)
- Évaluation du modèle : par calcul de la déviance des log-vraisemblances (analogue de la somme des carrés des résidus de la régression) et test du χ^2
- Existence d'une régression logistique à mesures répétées (proc GENMOD de SAS)
- Variable V « offset » : sert à tarer un modèle si la variable cible dépend linéairement de V
  - le nb de sinistres dans une compagnie d'assurance doit être équilibré par la variable offset « nb de contrats »
- Source : Nelder-Wedderburn (1972)

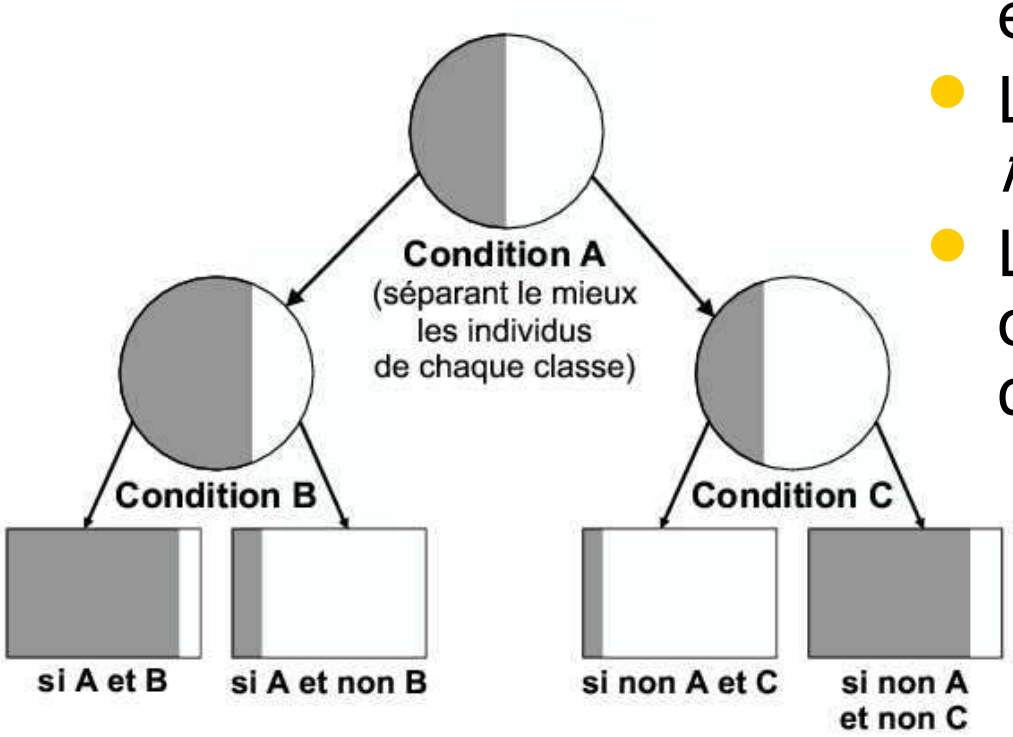
Modèle additif généralisé (GAM)

- On écrit $g(E(Y/X=x)) = \beta_0 + \Sigma_i f_i(x_i)$
- g : fonction de lien (g^{-1} : fonction de transfert)
- f_i : fonction quelconque (non-paramétrique : on n'a plus un simple paramètre comme le coefficient β_i) de x_i
  - par ex : f_i = fonction spline
- Mais le modèle reste additif (c'est Σ_i qui combine les f_i)
- La distribution de Y peut être normale, poissonienne ou binomiale
  - ex : modèle logistique additif généralisé si $g(\mu) = \log(\mu/1- \mu)$
- Modélisation puissante mais attention au sur-apprentissage et à l'interprétabilité des résultats
- Source : Hastie - Tibshirani (1990)

Technique de classement ou prédiction :

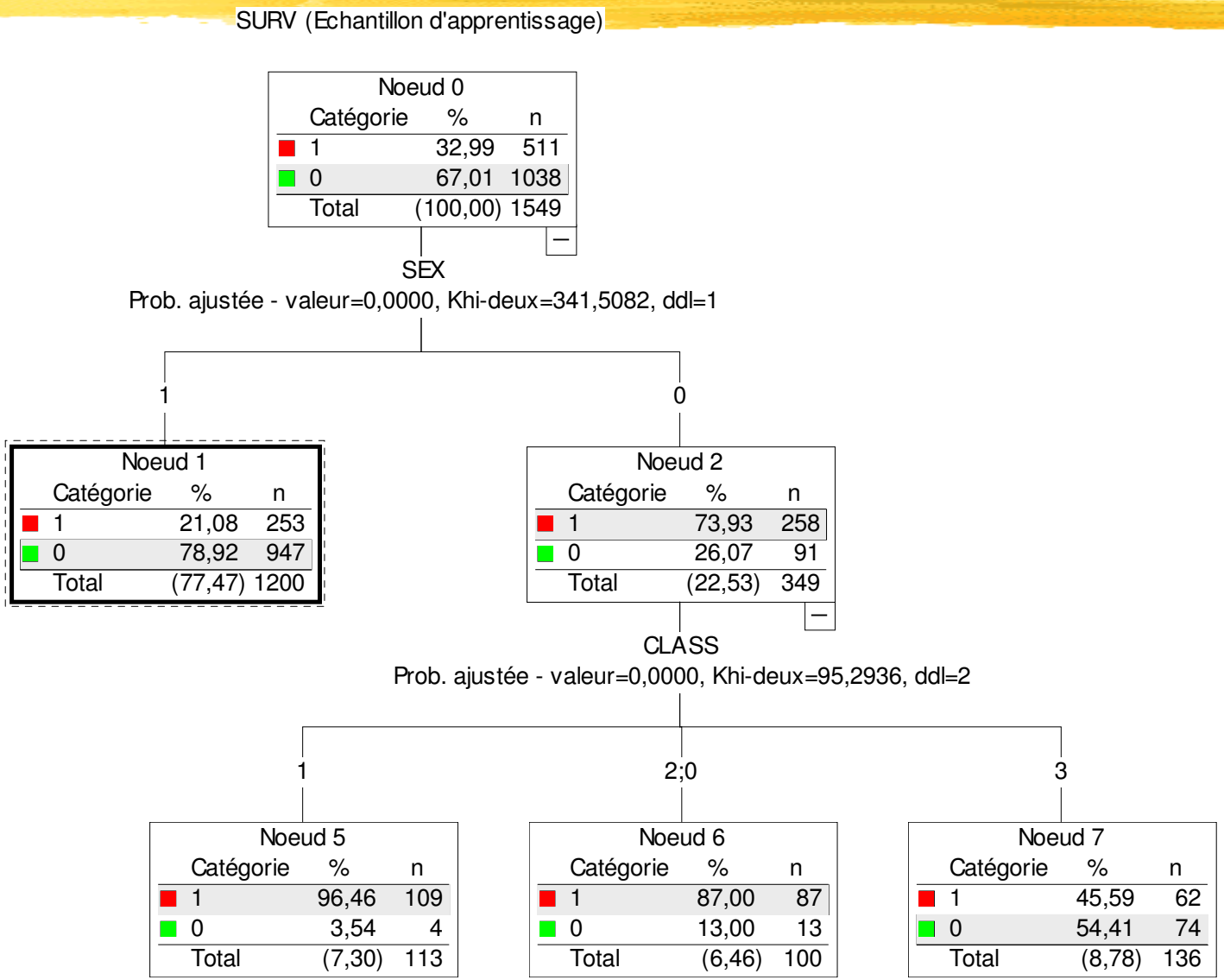
Arbres de décision

Classement par arbre de décision



- Le premier nœud de l'arbre est la *racine*
- Les nœuds terminaux sont les *feuilles*
- Le chemin entre la racine et chaque feuille est l'expression d'une règle
  - par exemple : les clients dont l'âge est < *x*, les revenus < *y* et le nombre de comptes > *z* appartiennent dans *n* % des cas à la classe C
- Si chaque nœud de l'arbre a au plus deux nœuds fils, on dit que l'arbre est *binaire*

Arbre de classement

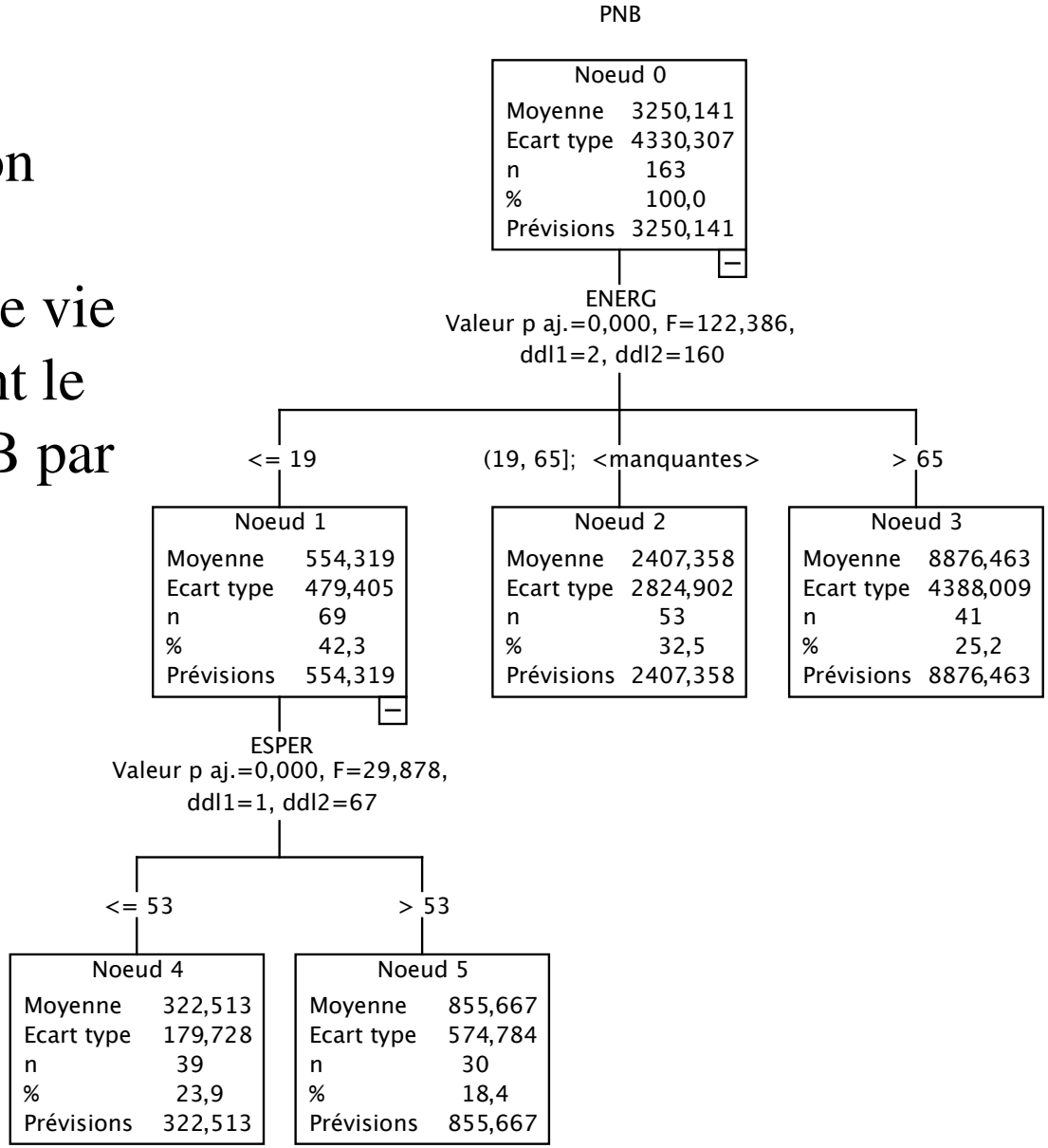


Prédiction par arbre de décision

- Les arbres peuvent s'appliquer à la prédiction : la variable à expliquer X est continue
- C'est une alternative à la régression linéaire multiple
- Principe :
  - la variable X doit avoir une variance + faible dans les nœuds fils
  - la variable X doit avoir une moyenne la + distincte possible d'un nœud fils à un autre

Arbre de régression

Ce sont la consommation d'énergie et l'espérance de vie qui expliquent le mieux le PNB par habitant



Classement par arbre de décision

- Pour répartir les individus d'une population en *n* classes, on commence par choisir la variable séparant le mieux les individus de chaque classe en fonction de la variable cible, en sous-populations appelées *nœuds* : le critère précis (C1) de choix de la variable et de sa valeur testée dépend de chaque type d'arbre

- Pour chaque nœud, on répète la même opération, ce qui donne naissance à un ou plusieurs *nœuds fils*. Chaque nœud fils donne à son tour naissance à un ou plusieurs nœuds, et ainsi de suite, jusque ce que :
  - la séparation des individus ne soit plus possible
  - OU un certain critère (C2) d'arrêt d'approfondissement de l'arbre soit satisfait

Critère d'arrêt d'un arbre (C2)

- Le critère d'arrêt (C2) dépend du type et du paramétrage de l'arbre. Souvent (C2) combine plusieurs règles :
  - la profondeur de l'arbre a atteint une limite fixée
  - OU le nombre de feuilles (c'est-à-dire de règles) a atteint un maximum fixé
  - OU l'effectif de chaque nœud est inférieur à une valeur fixée en deçà de laquelle on estime qu'il ne faut plus diviser un nœud (au moins 75 à 100 pour de bons résultats)
  - OU la division ultérieure de tout nœud provoquerait la naissance d'un fils d'effectif inférieur à une valeur fixée
  - OU la qualité de l'arbre est suffisante
  - OU la qualité de l'arbre n'augmente plus de façon sensible.
- C'est bien entendu sur cette dernière règle que les arbres diffèrent le plus
  - exemple : dans les arbres CART la qualité se mesure par la pureté des nœuds

Principaux critères de scission (C1)

- Le *critère du* χ^2
  - lorsque les variables explicatives sont qualitatives
  - utilisé dans l'arbre CHAID
- L'*indice de Gini, l'indice Twoing et l'entropie*
  - pour tous types de variables explicatives
  - l'indice de Gini est utilisé dans l'arbre CART
  - l'indice Twoing est utilisé dans l'arbre CART lorsque la variable cible a ≥ 3 modalités
  - l'entropie est utilisée dans les arbres C4.5 et C5.0
  - plus les classes sont uniformément distribuées dans un nœud, plus l'indice de Gini et l'entropie sont élevés ; plus le nœud est pur, plus ils sont bas

Les principaux arbres de décision

- CHAID (CHi-Square Automation Interaction Detection)
  - utilise le test du χ^2 pour définir la variable la plus significative et le découpage de ses modalités
  - adapté à l'étude des variables explicatives discrètes
- CART (Classification and Regression Tree)
  - cherche à maximiser la pureté des nœuds
  - adapté à l'étude de tout type de variables explicatives
- C5.0 de J.R. Quinlan
  - cherche à maximiser le gain d'information réalisé en affectant chaque individu à une branche de l'arbre
  - adapté à l'étude de tout type de variables explicatives

Arbre CHAID – Algorithme 1/2

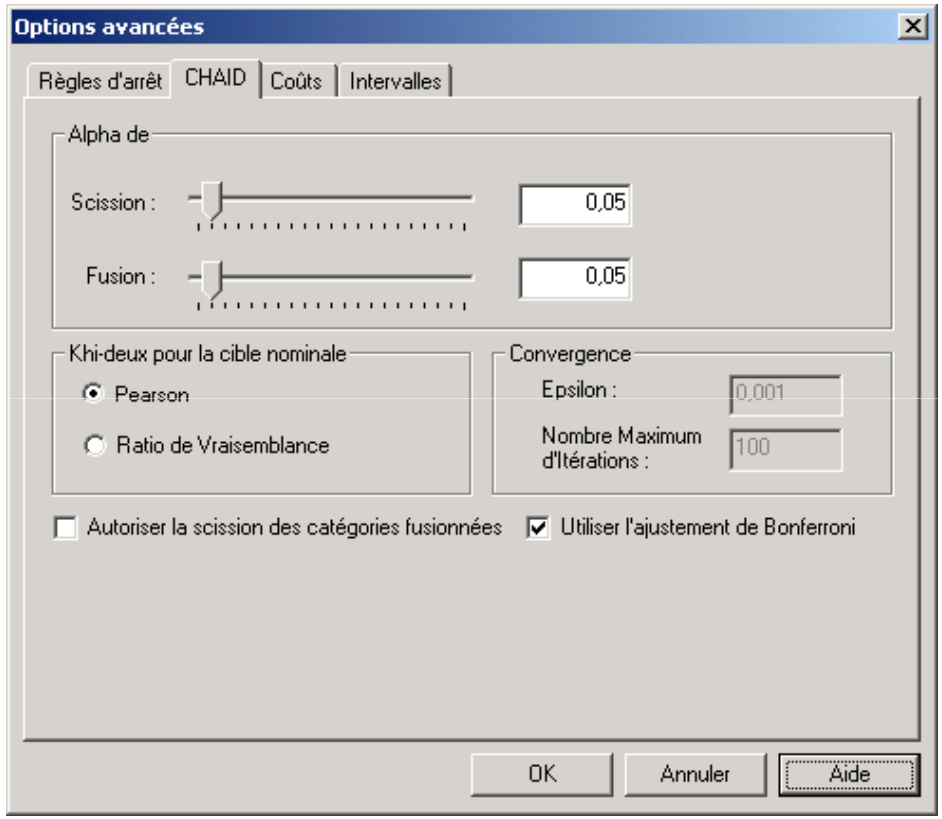
- Cet arbre est de conception plus ancienne (principe : 1975, Hartigan ; algorithme : 1980, Kass)
- Il discrétise automatiquement les variables explicatives continues
- La cible est une variable qualitative à k modalités
- Utilise plusieurs fois la statistique du χ^2 :
  1. On construit pour chaque prédicteur X_i, le tableau de contingence X_i x Y et on effectue les étapes 2 et 3
  2. On sélectionne la paire de modalités de X_i dont le sous-tableau (2 x k) a le plus petit χ^2. Si ce χ^2 n'est pas significatif, on fusionne les 2 modalités et on répète cette étape

Arbre CHAID – Algorithme 2/2

3. Eventuellement, pour chaque modalité composée de plus de 3 modalités originales, on détermine la division binaire au χ^2 le plus grand. S'il est significatif, on effectue cette division
4. On calcule la significativité (probabilité associée au χ^2) de chaque prédicteur X_i dont les modalités ont été précédemment regroupées et on retient le plus significatif. Si ce χ^2 est plus significatif que le seuil choisi, on peut diviser le nœud en autant de nœuds-fils qu'il y a de modalités après regroupement. Si ce χ^2 n'atteint pas le seuil spécifié, le nœud n'est pas divisé

Arbre CHAID – Ajustement de Bonferroni

- Lors du calcul de la significativité de tous les prédicteurs (étape 4), on peut multiplier la valeur de la probabilité du χ^2 par le coefficient de Bonferroni, qui est le nombre de possibilités de regrouper les m modalités d'un prédicteur en g groupes ($1 \leq g \leq m$)
- Ce calcul permet d'éviter la surévaluation de la significativité des variables à modalités multiples



Arbre CHAID – Caractéristiques

- CHAID traite l'ensemble des valeurs manquantes comme une seule catégorie (qu'il fusionne éventuellement avec une autre)
  - pas d'utilisation de variables de substitution
- Il n'est pas binaire et produit des arbres souvent plus larges que profonds
  - utile pour la discrétisation de variables continues
- Il souffre de l'absence de dispositif automatique d'optimisation par élagage : quand l'arbre maximum est élaboré, les critères d'arrêt étant rencontrés, sa construction s'achève
- Il est utile pour discrétiser les variables continues
- Le nb de classes obtenues dépend des seuils fixés pour le test du χ^2

Discrétisation avec CHAID 1/4

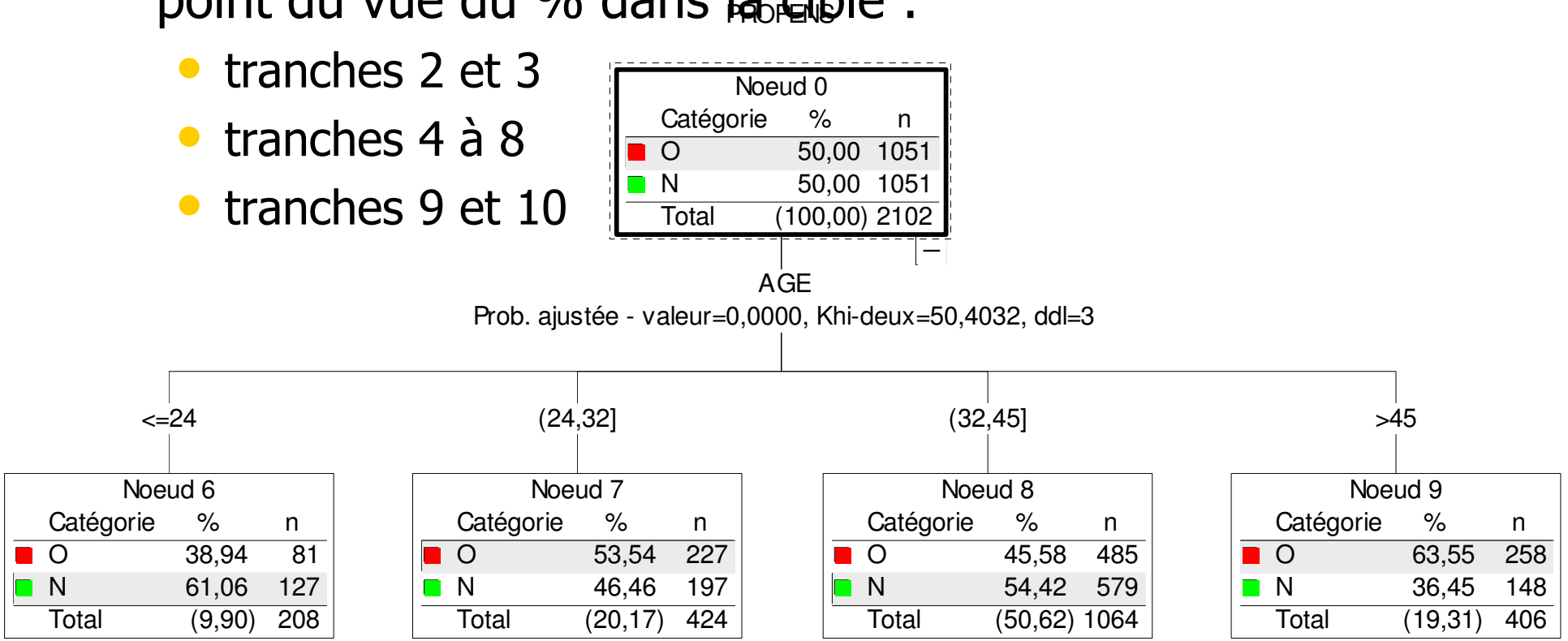
- Supposons que nous voulions prédire une variable cible à l'aide de certaines variables, dont l'âge, et que nous voulions découper l'âge en classes pour les raisons déjà indiquées :
  - prise en compte de la non-monotonie ou non-linéarité de la réponse en fonction de l'âge
  - suppression du problème des extrêmes
  - modèle plus robuste
- Nous allons découper l'âge en 10 tranches (ou plus, si le nb d'individus est grand) et regarder le % d'individus dans la cible pour chaque classe d'âge

Discrétisation avec CHAID 2/4

			cible		
			non	oui	Total
tranche d'âge	18-25 ans	Effectif	127	81	208
		% dans tranche d'âge	61,1%	38,9%	100,0%
	25-29 ans	Effectif	104	126	230
		% dans tranche d'âge	45,2%	54,8%	100,0%
	29-32 ans	Effectif	93	101	194
		% dans tranche d'âge	47,9%	52,1%	100,0%
	32-35 ans	Effectif	113	99	212
		% dans tranche d'âge	53,3%	46,7%	100,0%
	35-38 ans	Effectif	93	94	187
		% dans tranche d'âge	49,7%	50,3%	100,0%
	38-40 ans	Effectif	149	123	272
		% dans tranche d'âge	54,8%	45,2%	100,0%
	40-42 ans	Effectif	108	72	180
		% dans tranche d'âge	60,0%	40,0%	100,0%
	42-45 ans	Effectif	116	97	213
		% dans tranche d'âge	54,5%	45,5%	100,0%
	45-51 ans	Effectif	77	113	190
		% dans tranche d'âge	40,5%	59,5%	100,0%
	> 51 ans	Effectif	71	145	216
		% dans tranche d'âge	32,9%	67,1%	100,0%
Total		Effectif	1051	1051	2102
		% dans tranche d'âge	50,0%	50,0%	100,0%

Discrétisation avec CHAID 3/4

- Nous voyons que certaines classes sont proches du point du vue du % dans la cible :
  - tranches 2 et 3
  - tranches 4 à 8
  - tranches 9 et 10

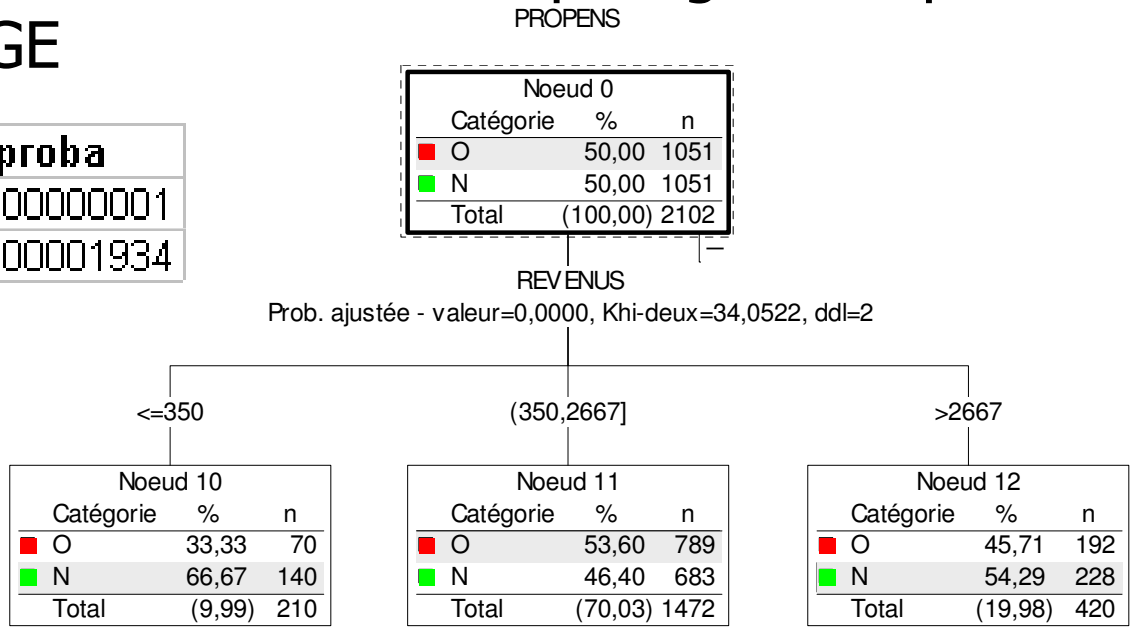


- Nous voyons que CHAID a fait automatiquement ce que nous avons fait manuellement

Discrétisation avec CHAID 4/4

- Pour la scission de la racine de l'arbre, la variable AGE est retenue devant la variable REVENUS car la proba associée au χ^2 des REVENUS est plus grande que celle associée à l'AGE

variable	chi-2	ddl	proba
âge	50,40	3	0,0000000001
revenus	34,05	2	0,0000001934



- NB : si le nb de ddl n'est pas le même pour 2 variables, il faut comparer les probas et non les χ^2 eux-mêmes

Indice de Gini

- Indice de Gini d'un nœud = $1 - \Sigma_i f_i^2$
  - où les f_i, $i = 1$ à p, sont les fréquences relatives dans le nœud des p classes à prédire (variable cible)
  - = probabilité que 2 individus, choisis aléatoirement dans un nœud, appartiennent à 2 classes différentes
- Plus les classes sont uniformément distribuées dans un nœud, plus l'indice de Gini est élevé ; plus le nœud est pur, plus l'indice de Gini est bas
  - Dans le cas de 2 classes, l'indice va de 0 (nœud pur) à 0,5 (mélange maximal). Avec 3 classes, l'indice va de 0 à 2/3.
- Chaque séparation en k nœuds fils (d'effectifs n_1, n_2 ... n_k) doit provoquer la plus grande hausse de la pureté, donc la plus grande baisse de l'indice de Gini. Autrement dit, il faut minimiser :
  - Gini (séparation) = $\sum_{i=1}^{k} \frac{n_k}{n} Gini(k^e noeud)$

Arbre CART 1/2

- Le critère de division est basé sur l'indice de Gini
- Optimal : toutes les scissions possibles sont examinées
- Optimal : élagage supérieur à celui de CHAID
  - une fois l'arbre maximum construit, l'algorithme en déduit plusieurs sous-arbres par élagages successifs, qu'il compare entre eux, avant de retenir celui pour lequel le taux d'erreur mesuré en test est le plus bas possible
- Général : variable cible quantitative ou qualitative

  $\Rightarrow$ CART sert à la prédiction comme au classement
- Général : CART permet la prise en compte de coûts C_{ij} de mauvaise affectation (d'un individu de la classe j dans la classe i) en les intégrant dans le calcul de l'indice de Gini
  - Gini (nœud) = $\sum_{i \neq j} C_{ij} f_i f_j$

Arbre CART 2/2

- Un nœud est considéré comme une feuille lorsque aucune séparation ne conduit à une baisse significative de l'indice de Gini
- Une feuille est affectée à la classe C :
  - la mieux représentée dans la feuille
  - ou la plus probable dans la feuille (si cette probabilité est différente de la proportion – cela dépend du paramétrage)
  - ou la moins coûteuse si des coûts de mauvais classement ont été fixés
- Dans sa version de base, CART est binaire
  - il est moins large que profond, mais parfois trop profond
- Gère les valeurs manquantes en recourant aux variables équidivisantes ou équiréductrices
  - différent de CHAID

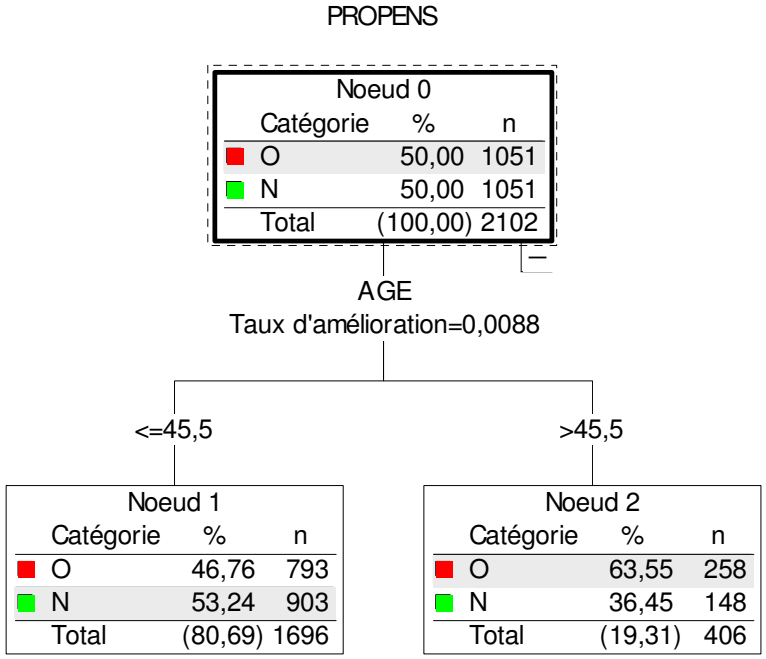
Traitements des valeurs manquantes

- Variables équidivisantes :
  - celles qui assurent (à peu près) la même pureté des nœuds que la variable optimale
- Variables équiréductrices :
  - celles qui répartissent les individus (à peu près) de la même façon que la variable optimale

- Ces variables servent de variables « de rechange » lorsque la variable optimale a une valeur manquante.
  - Par cohérence, il vaut mieux utiliser les variables équiréductrices

Exemple précédent avec CART

- La scission de la racine se fait par l'AGE, comme avec CHAID, mais l'arbre binaire est moins équilibré :



- On peut aussi pénaliser les scissions déséquilibrées
- CART est surtout apte à détecter rapidement des profils très marqués

Mécanisme de scission des nœuds avec Gini (ex : catalogue avec prix article et achat)

Article	Prix	Achat
1	125	N
2	100	N
3	70	N
4	120	N
5	95	O
6	60	N
7	220	N
8	85	O
9	75	N
10	90	O

Mécanisme de scission des nœuds avec Gini

Achat	N		N		N		O		O		O		N		N		N		N		N	
Prix	60		70		75		85		90		95		100		120		125		220			
Seuil	55		65		72		80		87		92		97		110		122		172		230	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
O	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
N	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0,420		0,400		0,375		0,343		0,417		0,400		**0,300**		0,343		0,375		0,400		0,420	

$6/10.(1-0,5^2-0,5^2)+4/10.(1-0^2-1^2)=6/10*0,5=0,3$

CART et complexité du choix (C1)

- Si une **variable explicative qualitative** X a un ensemble E de n valeurs possibles x_1, ..., x_n, toute condition de séparation sur cette variable sera de la forme
  - $X \in E'$, où $E' \subset E - \{0\}$

> 2^{n-1} — 1 conditions de séparation possibles

- Pour une **variable explicative continue** X, la complexité est liée au tri des valeurs x_1, ..., x_n de X, puisqu'une fois les variables dans l'ordre $x_1 \leq ... \leq x_n$,
- il suffit de trouver l'indice k tel que la condition
  - $X \leq$ moyenne (x_k , x_{k+1})
- soit la meilleure (selon le critère choisi, par exemple Gini).

Entropie

- Entropie (ou « information ») d'un nœud = $\Sigma f_i.\log(f_i)$
  - où les f_i, i = 1 à p, sont comme ci-dessus les fréquences relatives dans le nœud des p classes à prédire

- Plus les classes sont uniformément distribuées dans un nœud, plus l'entropie est élevée ; plus le nœud est pur, plus l'entropie est basse (elle vaut 0 lorsque le nœud ne contient qu'une seule classe)

- Comme précédemment, il faut minimiser l'entropie dans les nœuds-fils

Arbre C5.0

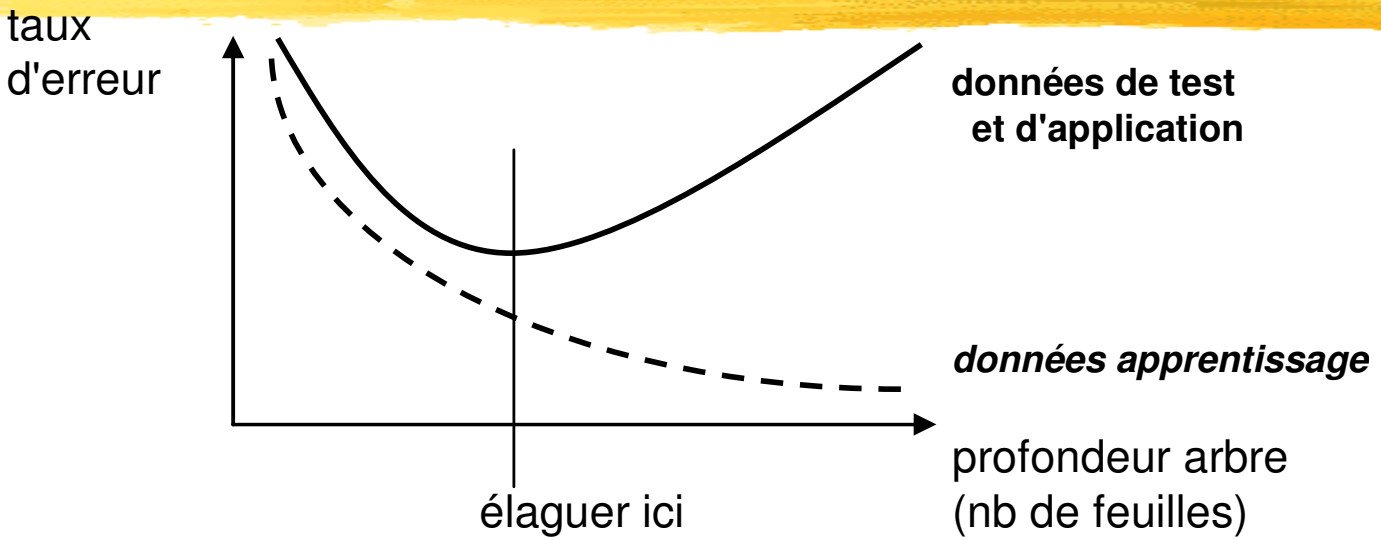
- C5.0 (successeur de C4.5) est adapté comme CART à tout type de variables
- Dispositif d'optimisation de l'arbre par construction puis élagage d'un arbre maximum
  - le procédé d'élagage est différent de celui de CART et il est lié à l'intervalle de confiance du taux d'erreur donc à l'effectif du nœud
- C5.0 cherche à minimiser l'entropie dans les nœuds-fils
- C5.0 n'est pas binaire. Les variables qualitatives, au niveau d'un nœud père, donnent naissance à un nœud fils par modalité
  - inconvénient : les nœuds voient plus rapidement leurs effectifs baisser (moindre fiabilité statistique)

Pré-élagage et Post-élagage

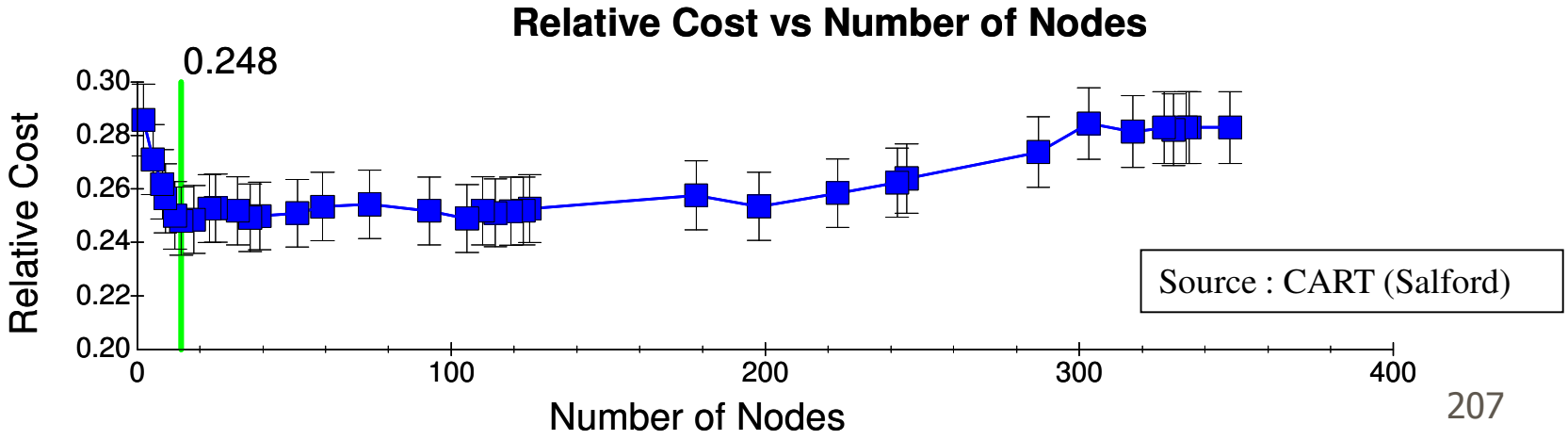
- Certains arbres (CHAID) effectuent un pré-élagage :
  - si un approfondissement d'une branche dégrade la qualité de l'arbre : on s'arrête là

- D'autres arbres (CART, C5.0) effectuent un post-élagage :
  - l'approfondissement de l'arbre est mené à son terme
  - AVANT d'élaguer l'arbre
  - > ce qui est + efficace, car parfois le sur-apprentissage a commencé avant d'être détecté par le taux d'erreur en test
  - > l'arbre peut ainsi découvrir des informations importantes plus profondes que ce que montre un élagage prématuré

Élagage et sur-apprentissage



- Un bon arbre doit être élagué pour éviter la remontée du taux d'erreur due au sur-apprentissage



Validation croisée

- Lorsque la population est trop petite pour en extraire un échantillon d'apprentissage et un de test (courant en pharmacie) :
- On a recours à la **validation croisée** (leave-one-out)
  - La population est scindée en, disons, 10 échantillons de tailles égales, ayant chacun la même distribution pour la classe ou la variable à prédire.
  - On utilise les 9 premiers échantillons comme échantillon d'apprentissage, et le 1/10^{e} restant comme échantillon de test. On obtient ainsi un taux d'erreur en test.
  - On répète ensuite 9 fois la même opération sur chaque 9/10^{e} possible, en prenant chaque 1/10^{e} restant pour échantillon de test.
  - On combine enfin les 10 taux d'erreur obtenus.

Avantages des arbres de décision 1

- Ils fournissent des règles :
  - explicites (contrairement aux réseaux de neurones)
  - qui s'écrivent directement avec les variables d'origine
- Méthode non paramétrique, non perturbée par :
  - la distribution non linéaire ou non monotone des prédicteurs par rapport à la variable cible
  - la colinéarité des prédicteurs
  - les interactions entre les prédicteurs
  - les individus « hors-normes » (isolés dans des règles spécifiques)
  - les fluctuations des prédicteurs non discriminants (l'arbre sélectionne les plus discriminantes)

Avantages des arbres de décision 2

- Beaucoup traitent (sans recodification) des données hétérogènes (numériques et non numériques, voire manquantes)
  - CART traite les valeurs manquantes en remplaçant les variables concernées par des variables équidivisantes
  - CHAID traite l'ensemble des valeurs manquantes d'une variable comme une modalité à part ou pouvant être associée à une autre
  - éviter d'avoir plus de 15 % de valeurs manquantes
- Durée de traitement
  - leur apprentissage peut être un peu long, mais beaucoup moins que pour les réseaux de neurones
  - leur application est très rapide d'exécution

Inconvénients des arbres de décision

- Les nœuds du niveau *n+1* dépendent fortement de ceux du niveau *n*
  - un arbre détecte des optimums locaux et non globaux
  - > la modification d'une seule variable, si elle est placée près du sommet de l'arbre, peut entièrement modifier l'arbre
  - les variables sont testées séquentiellement et non simultanément
  - > manque de robustesse
- L'apprentissage nécessite un nombre suffisant d'individus (pour avoir au moins 30 à 50 individus / nœud)
- Discontinuité de la réponse de la variable cible en fonction des variables explicatives (nb de valeurs du score = nb de feuilles)
- Valeurs du score non uniformément distribuées

Pour améliorer les résultats :

Le rééchantillonnage

Rééchantillonnage Bootstrap

- Pour estimer un paramètre statistique dont on ne connaît pas la loi dans un échantillon de n individus
  - ou quand son calcul exige une distribution normale non vérifiée
- On l'approche par une suite de B (souvent B ≥ 100) tirages aléatoires de *n* individus avec remise
  - en mesurant le paramètre pour chaque échantillon simulé
  - puis en établissant la distribution des fréquences des valeurs de ce paramètre
  - puis en calculant l'intervalle de confiance du paramètre
  - $(2n-1)!/[n!(n-1)!]$ échantillons bootstrap différents
- Inventé par Bradley Efron (1979)

Principe du bootstrap 1/4

- Pb : estimation d'un paramètre statistique défini dans une population globale Ω et fonction d'une loi statistique F
  - ex : la moyenne = E(F)
- Or, la population Ω et la loi F sont généralement inconnues
  - d'autant que la population peut être en évolution perpétuelle ou qu'il peut exister des erreurs de mesure, de saisie…
- Quand nous travaillons sur un jeu de données, il s'agit presque toujours d'un échantillon $S = \{x_1, x_2, \ldots, x_n\}$ tiré d'une population globale Ω inconnue
- et on cherche à approcher le paramètre par un estimateur défini sur S, cet estimateur étant obtenu en remplaçant la loi inconnue F par la loi « empirique », qui est la loi discrète donnant une probabilité $1/n$ à chaque x_i

Principe du bootstrap 2/4

- Cet estimateur est appelé estimateur « plug-in »
- On le note $\hat{\theta} = s(x)$ pour signifier qu'il dépend de l'échantillon
  - ex : $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est un estimateur « plug-in » de la moyenne
- Si F est la loi normale de moyenne μ_F et de d'écart-type σ_F, on connaît la distribution des estimateurs $\hat{\mu}$: elle suit la loi normale de moyenne μ_F et de d'écart-type $\sigma_F / \sqrt{n}$
  - $E(\hat{\mu}) = \mu \Rightarrow$ on dit que $\hat{\mu}$ est un estimateur sans biais.
  - ici, de plus, il est donné par une formule explicite, de même que son écart-type
- Plus généralement se pose la question de la précision et de la robustesse d'un estimateur, i.e. de son biais et de son écart-type, généralement non explicites

Principe du bootstrap 3/4

- Pour calculer l'écart-type de l'estimateur, il faudrait pouvoir déterminer l'estimateur sur un grand nombre d'échantillons S', S"…
- Or, souvent un seul échantillon S nous est donné
- Idée de Bradley Efron (1979) : reproduire le passage de la population Ω à l'échantillon S étudié, en faisant jouer à $S = \{x_1, x_2, \ldots, x_n\}$ le rôle d'une nouvelle population et en obtenant les échantillons souhaités S', S"… par des tirages aléatoires avec remise des n individus $x_1, x_2, \ldots, x_n$
- Échantillon bootstrap = échantillon obtenu par tirage avec remise de n individus parmi n
- Chaque x_i peut être tiré plusieurs fois ou ne pas être tiré. Sa probabilité d'être tiré est $p = 1 - (1 - 1/n)^n$, $p \rightarrow 0{,}632$ $(n \rightarrow +\infty)$

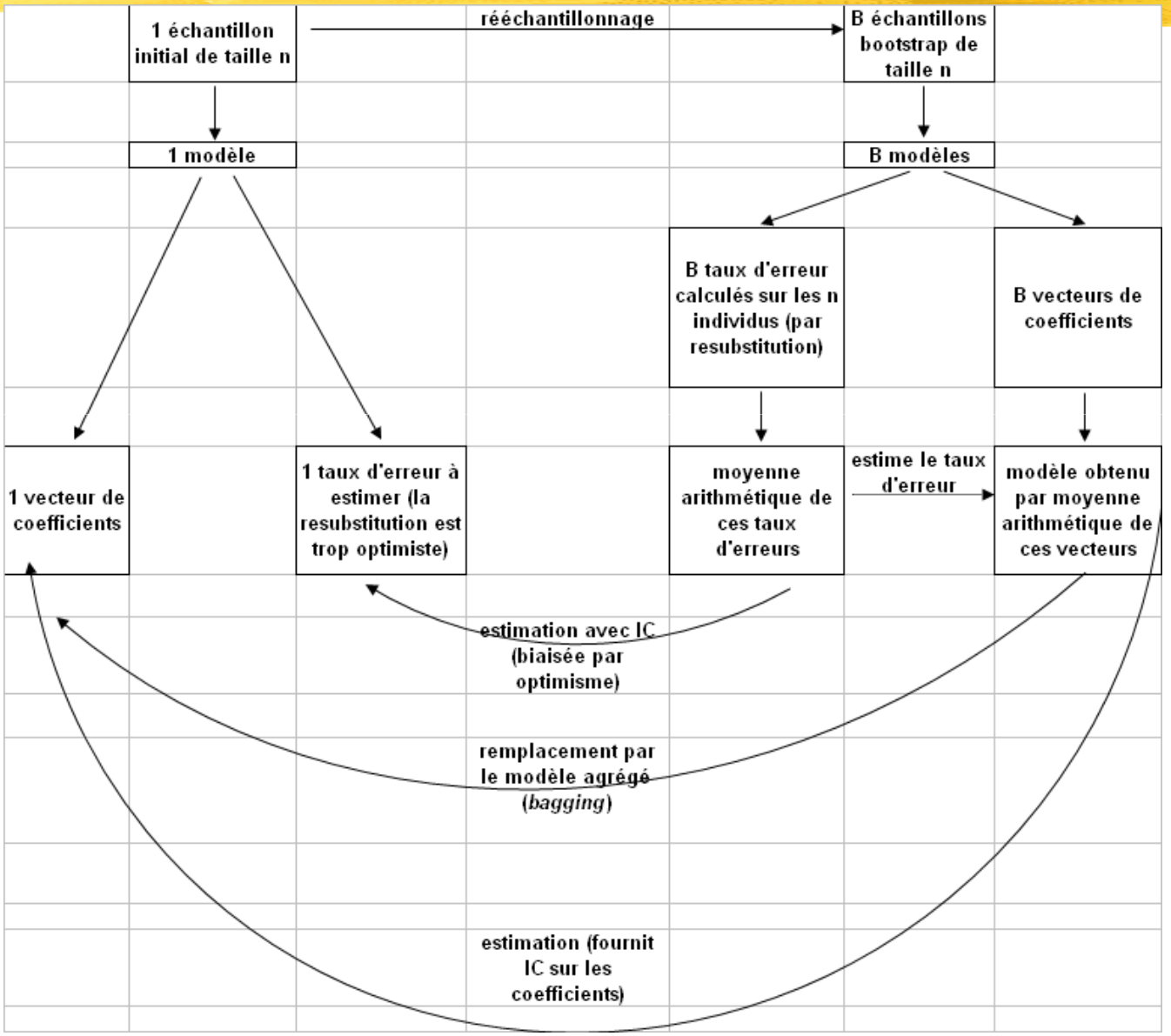
Principe du bootstrap 4/4

- Pour avoir le biais et l'écart-type de l'estimateur d'un paramètre statistique avec Ω et F inconnues
- On tire B (souvent B ≥ 100) échantillons bootstrap
  - on calcule sur chacun d'eux l'estimateur « plug-in »
  - on obtient une distribution des estimateurs « plug-in » centrée autour de la moyenne $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$
  - on déduit un écart-type qui fournit l'approximation recherchée de l'écart-type de l'estimateur
    - on peut déduire un intervalle de confiance $[Q_{2,5}\ ;\ Q_{97,5}]$ à 95 % de l'estimateur en regardant la 25e plus faible valeur $Q_{2,5}$ et la 25e plus forte valeur $Q_{97,5}$ de l'estimateur bootstrap
  - le biais = différence entre l'estimateur calculé sur S et la moyenne des estimateurs bootstrap

Application aux problèmes de scoring

- Les paramètres θ que l'on cherche à estimer sont :
  - le taux d'erreur (ou de bon classement) ou une autre mesure de performance du modèle de score (aire sous la courbe ROC, indice de Gini…)
  - les coefficients de la fonction de score
  - les prédictions (probabilités *a posteriori* d'appartenance à chaque classe à prédire)
- La population globale sur laquelle devrait être construit le modèle est inconnue :
  - on tire B échantillons bootstrap à partir de l'échantillon initial
  - puis on construit un modèle sur chaque échantillon
  - on obtient des intervalles de confiance des indicateurs de performance (ex : aire sous la courbe ROC) du modèle

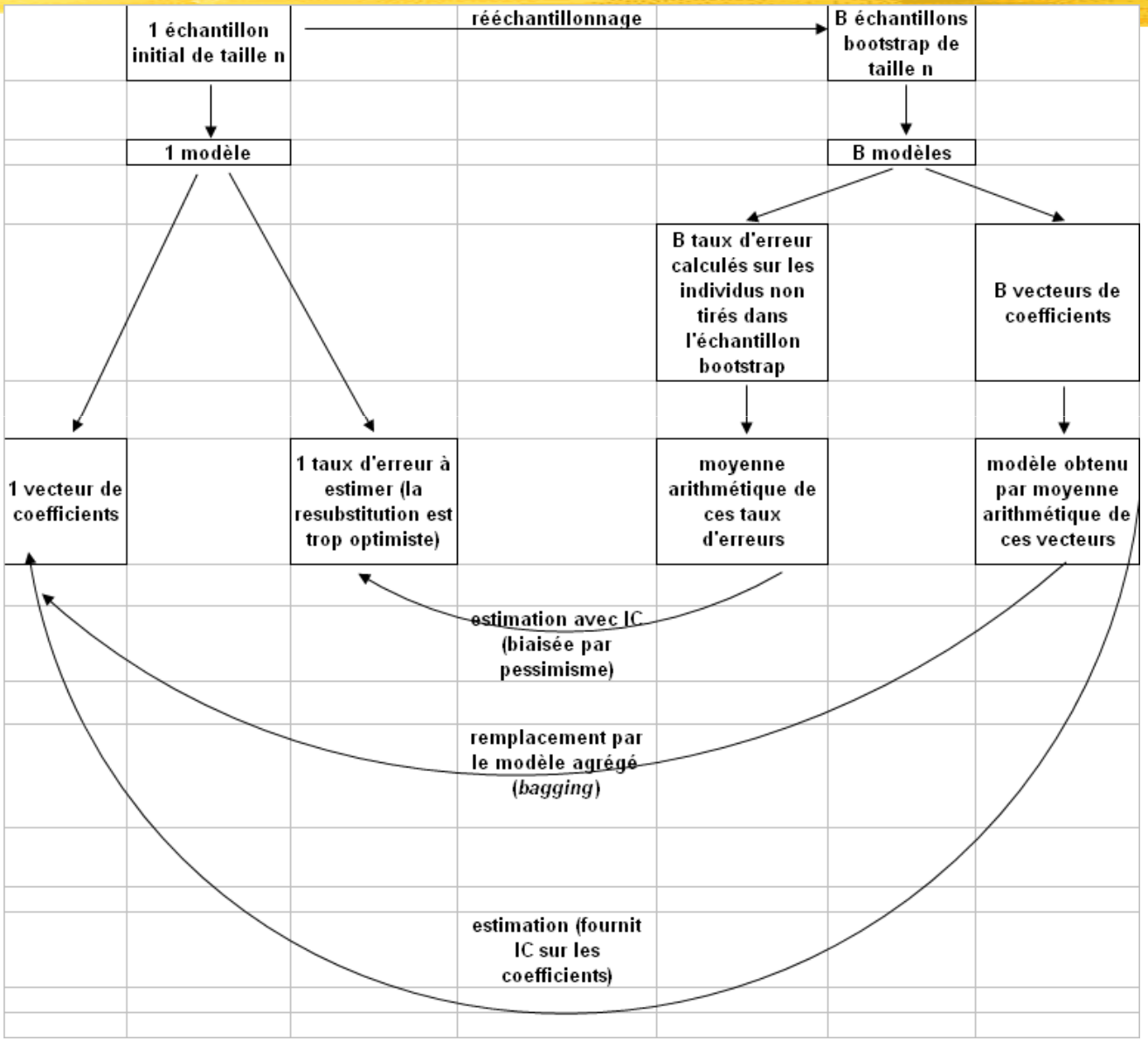
Rééchantillonnage boostrap et bagging



Biais des estimations

- NB : la moyenne des taux d'erreur sur les échantillons bootstrap est une estimation biaisée par optimisme
- Une variante consiste à calculer les erreurs sur les seuls individus n'appartenant pas à l'échantillon bootstrap : c'est l'estimation « out-of-bag »
- Comme cette estimation est cette fois-ci biaisée par pessimisme, Efron et Tibshirani ont proposé de pallier simultanément le biais optimiste de l'estimation de la resubstitution et le biais pessimiste du bootstrap « out-of-bag » par la « formule magique » du « .632-bootstrap » :
- $\text{Estimation}_{.632}$ = 0,368 x estimation(resubstitution) + 0,632 x estimation(bootstrap-oob)

Rééchantillonnage boostrap avec estimation « out-of-bag »



Agrégation de modèles : le bagging

- BAGGING : bootstrap aggregating, Breiman, 1996
- Construction d'une famille de modèles sur n échantillons *bootstrap* (tirages avec remise)
- Ensuite agrégés par un vote ou une moyenne des estimations (ou une moyenne des probabilités en régression logistique)

- FORETS ALEATOIRES, Breiman, 2001
- = Bagging pour les arbres de décision en ajoutant un tirage aléatoire parmi les variables explicatives
- Évite de voir apparaître toujours les mêmes variables
- Efficace sur les souches (« stumps »), arbres à 2 feuilles
  - contrairement au simple bagging

Agrégation de modèles : le boosting

- BOOSTING, Freund et Shapire, 1996
- Version adaptative et généralement déterministe du Bagging :
  - on travaille sur toute la population
  - et à chaque itération, on augmente le poids des individus mal classés dans les itérations précédentes
    - tandis que le poids des bien classés n'augmente pas
- Plusieurs algorithmes : Discrete AdaBoost, Real AdaBoost, LogitBoost, Gentle AdaBoost et ARCING (Adaptative Resampling and Combining)
- Avec CART, le nb de feuilles est à prendre dans [4,8] ou = $\sqrt{p}$, où p = nb de variables explicatives

Différence entre bagging et boosting

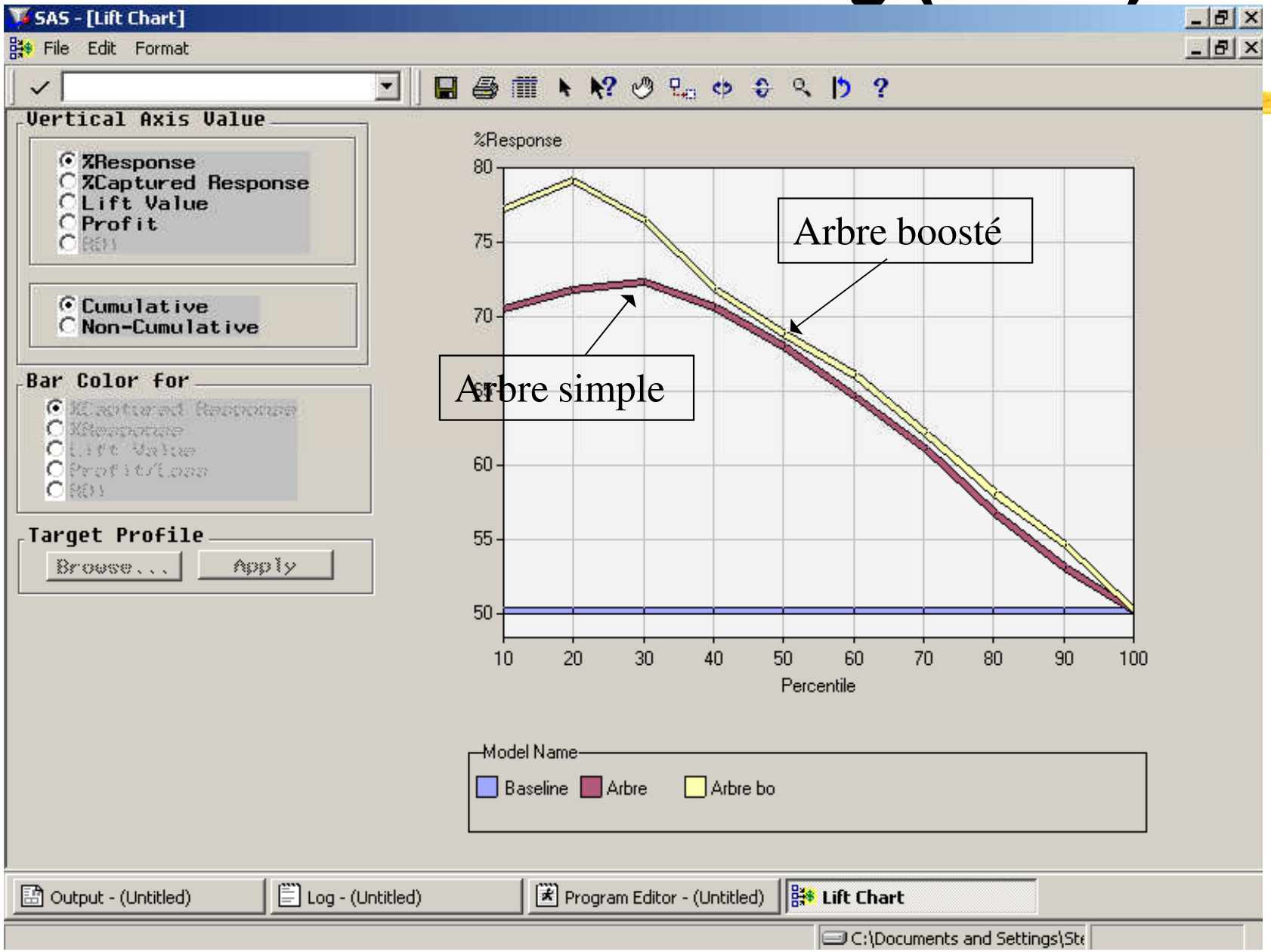
- En boosting :
  - on construit un ensemble de modèles dont on agrège ensuite les prédictions
- Mais :
  - on n'utilise pas nécessairement des échantillons bootstrap mais plus souvent l'échantillon initial complet à chaque itération (sauf dans quelques versions des algorithmes AdaBoost et Arcing)
  - chaque modèle est une version adaptative du précédent, l'adaptation consistant à augmenter le poids des individus précédemment mal classés tandis que le poids des bien classés n'augmente pas
  - l'agrégation finale des modèles est réalisée par une moyenne de tous les modèles dans laquelle chacun est généralement (sauf dans l'algorithme Arcing) pondéré par sa qualité d'ajustement

BAGGING	BOOSTING
Caractéristiques	
Le *bagging* est aléatoire	Le *boosting* est adaptatif et généralement déterministe
On utilise des échantillons bootstrap	On utilise généralement l'échantillon initial complet
Chaque modèle produit doit être performant sur l'ensemble des observations	Chaque modèle produit doit être performant sur certaines observations ; un modèle performant sur certains *outliers* sera moins performant sur les autres individus
Dans l'agrégation, tous les modèles ont le même poids	Dans l'agrégation, les modèles sont généralement pondérés selon leur qualité d'ajustement (sauf l'Arcing)
Avantages et inconvénients	
Technique de réduction de la variance par moyenne de modèles	Peut diminuer la variance et le biais du classifieur de base. Mais la variance peut augmenter avec un classifieur stable
Perte de lisibilité quand le classifieur de base est un arbre de décision	Perte de lisibilité quand le classifieur de base est un arbre de décision
Inopérant sur les « stumps »	Efficace sur les « stumps »
Possibilité de paralléliser l'algorithme	Algorithme séquentiel ne pouvant être parallélisé
Pas de sur-apprentissage : supérieur au *boosting* en présence de « bruit »	Risque de sur-apprentissage mais globalement supérieur au *bagging* sur des données non bruitées (l'Arcing est moins sensible au bruit)
Le *bagging* fonctionne souvent mieux que le *boosting*…	… mais quand le *boosting* fonctionne, il fonctionne mieux

Questions sur le boosting

- Utiliser des échantillons bootstrap ou l'échantillon initial complet ?
- Quelle fonction d'erreur pour pondérer les individus (résidu de la déviance pour un modèle linéaire généralisé) ?
- Faut-il à chaque itération n'utiliser que l'erreur de l'itération précédente, ou la multiplier par l'erreur de toutes les itérations antérieures (risque : « zoomer » excessivement sur les individus outliers mal classés) ?
- Que faire des individus très mal classés à l'itération i : borner leur erreur (ex : limiter à 2 le résidu de la déviance), leur interdire de participer à l'itération i+1, ou ne rien faire ?
- Comment réaliser l'agrégation finale ? Prendre en compte tous les modèles ou écarter ceux qui s'ajustent trop mal ?

Résultat d'un boosting (arbre)



Agrégation de modèles : Conclusion

- Ces techniques permettent d'améliorer parfois très nettement la qualité (tx de biens classés) et la robustesse (sur un autre échantillon) des prédictions
  - même avec seulement une centaine d'itérations
  - mais surtout sur les arbres de décision ! et non sur les classifieurs forts (analyse discriminante ou régression logistique) pour lesquels le gain est faible
- AVANTAGES
  - bonne résistance au bruit
  - bonne résistance au sur-apprentissage
- INCONVÉNIENTS
  - perte de lisibilité
  - importance du temps machine de traitement
- Objet de nombreux travaux théoriques en cours

Combinaison et agrégation de modèles

Appliquer :		Sur :	
		Le même échantillon	**Des échantillons différents**
Quoi :	**La même technique**	Modèle simple	Agrégation de modèles
	Des techniques différentes	Combinaison de modèles	Mélange (*)

(*) Il pourrait s'agir d'une suite d'échantillons bootstrap auxquels seraient chaque fois appliqués un arbre de décision et un réseau de neurones.

Choix d'une méthode de modélisation

Qualités attendues d'une méthode 1/2

- La précision
  - le taux d'erreur doit être le plus bas possible, et l'aire sous la courbe ROC la plus proche possible de 1
- La robustesse
  - être le moins sensible possible aux fluctuations aléatoires de certaines variables et aux valeurs manquantes
  - ne pas dépendre de l'échantillon d'apprentissage utilisé et bien se généraliser à d'autres échantillons
- La concision
  - les règles du modèle doivent être les plus simples et les moins nombreuses possible

Qualités attendues d'une méthode 2/2

- Des résultats explicites
  - les règles du modèle doivent être accessibles et compréhensibles
- La diversité des types de données manipulées
  - toutes les méthodes ne sont pas aptes à traiter les données qualitatives, discrètes, continues et… manquantes
- La rapidité de calcul du modèle
  - un apprentissage trop long limite le nombre d'essais possibles
- Les possibilités de paramétrage
  - dans un classement, il est parfois intéressant de pouvoir pondérer les erreurs de classement, pour signifier, par exemple, qu'il est plus grave de classer un patient malade en « non-malade » que l'inverse

Choix d'une méthode : nature des données

- La régression linéaire traite les variables continues
- L'analyse discriminante traite les variables à expliquer nominales et les variables explicatives continues
- L'analyse discriminante DISQUAL traite les variables à expliquer nominales et les variables explicatives qualitatives
- La régression logistique traite les variables à expliquer qualitatives (nominales ou ordinales) et les variables explicatives continues ou qualitatives
- Les réseaux de neurones traitent les variables continues dans [0,1] et transforment les autres
- Certains arbres de décision (CHAID) traitent nativement les variables discrètes et qualitatives (et transforment les autres)
- CART, C5.0 peuvent aussi traiter les variables continues

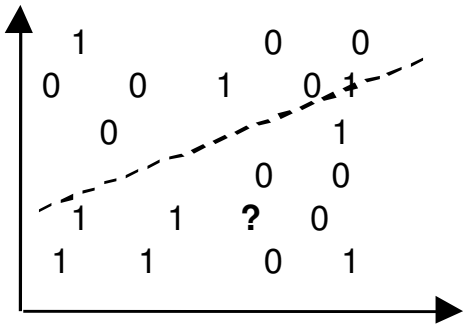
Choix d'une méthode : précision, robustesse, concision, lisibilité

- Précision : privilégier la régression linéaire, l'analyse discriminante et la régression logistique, et parfois les réseaux de neurones en prenant garde au sur-apprentissage (ne pas avoir trop de neurones dans la ou les couches cachées)
- Robustesse : éviter les arbres de décision et se méfier des réseaux de neurones, préférer une régression robuste à une régression linéaire par les moindres carrés
- Concision : privilégier la régression linéaire, l'analyse discriminante et la régression logistique, ainsi que les arbres sans trop de feuilles
- Lisibilité : préférer les arbres de décision et prohiber les réseaux de neurones. La régression logistique, DISQUAL, l'analyse discriminante linéaire et la régression linéaire fournissent aussi des modèles faciles à interpréter

Choix d'une méthode : autres critères

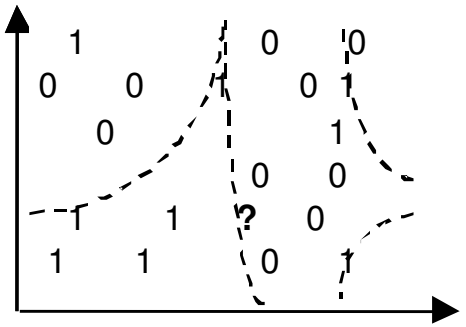
- Peu de données : éviter les arbres de décision et les réseaux de neurones
- Données avec des valeurs manquantes : essayer de recourir à un arbre, à une régression PLS, ou à une régression logistique en codant les valeurs manquantes comme une classe particulière
- Les valeurs extrêmes de variables continues n'affectent pas les arbres de décision, ni la régression logistique et DISQUAL quand les variables continues sont découpées en classes et les extrêmes placés dans 1 ou 2 classes
- Variables explicatives très nombreuses ou très corrélées : utiliser les arbres de décision ou la régression PLS
- Mauvaise compréhension de la structure des données : réseaux de neurones (sinon exploiter la compréhension des données par d'autres types de modèles)

Choix d'une méthode : topographie des classes à discriminer



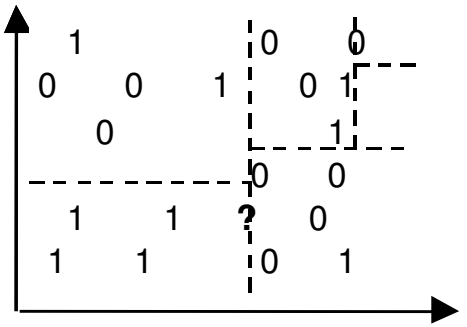
? est classé en "1"

Analyse discriminante



? est classé en "0"

Réseau de neurones



? est classé en "0"

Arbre de décision

- Toutes les méthodes inductives de classement découpent l'espace des variables en régions, dont chacune est associée à une des classes
- La forme de ces régions dépend de la méthode employée

Influence des données et méthodes

- Pour un jeu de données fixé, les écarts entre les performances de différents modèles sont souvent faibles
  - exemple de Gilbert Saporta sur des données d'assurance automobile (on mesure l'aire sous la courbe ROC) :
    - régression logistique : 0,933
    - régression PLS : 0,933
    - analyse discriminante DISQUAL : 0,934
    - analyse discriminante barycentrique : 0,935
  - le choix de la méthode est parfois affaire d'école
- Les performances d'un modèle dépendent :
  - un peu de la technique de modélisation employée
  - beaucoup plus des données !
- D'où l'importance de la phase préliminaire d'exploration et d'analyse des données
  - et même le travail (informatique) de collecte des données

Les 8 principes de base de la modélisation

- La préparation des données est la phase la plus longue, pas la plus passionnante mais la plus importante
- Il faut un nombre suffisant d'observations pour en inférer un modèle
- Validation sur un échantillon de test distinct de celui d'apprentissage (ou validation croisée)
- Arbitrage entre la précision d'un modèle et sa robustesse (« dilemme biais – variance »)
- Limiter le nb de variables explicatives et surtout éviter leur colinéarité
- Perdre de l'information pour en gagner
  - découpage des variables continues en classes
- La performance d'un modèle dépend plus de la qualité des données et du type de problème que de la méthode
  - cf. match « analyse discriminante vs régression logistique »
- On modélise mieux des populations homogènes
  - intérêt d'une classification préalable à la modélisation

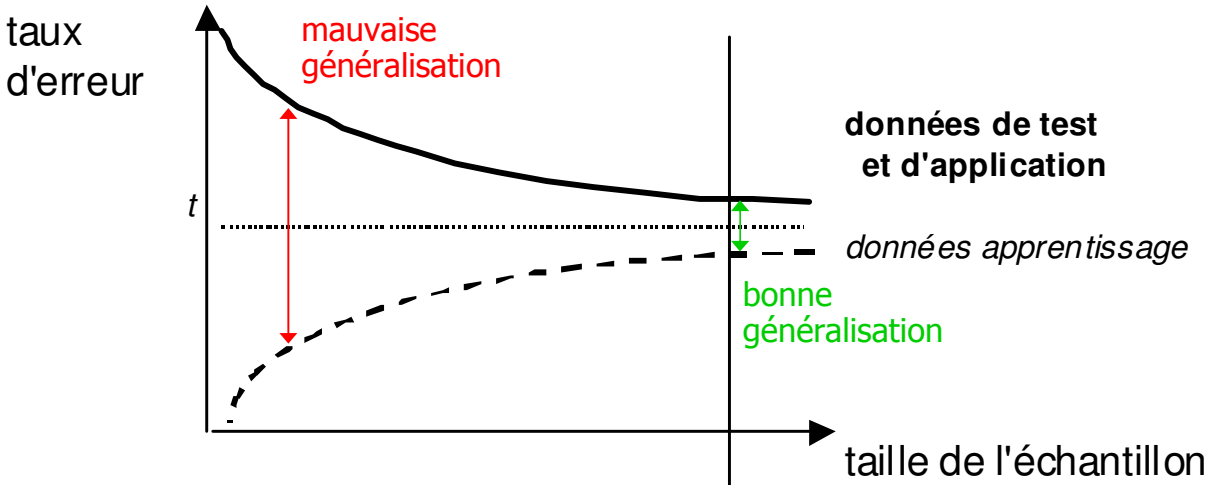
Théorie de l'apprentissage de Vapnik

Fonction de perte et risque d'un modèle

- L'erreur de prédiction d'un modèle se mesure par une fonction de perte :
  - y continue $\Rightarrow L(y,f(x)) = (y - f(x))^2$
  - $y = 0/1 \Rightarrow L(y,f(x)) = ½ |y - f(x)|$
- Risque (ou risque réel) = espérance de la fonction de perte sur l'ensemble des valeurs possibles des données (x,y)
  - comme on ne connaît pas la loi de probabilité conjointe de x et y, on ne peut qu'estimer le risque
  - l'estimation la plus courante est le risque empirique $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ ou $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}|y_i - f(x_i)|$
  - on retrouve le taux d'erreur pour $y = 0/1$ (n = effectif)
- Biais lorsque le risque empirique est mesuré sur l'échantillon d'apprentissage : mieux vaut l'échantillon de test qui approche mieux le risque réel

Risque empirique en apprentissage et test

- Si les courbes de risque empirique sur les données d'apprentissage et de test convergent à partir d'une taille *n* de l'échantillon d'apprentissage, le pouvoir discriminant du modèle se généralisera probablement bien



- Cette convergence a souvent lieu mais pas toujours
- S'il y a convergence, on dit que le modèle est *consistent*

Complexité et VC-dimension

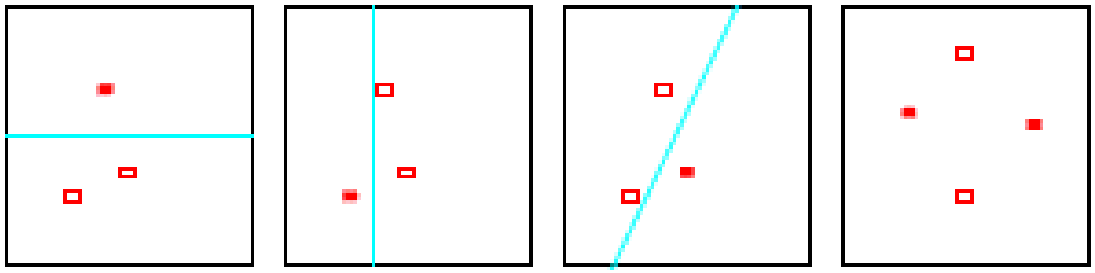
- Plus généralement, Vladimir Vapnik (*The Nature of Statistical Learning Theory*, 1995) s'est intéressé à la convergence du risque empirique sur l'échantillon d'apprentissage vers le risque réel (approché par le risque empirique sur l'échantillon de test)
- Il a démontré deux résultats fondamentaux :
  - sur l'existence d'une convergence
  - sur la vitesse de convergence
- Pour les énoncer, il faut introduire une caractéristique du modèle appelée *dimension de Vapnik-Chernovenkis* (= VC-dimension).
- La VC-dimension est une mesure de complexité d'un modèle
  - définie pour toute famille de fonctions $\mathbf{R}^p \rightarrow \mathbf{R}$ (donc en particulier pour les modèles de classement $\{f(x) \geq 0$, oui ou non$\}$)
  - dont elle mesure le pouvoir séparateur des points de $\mathbf{R}^p$

Hachage de points

- Soit un échantillon de points ($x_1, .. , x_n$) de $\mathbf{R}^p$
- Il existe 2^n différentes manières de séparer cet échantillon en deux sous-échantillons
- Chaque manière correspond à un ensemble (x_1, y_1) , ..., (x_n, y_n) , avec $y_i = +1$ ou -1
- Un ensemble F de fonctions $f(x,\theta)$ « hache » l'échantillon si les 2^n séparations peuvent être faites par des $f(x,\theta) \in F$, c.a.d si on peut toujours trouver θ tel que signe($f(x_i,\theta)$) = y_i pour tout i
- Cela signifie que F peut discriminer n'importe quelle configuration de l'échantillon : problème de classement
- Les droites du plan peuvent « hacher » certains échantillons de trois points (ceux qui sont non alignés) mais aucun échantillon de quatre points

VC-dimension

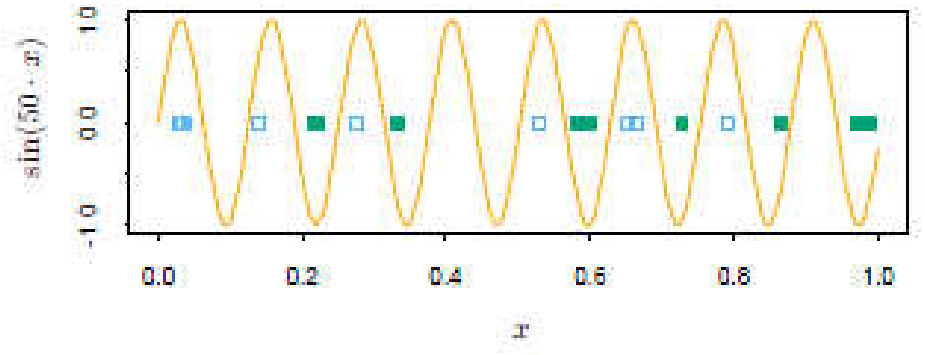
- La VC-dimension de F est le plus grand nombre de points qui peuvent être hachés par les fonctions de F
- Autrement dit, la VC-dimension de F vaut h si :
  - il existe un échantillon $(x_1, .. , x_h)$ de $\mathbf{R}^p$ qui peut être haché
  - aucun échantillon $(x_1, .. , x_{h+1})$ de $\mathbf{R}^p$ ne peut être haché par F
- Cela ne signifie pas que tout échantillon $(x_1, .. , x_h)$ de $\mathbf{R}^p$ puisse être haché (exemple de 3 points alignés dans le plan)
- La VC-dimension des droites du plan vaut 3
- La VC-dim



(source : Hastie – Tibshirani – Friedman)

Exemples de VC-dimension

- La VC-dimension de l'ensemble des hyperplans de $\mathbf{R}^p$ est $p+1$
- Mais la VC-dimension d'une classe de fonctions n'est pas toujours égale à son nombre de paramètres
- Exemple : la VC-dimension de l'ensemble de fonctions
  $f(x,\theta) =$ signe $(sin(\theta.x))$, $x \in [0,1]$,
  avec un paramètre θ, est infinie

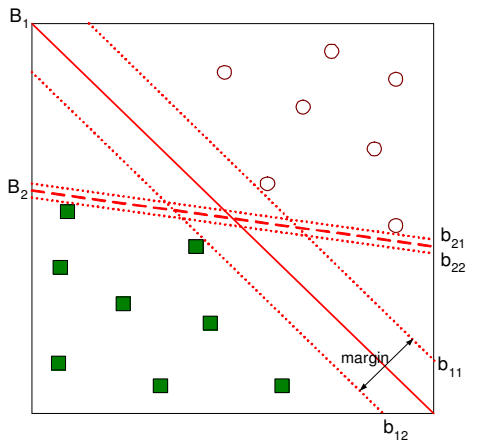


(source : Hastie – Tibshirani – Friedman)

- En effet, si grand soit le nombre de points, on pourra toujours trouver un θ assez grand pour les séparer

Hyperplans avec contrainte

- L'ensemble des hyperplans de $\mathbf{R}^p$ de marge donnée 2M, c'est-à-dire d'équation contrainte par $||\beta|| \leq 1/M$, a une VC-dimension bornée par R^2/M^2 (et bien sûr par p+1) si les observations sont dans une sphère de rayon R
- Plus précisément $h \leq$ min [partie entière $(R^2/M^2), p] + 1$
- Cette formule montre que h n'est pas un majorant défini *a priori* mais qu'il dépend de la configuration des données
- Maximiser la marge $\Rightarrow$ minimiser la VC-dimension
- La marge est le couloir qui sépare les observations. Elle vaut $2/ ||\beta||$ si l'éq. de l'hyperplan est $<\beta.x> + \beta_0$
- Si M > R, visiblement deux points ne peuvent jamais être séparés ($h \leq 1$)



Théorèmes de convergence

- Les deux théorèmes de Vladimir Vapnik :
  - le risque empirique sur l'échantillon d'apprentissage R_{emp} d'un modèle converge vers son risque réel $R \Leftrightarrow$ sa VC-dimension est finie
  - lorsque la VC-dimension h d'un modèle est finie, on a, avec une probabilité d'erreur α :

$$(*) \quad R < R_{emp} + \sqrt{\frac{h\,(\log(2n/h)+1) - \log(\alpha/4)}{n}}$$

- Cette majoration est universelle : elle s'applique à tous les modèles, sans hypothèse sur la loi conjointe de x et y
- La majoration (*) n'est vraie qu'avec une probabilité d'erreur donnée α, et le majorant tend vers l'infini lorsque α tend vers 0

Conséquences

- Le meilleur modèle est celui qui minimise la somme de R_{emp} et de $\sqrt{\frac{h\,(\log(2n/h)+1)-\log(\alpha/4)}{n}}$
- C'est le modèle qui réalise le meilleur compromis entre ajustement et robustesse
- Pour une taille n fixée, lorsque h diminue, généralement R_{emp} augmente et $\sqrt{\frac{h\,(\log(2n/h)+1)-\log(\alpha/4)}{n}}$ diminue $\Rightarrow$ il faut trouver la valeur optimale de h
- Si n augmente, h peut augmenter aussi, car le terme $\sqrt{\frac{h\,(\log(2n/h)+1)-\log(\alpha/4)}{n}}$ tend vers 0 lorsque h/n tend vers 0
- A pouvoir prédictif égal, il privilégier le modèle qui a la plus faible VC-dimension

Cas des modèles avec contrainte

- Dans quelques cas simples, la VC-dimension d'un modèle est égale au nombre de paramètres
- Mais elle est le plus souvent difficile à calculer et même à majorer efficacement, ce qui limite l'intérêt pratique de la majoration (*)
- Les support vector machines (SVM) sont l'un des premiers types de modèles dont il fut possible de calculer la VC-dimension
- Comme la régression régularisée, il s'agit de modèles calculés en appliquant une contrainte $||\beta|| \leq 1/M$
- On a vu qu'en maximisant la marge 2M, on minimise h : cela permet d'assurer et de contrôler le pouvoir de généralisation du modèle
  - la régression ridge est généralement plus robuste que la régression linéaire ordinaire