LES MODELES DE SCORE

Stéphane TUFFERY

CONFERENCE GENDER DIRECTIVE

31 mai 2012

Plan

- Le scoring et ses applications
- L'élaboration d'un modèle de scoring
- La sélection des variables
- La modélisation
- La mesure du pouvoir discriminant
- Conclusion

Le scoring et ses applications

Variété de modèles prédictifs

- Variable à expliquer qualitative
  - souvent binaire, mais peut être polytomique (par exemple : faible, moyen, fort)
  - scores
  - régression logistique, analyse discriminante linéaire, arbres de décision, SVM…
  - chaque individu est affecté à une classe (« risqué » ou « non risqué », par exemple) en fonction de ses caractéristiques
- Variable à expliquer de comptage
  - nombre de sinistres en assurance ou risques opérationnels
  - régression de Poisson
- Variable à expliquer quantitative asymétrique
  - montant des sinistres
  - régression gamma, régression log-normale
- Variable à expliquer quantitative normale
  - modèle linéaire général

Quelques types de scores

- Score d'appétence (ou de propension)
  - prédire l'achat d'un produit ou service
- Score de (comportement) risque
  - prédire les impayés ou la fraude
- Score d'octroi (ou d'acceptation)
  - prédire en temps réel les impayés
- Score d'attrition
  - prédire le départ du client vers un concurrent
- Et aussi :
  - En médecine : diagnostic (bonne santé : oui / non) en fonction du dossier du patient et des analyses médicales
  - Courriels : spam (oui / non) en fonction des caractéristiques du message (fréquence des mots…)

Le scoring dans l'assurance de risque

- Des produits obligatoires (automobile, habitation) :
  - soit prendre un client à un concurrent
  - soit faire monter en gamme un client que l'on détient déjà
- D'où les sujets dominants :
  - attrition
  - ventes croisées (*cross-selling*)
  - montées en gamme (*up-selling*)
- Besoin de décisionnel dû à :
  - concurrence des nouveaux entrants (bancassurance)
  - bases clients des assureurs traditionnels mal organisées :
    - compartimentées par agent général
    - ou structurées par contrat et non par client

Le scoring dans la banque

- Naissance du score de risque en 1941 (David Durand)
  - utilisation de l'analyse discriminante linéaire par David Durand pour modéliser le risque de défaut d'un emprunteur à partir de quelques caractéristiques telles que son âge et son sexe
- Multiples techniques appliquées à la banque de détail et la banque d'entreprise
- Surtout la banque de particuliers :
  - grand nombre de dossiers
  - dossiers relativement standards
  - montants unitaires modérés
- Essor dû à :
  - développement des nouvelles technologies
  - nouvelles attentes de qualité de service des clients
  - pression mondiale pour une plus grande rentabilité
  - surtout : ratio de solvabilité Bâle 2

Quelques problématiques dans l'élaboration d'un modèle de scoring

Définition de la variable à expliquer

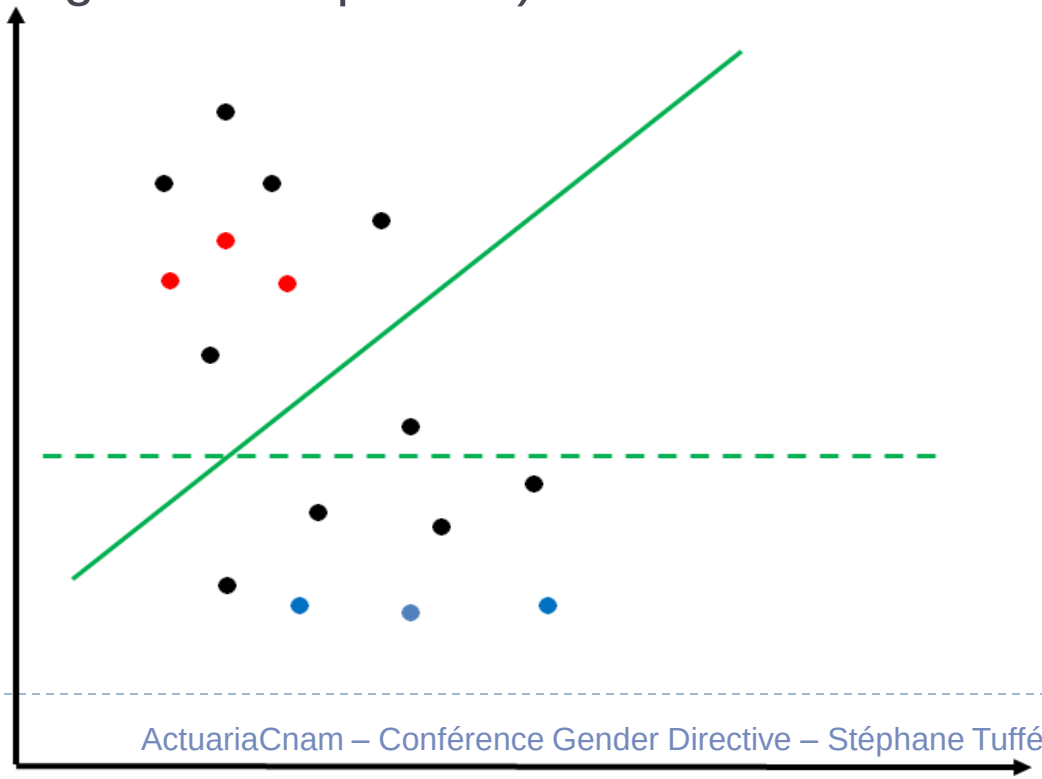
- En médecine : définition souvent naturelle
  - un patient a ou non une tumeur (et encore faut-il distinguer les différents stades d'une tumeur)
- Dans la banque : qu'est-ce qu'un client non risqué ?
  - aucun impayé, 1 impayé, *n* impayés mais dette apurée ?
- Dans certains modèles, on définit une « zone indéterminée » non modélisée :
  - 1 impayé $\Rightarrow$ variable à expliquer non définie
  - aucun impayé $\Rightarrow$ variable à expliquer = 0
  - $\geq$ 2 impayés $\Rightarrow$ variable à expliquer = 1
- Définition parfois encore plus problématique en appétence et en attrition
  - mais dans l'assurance, contrairement à la banque, le départ est toujours net (mais le client ne résilie pas tous ses contrats simultanément)

Biais de sélection

- En risque : certaines demandes sont refusées et on ne peut donc pas mesurer la variable à expliquer
  - certaines populations ont été exclues de la modélisation et on leur applique pourtant le modèle
  - il existe des méthodes « d'inférence des refusés », mais dont aucune n'est totalement satisfaisante
  - et parfois aucune trace n'est conservée des demandes refusées !
- En appétence : certaines populations n'ont jamais été ciblées et on ne leur a pas proposé le produit
  - si on les modélise, elles seront présentes dans l'échantillon des « mauvais » (clients sans appétence) peut-être à tort
  - contrairement au cas précédent, on peut mesurer la variable à expliquer car il y a des souscriptions spontanées
  - envisager de limiter le périmètre aux clients ciblés

Inférence des refusés

- Plusieurs méthodes ajoutent une étape préliminaire d'étiquetage des refusés :
  - en fonction des acceptés qui ont un score proche (ex : parcelling)
  - en fonction des acceptés qui ont un profil proche (ex : apprentissage semi-supervisé)



Construction de la base d'analyse

n° client	variable cible : acheteur (O/N)	âge	PCS	situation famille	nb achats	montant achats	…	variable explicative *m*	échantillon
1	O	58	cadre	marié	2	40	…	…	apprentissage
2	N	27	ouvrier	célibataire	3	30	…	…	test
…	…	…		…	…	…	…	…	…
…	…	…		…	…	…	…	…	…
k	O	46	technicien	célibataire	3	75	…	…	test
…	…	…		…	…	…	…	…	…
…	…	…		…	…	…	…	…	…
1000	N	32	employé	marié	1	50	…	…	apprentissage
…	…	…		…	…	…		…	…

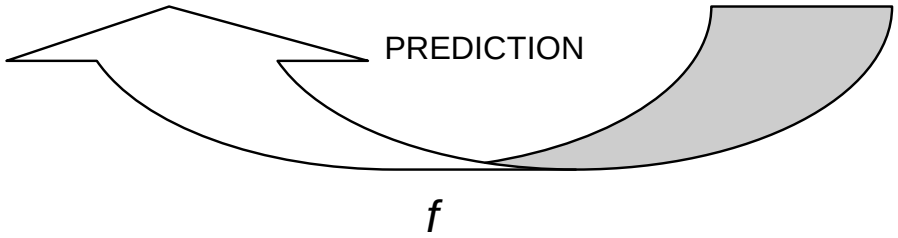
au moins 1000 cas

variable à expliquer observée année n

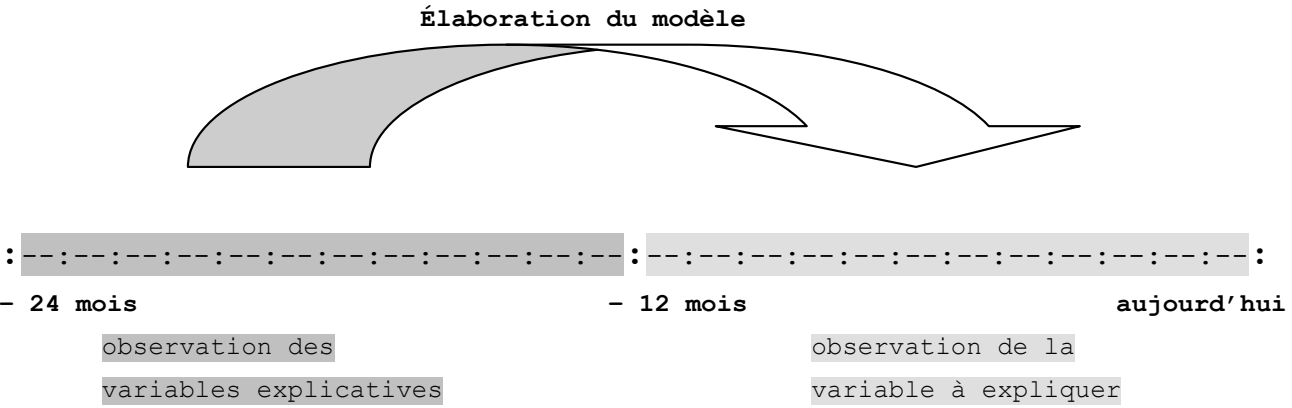
variables explicatives observées année n-1

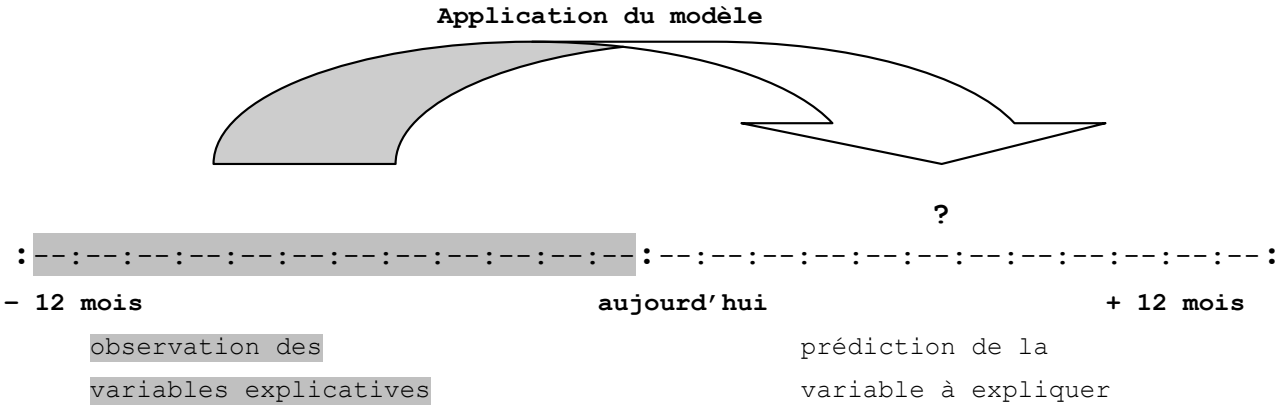
répartition aléatoire des clients entre les 2 échantillons

O : au moins 500 clients ciblés dans l'année *n* et acheteurs
N : au moins 500 clients ciblés dans l'année *n* et non acheteurs



Sélection des périodes d'observation



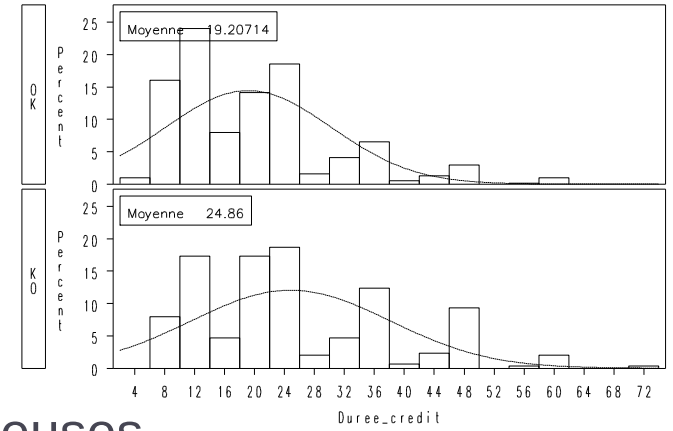


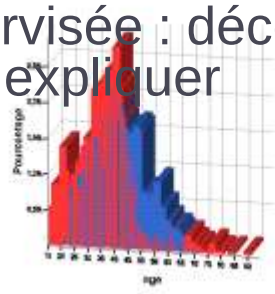
Le **modèle** sera par exemple une fonction *f* telle que :

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

Analyse exploratoire des données 1/2

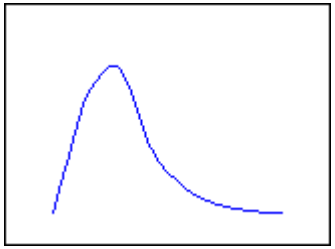
- Explorer la distribution des variables
- Vérifier la fiabilité des variables
  - valeurs incohérentes ou manquantes
    - suppression ou imputation ou isolement
  - valeurs extrêmes
    - voir si valeurs aberrantes à éliminer
  - certaines variables sont fiables mais trompeuses
    A. le profil de souscripteurs peut être faussé par une campagne commerciale ciblée récente
- Variables continues
  - détecter la non-monotonie ou la non-linéarité justifiant la discrétisation
  - tester la normalité des variables (surtout si petits effectifs) et les transformer pour augmenter la normalité
  - éventuellement effectuer une discrétisation supervisée : découper la variable en tranches en fonction de la variable à expliquer
    - et isoler les valeurs manquantes ou aberrantes

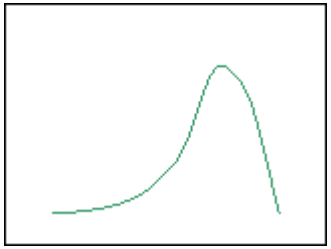




Normalisation : transformations

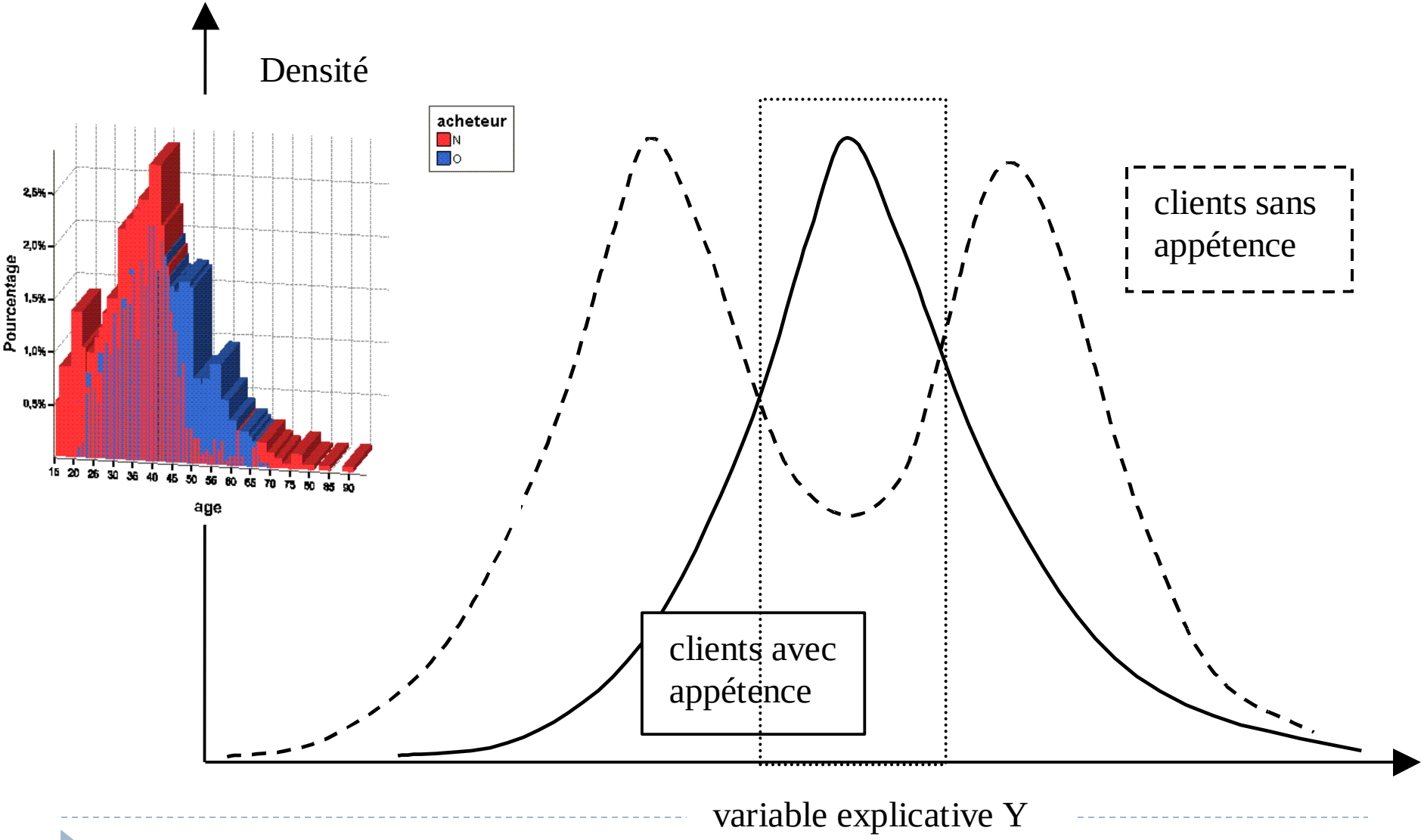
- Log (V)
  - transformation la plus courante pour corriger un coefficient d'asymétrie > 0
  - Si V ≥ 0, on prend Log (1 + V)
- Racine carrée (V) si coefficient d'asymétrie > 0
- -1/V ou $-1/V^2$ si coefficient d'asymétrie > 0
- V^2 ou V^3 si coefficient d'asymétrie < 0
- Arc sinus (racine carrée de V/100)
  - si V est un pourcentage compris entre 0 et 100
- La normalisation peut améliorer certains modèles prédictifs en raison de leurs hypothèses (analyse discriminante linéaire)





Transformation	exp(V)	V^3	V^2	V	√V	log(V)	-1/V	$-1/V^2$
Correction	asymétrie à gauche			pas de correction	asymétrie à droite			
Effet	fort ← moyen				moyen → fort			

Discrétisation en tranches naturelles



Pourquoi discrétiser ?

- Appréhender des liaisons non linéaires (de degré >1), voire non monotones, entre les variables continues et la variable à expliquer
  - par une ACM, une régression logistique ou une analyse discriminante DISQUAL
- Neutraliser les valeurs extrêmes (« outliers »)
  - qui sont dans la 1ère et la dernière tranches
- Gérer les valeurs manquantes (imputation toujours délicate)
  - rassemblées dans une tranche supplémentaire spécifique
- Gérer les ratios dont le numérateur et le dénominateur peuvent être tous deux > 0 ou < 0
  - EBE / capital économique (rentabilité économique), résultat net / capitaux propres (rentabilité financière ou ROE)
- Traiter simultanément des données quantitatives et qualitatives

Analyse exploratoire des données 2/2

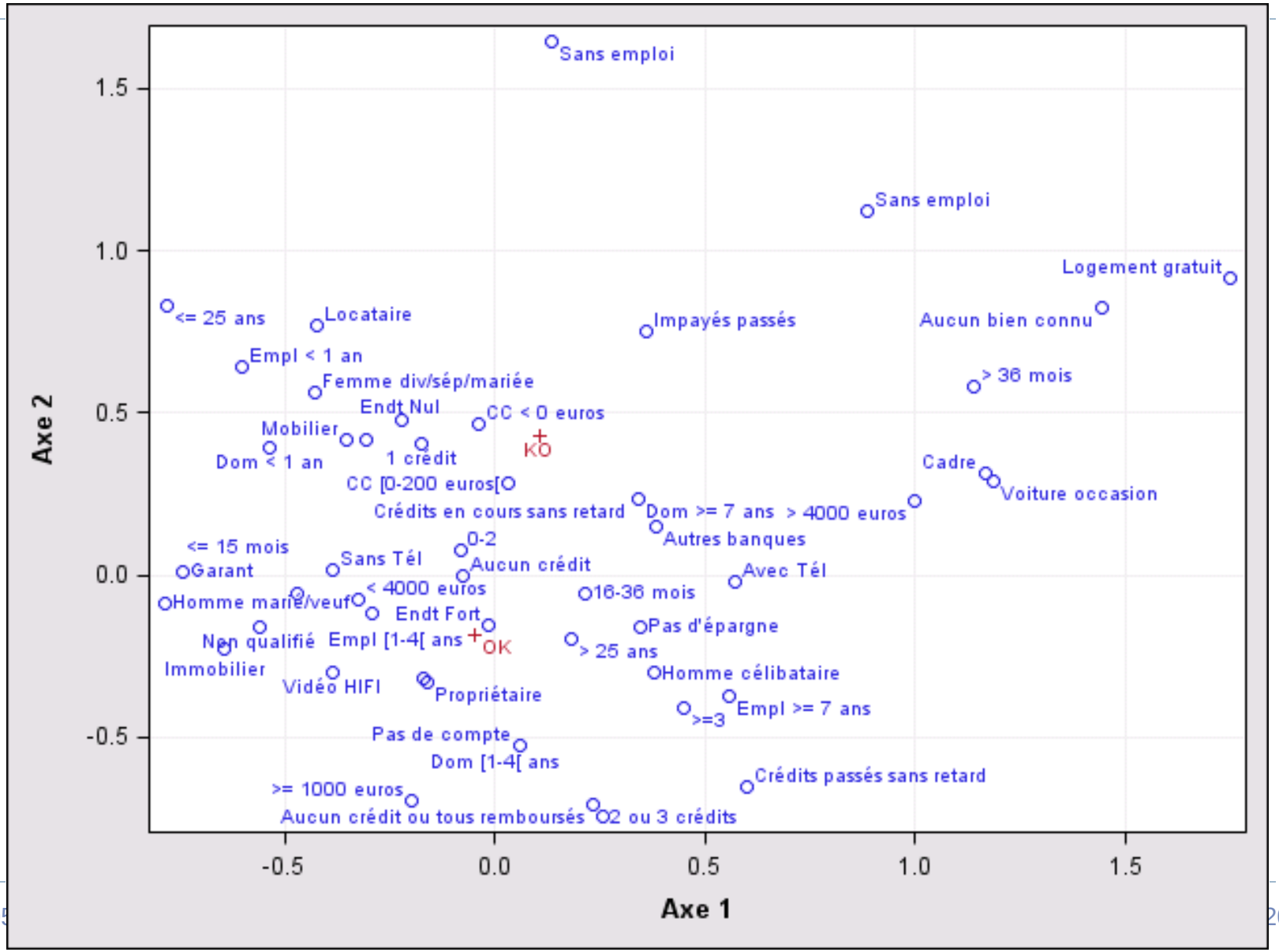
- Variables qualitatives ou discrètes
  - regrouper certaines modalités aux effectifs trop petits
  - représenter les modalités dans une analyse des correspondances multiples
- Créer des indicateurs pertinents d'après les données brutes
  - prendre l'avis des spécialistes du secteur étudié
  - création d'indicateurs pertinents (maxima, moyennes, présence/absence…)
  - utiliser des ratios plutôt que des variables absolues
  - calcul d'évolutions temporelles de variables
  - croisement de variables, interactions
  - utilisation de coordonnées factorielles
- Détecter les liaisons entre variables
  - entre variables explicatives et à expliquer (bon)
  - entre variables explicatives entre elles (colinéarité à éviter dans certaines méthodes)

Exemple de regroupement de modalités

- Le regroupement des modalités « Locataire » et « Logement gratuit » est évident
- Elles sont associées à des taux d'impayés proches et élevés (39,11% et 40,74%)
- Les propriétaires sont moins risqués, surtout s'ils ont fini leur emprunt, mais pas seulement dans ce cas, car ils sont généralement plus attentifs que la moyenne au bon remboursement de leur emprunt

Table de Statut_domicile par Cible			
Statut_domicile	**Cible**		
FREQUENCE **Pourcentage** **Pct en ligne**	**OK**	**KO**	**Total**
Locataire	109 10.90 60.89	70 7.00 39.11	179 17.90
Propriétaire	527 52.70 73.91	186 18.60 26.09	713 71.30
Logement gratuit	64 6.40 59.26	44 4.40 40.74	108 10.80
Total	700 70.00	300 30.00	1000 100.00

Exploration avec une ACM

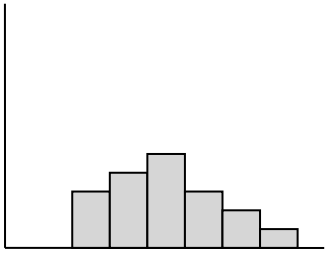


Traitement des valeurs manquantes

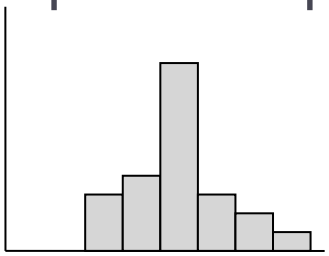
- D'abord vérifier que les valeurs manquantes ne proviennent pas :
  - d'un problème technique dans la constitution de la base
  - d'individus qui ne devraient pas se trouver dans la base
- Sinon, plusieurs solutions sont envisageables selon les cas :
  - supprimer les observations (si elles sont peu nombreuses ou si le non renseignement de la variable est grave et peut laisser suspecter d'autres anomalies dans l'observation)
  - ne pas utiliser la variable concernée (surtout si elle est peu discriminante) ou la remplacer par une variable proche mais sans valeur manquante
    - mieux vaut supprimer une variable *a priori* peu utile, mais qui est souvent non renseignée et conduirait à exclure de nombreuses observations de la modélisation
  - traiter la valeur manquante comme une valeur à part entière
  - imputation : remplacer la valeur manquante par une valeur par défaut ou déduite des valeurs des autres variables
  - remplacer les valeurs manquantes grâce à une source externe (rarement possible)
- Mais aucune solution n'est idéale ☹

Imputation des valeurs manquantes

- Imputation statistique
  - par le mode, la moyenne ou la médiane
  - par une régression ou un arbre de décision
  - imputation
    - simple (minore la variabilité et les intervalles de confiance des paramètres estimés)
    - ou multiple (remplacer chaque valeur manquante par *n* valeurs, puis faire les analyses sur les *n* tables et combiner les résultats pour obtenir les paramètres avec leurs écart-types
- Mais l'imputation n'est jamais neutre
  - Surtout si les données ne sont pas manquantes au hasard



avant imputation



après imputation par la moyenne

La sélection des variables

Problèmes posés par l'absence d'une variable

- Diminution du pouvoir prédictif du modèle
- Coefficients du modèle moins interprétables
- Hétéroscédasticité des résidus dans une régression linéaire
- Sur-dispersion dans un modèle linéaire généralisé
  - sous-estimation de la variance des estimateurs
- Paradoxe de Simpson
- Peut-on se passer d'une variable (le genre) pour la tarification mais l'utiliser pour le calcul des risques ?

Le paradoxe de Simpson : exemple

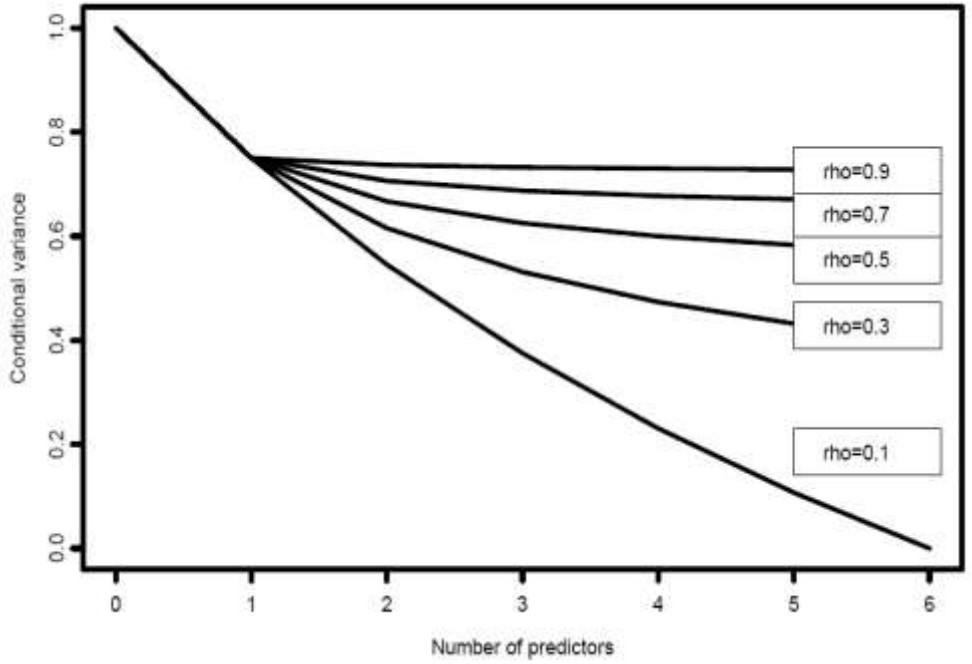
Hommes				
	sans achat	*avec achat*	*TOTAL*	*taux d'achat*
courriel	950	50	1 000	5,00%
téléphone	475	25	500	5,00%
TOTAL	1 425	75	1 500	5,00%
Femmes				
	sans achat	*avec achat*	*TOTAL*	*taux d'achat*
courriel	450	50	500	10,00%
téléphone	900	100	1 000	10,00%
TOTAL	1 350	150	1 500	10,00%
Tous clients				
	sans achat	*avec achat*	*TOTAL*	*taux d'achat*
courriel	1 400	100	1 500	6,67%
téléphone	1 375	125	1 500	8,33%
TOTAL	2 775	225	3 000	7,50%

Le paradoxe de Simpson : explication

- Dans le dernier exemple :
  - les hommes ne répondent pas mieux au téléphone qu'au courriel
  - de même pour les femmes
  - et pourtant, le téléphone semble avoir globalement un meilleur taux d'achat
- Explication :
  - un individu pris au hasard ne répond pas mieux au téléphone
  - mais les femmes achètent plus et on a privilégié le téléphone pour les contacter
    - liaison entre les variables « sexe » et « canal de vente »
- Autre exemple publié dans le *Wall-Street Journal* du 2/12/2009 :
  - le taux de chômage est globalement plus faible en octobre 2009 (10,2 %) qu'en novembre 1982 (10,8 %)
  - et pourtant, ce taux de chômage est plus élevé en 2009 à la fois pour les diplômés et pour les non-diplômés !
  - l'explication est l'existence d'une liaison entre l'année et le niveau d'étude : le niveau moyen d'étude est plus élevé en 2009, et le taux de chômage est plus faible chez ceux dont le niveau d'étude est plus élevé

Importance de la sélection des variables

- Exemple de David Hand (2005) : régression avec un coefficient de corrélation 0,5 entre chaque prédicteur et la variable à expliquer, et un coefficient de corrélation ρ entre chaque prédicteur
  - Les courbes représentent 1-R^2 (proportion de la somme des carrés non expliquée) en fonction du nb de prédicteurs



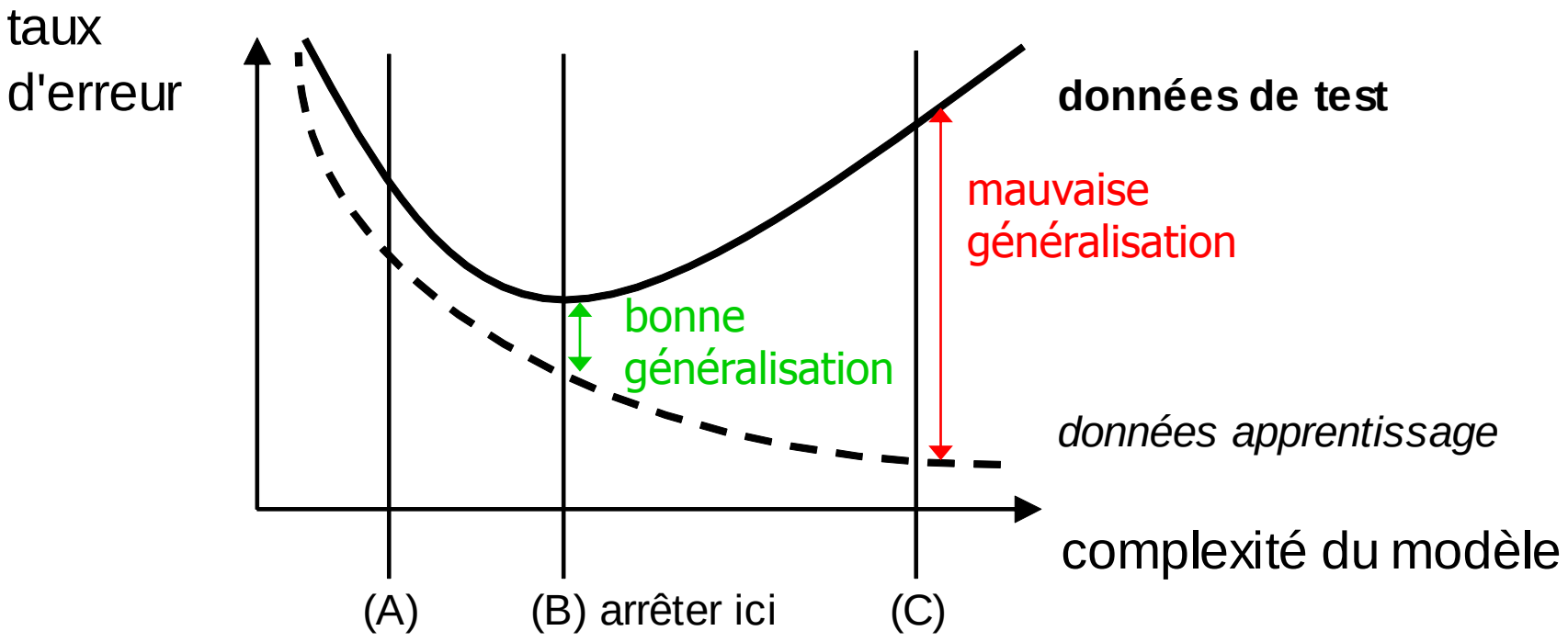
Problèmes posés par des variables corrélées

- En présence de corrélation entre les prédicteurs, l'apport marginal de chaque prédicteur décroît très vite
- Il peut même altérer le modèle (inversions de signes des paramètres) et réduire son pouvoir prédictif
- Augmentation de la variance des estimateurs
- Et pourtant, dans l'exemple de D. Hand, chaque prédicteur est supposé avoir la même liaison avec la cible, ce qui n'est pas le cas dans une sélection pas à pas réelle où la liaison décroît !
- On préfère souvent un prédicteur moins lié à la variable à expliquer s'il est moins corrélé aux autres prédicteurs
- Effectuer des tests statistiques de liaison
- Un test non-paramétrique (ex : Kruskal-Wallis) peut être plus apte qu'un test paramétrique (ex : ANOVA) à détecter les variables les plus pertinentes
  - notamment les variables hautement non normales que sont les ratios X/Y
- Limiter le nombre de prédicteurs
  - c'est plus facile si la population est homogène

Solutions au problème de variables corrélées

- Suppression des variables concernées
- Création d'une variable synthétique combinant et remplaçant les variables concernées (par exemple, le ratio de deux variables)
- Transformation (logarithme…) des variables concernées
- Régression avec régularisation (ridge, lasso…)
  - choix fin du paramètre de régularisation à l'aide du ridge plot ⇒ évolution continue des coefficients, contrairement à la PLS
- Régression sur composantes principales (passer ensuite de coefficients de régression des composantes principales à des coefficients sur les variables initiales
- Régression PLS (Partial Least Squares)
  - avec une seule composante : les signes de la régression sont toujours cohérents avec le sens des corrélations

Problèmes posés par des variables trop nombreuses : un modèle trop complexe

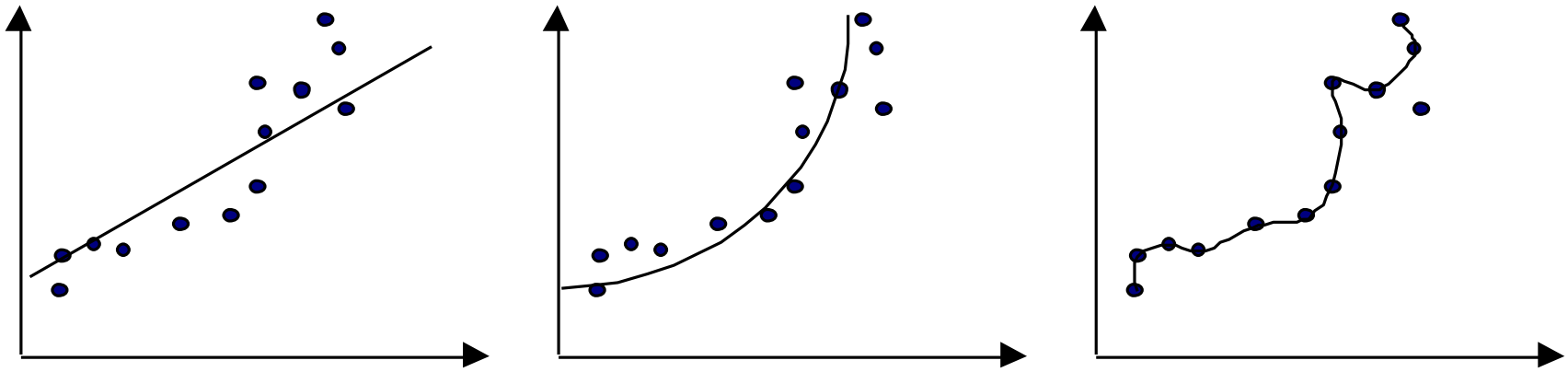


Théorème de Vapnik :

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

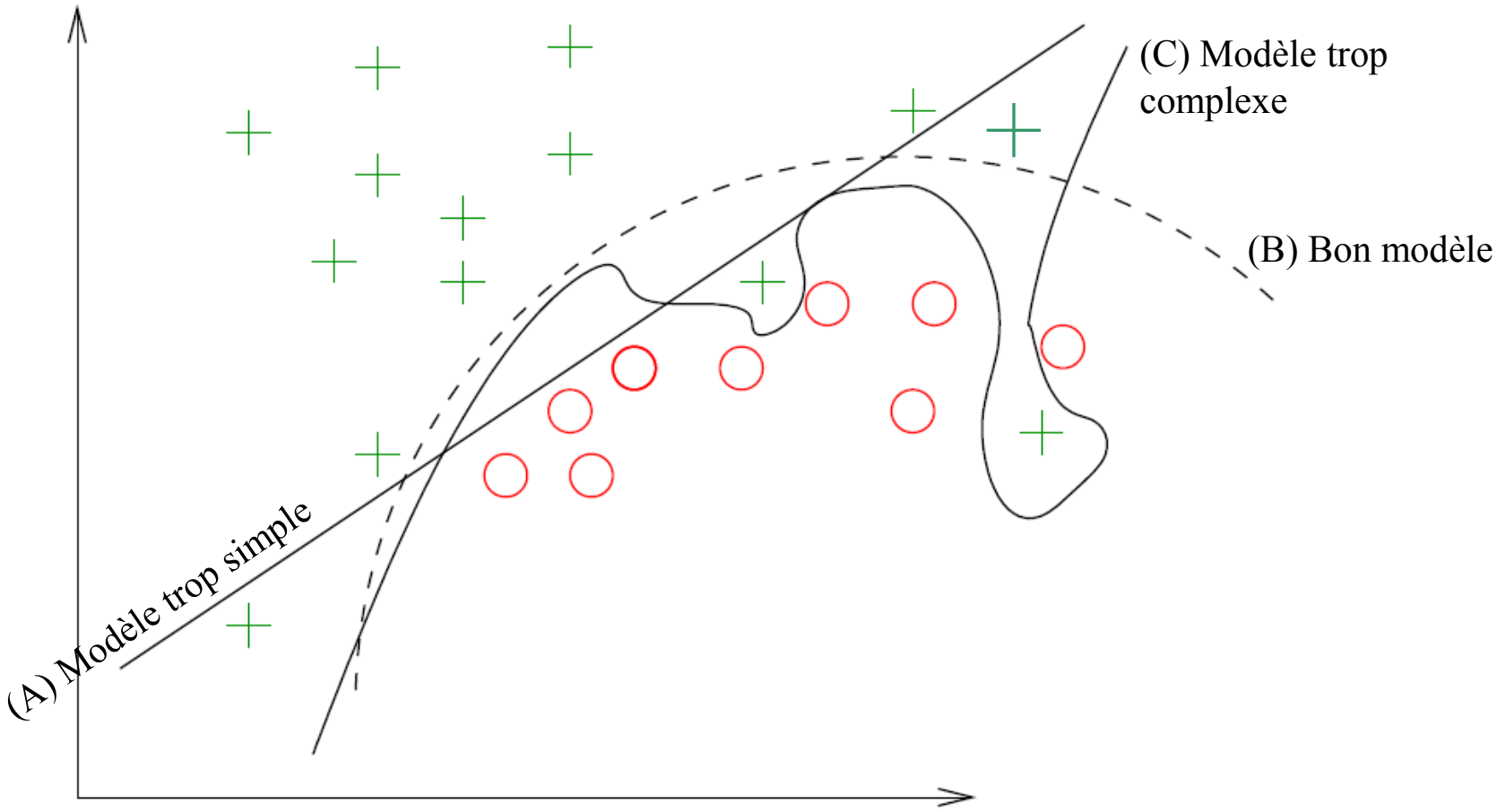
- Noter que le modèle peut être plus complexe si les données sont plus nombreuses

Sur-apprentissage en régression



- 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
- On parle de sur-apprentissage ou sur-ajustement

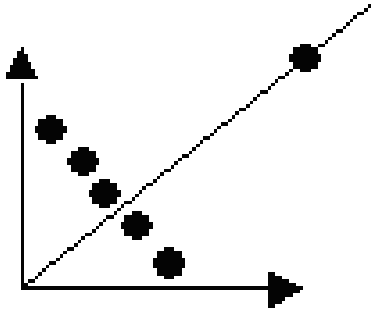
Sur-apprentissage en classement



Source : Olivier Bousquet

Rappel sur les tests

- Tests paramétriques
  - supposent que les variables suivent une loi particulière (normalité, homoscédasticité)
  - ex : test de Student, ANOVA
- Tests non-paramétriques
  - ne supposent pas que les variables suivent une loi particulière
  - se fondent souvent sur les rangs des valeurs des variables plutôt que sur les valeurs elles-mêmes
  - peu sensibles aux valeurs aberrantes
  - ex : test de Wilcoxon-Mann-Whitney, test de Kruskal-Wallis
- Exemple du *r* de Pearson et du ρ de Spearman :
  - $r > \rho \Rightarrow$ présence de valeurs extrêmes ?
  - $\rho > r \Rightarrow$ liaison non linéaire non détectée par Pearson ?
    - ex : x = 1, 2, 3… et y = e^1, e^2, e^3…



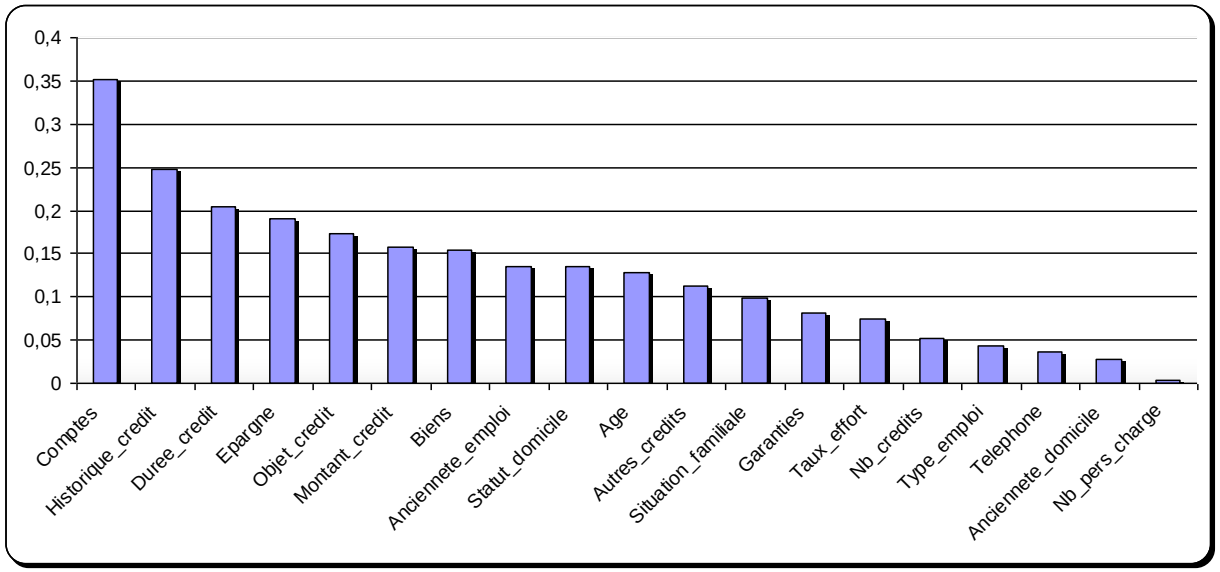
Liaison entre une variable continue et une variable de classe

lois suivies	**2 échantillons**	**3 échantillons et plus (***)**
normalité – homoscédasticité (*)	test T de Student	ANOVA
normalité – hétéroscédasticité	test T de Welch	Welch - ANOVA
non normalité – hétéroscédasticité (**)	Wilcoxon – Mann – Whitney	Kruskal – Wallis
non normalité – hétéroscédasticité (**)	test de la médiane	test de la médiane
non normalité – hétéroscédasticité (**)		test de Jonckheere-Terpstra (échantillons ordonnés)

(*) Ces tests supportent mieux la non-normalité que l'hétéroscédasticité.
(**) Ces tests travaillant sur les rangs et non sur les valeurs elles-mêmes, ils sont plus robustes et s'appliquent également à des variables ordinales
(***) ne pas comparer toutes les paires par des tests T ⇒ on détecte à tort des différences significatives (au seuil de 95 % : dans 27 % des cas pour 4 moyennes égales)

Exemple de liste des variables

- Liste des variables par liaison décroissante avec la variable à expliquer
- Ici les variables sont qualitatives et la liaison mesurée par le V de Cramer



Obs	V_Cramer	Variable
1	0.35174	Comptes
2	0.24838	Historique_credit
3	0.20499	Duree_credit
4	0.19000	Epargne
5	0.17354	Objet_credit
6	0.15809	Montant_credit
7	0.15401	Biens
8	0.13553	Anciennete_emploi
9	0.13491	Statut_domicile
10	0.12794	Age
11	0.11331	Autres_credits
12	0.09801	Situation_familiale
13	0.08152	Garanties
14	0.07401	Taux_effort
15	0.05168	Nb_credits
16	0.04342	Type_emploi
17	0.03647	Telephone
18	0.02737	Anciennete_domicile
19	0.00301	Nb_pers_charge

Pourquoi le V de Cramer ?

	Classe 1	Classe 2	Ensemble
Effectifs observés :			
A	55	45	100
B	20	30	50
Total	75	75	150
Effectifs attendus si la variable est indépendante de la classe :			
A	50	50	100
B	25	25	50
Total	75	75	150
Probabilité du χ^2 = 0,08326454			

	Classe 1	Classe 2	Ensemble
Effectifs observés :			
A	550	450	1000
B	200	300	500
Total	750	750	1500
Effectifs attendus si la variable est indépendante de la classe :			
A	500	500	1000
B	250	250	500
Total	750	750	1500
Probabilité du χ^2 = $4{,}3205.10^{-8}$			

- Quand la taille de la population augmente, le moindre écart devient significatif aux seuils usuels

Le V de Cramer

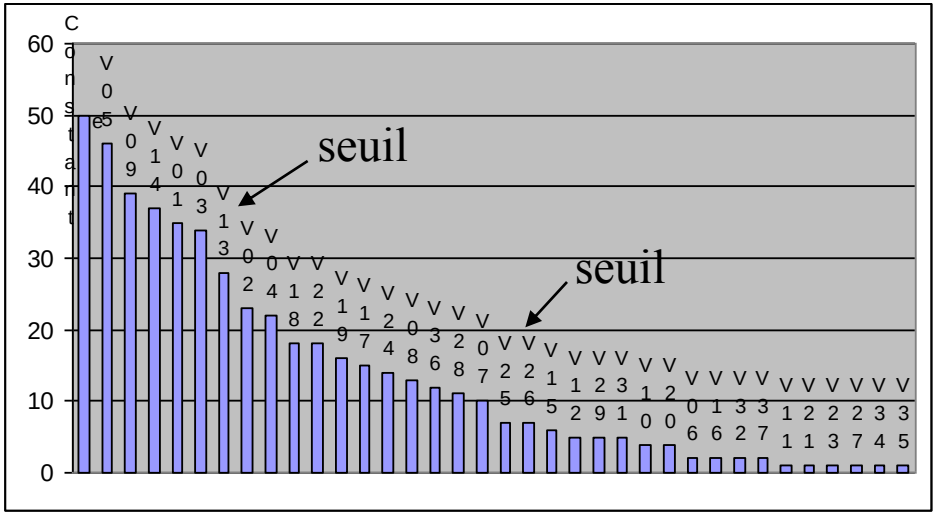
- V de Cramer = $\sqrt{\frac{\chi^2}{\chi^2_{max}}}$
  - mesure directement l'intensité de la liaison de 2 variables qualitatives, sans avoir recours à une table du χ^2
  - en intégrant l'effectif et le nombre de degrés de liberté, par l'intermédiaire de χ^2_{max}
  - χ^2_{max} = effectif x [min (nb lignes, nb colonnes) – 1]
  - V compris entre 0 (liaison nulle) et 1 (liaison parfaite)

Sélection des variables : bootstrap

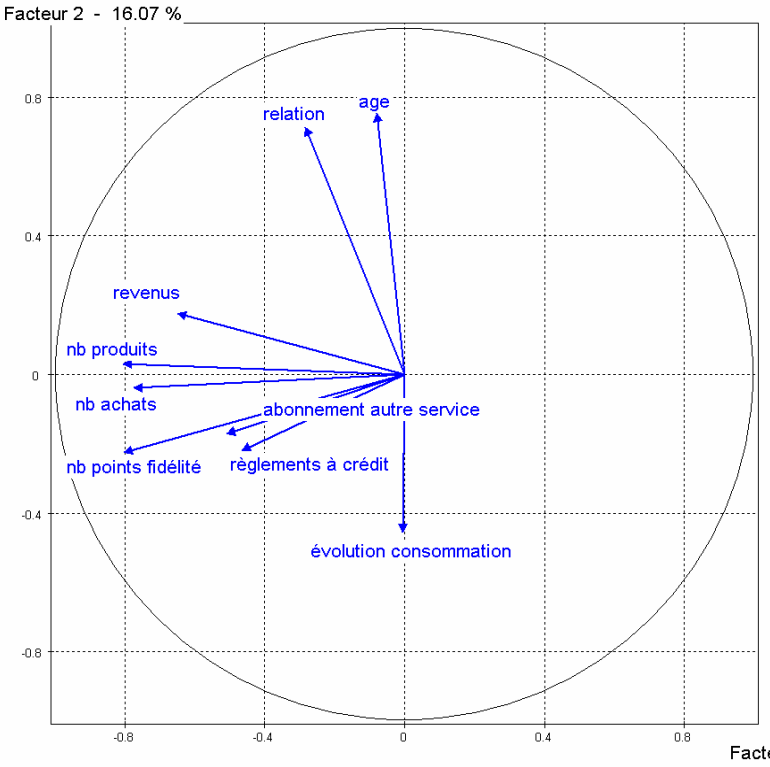
On effectue une régression logistique stepwise sur chacun des échantillons bootstrap

Variable	Nb occurrences		Variable	Nb occurrences
Constante	50		V25	7
V05	46		V26	7
V09	39		V15	6
V14	37		V12	5
V01	35		V29	5
V03	34		V31	5
V13	28		V10	4
V02	23		V20	4
V04	22		V06	2
V18	18		V16	2
V22	18		V32	2
V19	16		V37	2
V17	15		V11	1
V24	14		V21	1
V08	13		V23	1
V36	12		V27	1
V28	11		V34	1
V07	10		V35	1

Bootstrap : B tirages aléatoires avec remise de *n* individus parmi *n* et sélection de variables sur chacun des B échantillons bootstrap



Sélection des variables : classification à l'aide d'une ACP avec rotation

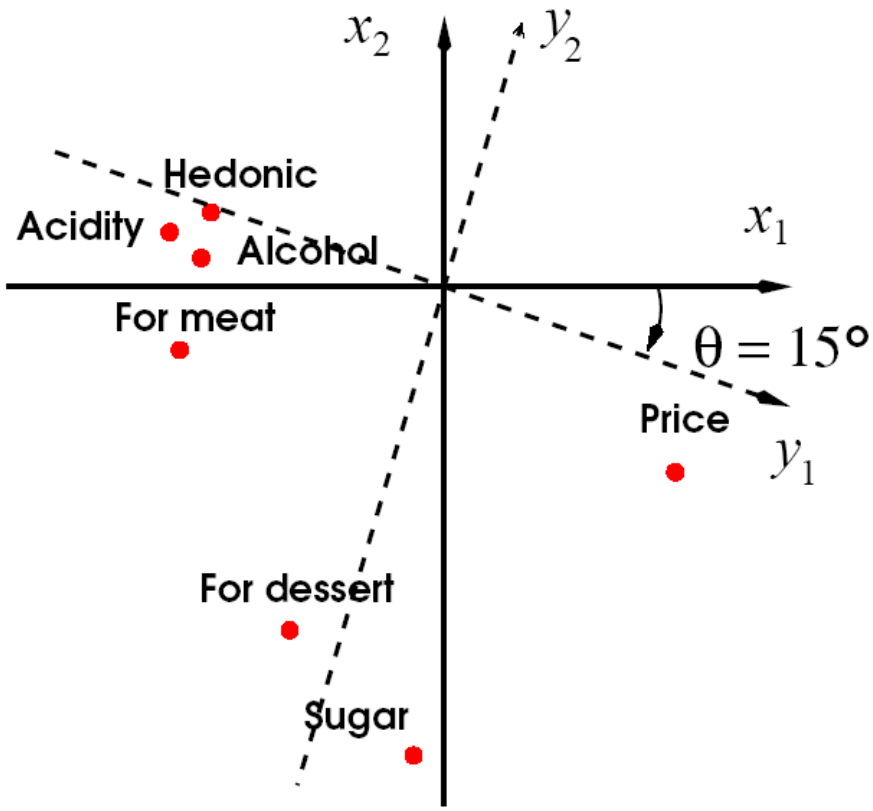


Cluster	Variable	R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	nbpoints	0.6546	0.0011	0.3458	nb points fidélité
	nbproduits	0.6189	0.0183	0.3882	nb produits
	nbachats	0.5950	0.0007	0.4053	nb achats
	revenus	0.4551	0.0234	0.5580	revenus du client
	abonnement	0.2537	0.0042	0.7495	abonnement autre service
	utilcredit	0.2312	0.0002	0.7689	règlements à crédit
Cluster 2	age	0.6033	0.0000	0.3967	âge
	relation	0.6461	0.0336	0.3662	relation (ancienneté client)
	evolconsom	0.2151	0.0027	0.7870	évolution consommation

```
PROC VARCLUS DATA=fichier_client;
VAR age relation nbpoints nbproduits nbachats revenus abonnement evolconsom
utilcredit;
RUN;
```

ACP avec rotation pour la classification de variables

- La procédure VARCLUS effectue une ACP avec rotation (quartimax) qui maximise la variance des coefficients de corrélation de chaque variable avec l'ensemble des facteurs
- Au lieu de maximiser la somme des carrés des coefficients de corrélation
- De façon à minimiser le nombre de facteurs nécessaires pour expliquer chaque variable



La modélisation

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 la courbe *ROC*), on recommence l'apprentissage
- **Validation du modèle** sur un 3e échantillon, éventuellement « out of time », pour avoir une idée du taux d'erreur non biaisé du modèle
- **Application du modèle** à l'ensemble de la population

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

Quelques méthodes de scoring

- Analyse discriminante linéaire
  - Résultat explicite $P(Y/ X_1, \dots , X_p)$ sous forme d'une formule
  - Requiert des X_i continues et des lois X_i/Y multinormales et homoscédastiques (attention aux « outliers »)
  - Optimale si les hypothèses sont remplies
- Régression logistique
  - Sans hypothèse sur les lois X_i/Y, X_i peut être discret, nécessaire absence de colinéarité entre les X_i
  - Méthode très souvent performante
  - Méthode la plus utilisée en scoring
- Arbres de décision
  - Règles complètement explicites
  - Traitent les données hétérogènes, éventuellement manquantes, sans hypothèses de distribution
  - Détection d'interactions et de phénomènes non linéaires
  - Mais moindre robustesse

Grille de score

- Passage de coefficients (« Estimation ») à des pondérations dont la somme est comprise entre 0 et 100

Analyse des estimations de la vraisemblance maximum						
Paramètre		**DF**	**Estimation**	**Erreur std**	**Khi 2 de Wald**	**Pr > Khi 2**
Intercept		1	-3.1995	0.3967	65.0626	<.0001
Comptes	**CC >= 200 euros**	1	1.0772	0.4254	6.4109	0.0113
Comptes	**CC < 0 euros**	1	2.0129	0.2730	54.3578	<.0001
Comptes	**CC [0-200 euros[**	1	1.5001	0.2690	31.1067	<.0001
Comptes	**Pas de compte**	0	0	.	.	.
Historique_credit	**Crédits en impayé**	1	1.0794	0.3710	8.4629	0.0036
Historique_credit	**Crédits sans retard**	1	0.4519	0.2385	3.5888	0.0582
Historique_credit	**Jamais aucun crédit**	0	0	.	.	.
Duree_credit	**> 36 mois**	1	1.4424	0.3479	17.1937	<.0001
Duree_credit	**16-36 mois**	1	1.0232	0.2197	21.6955	<.0001
Duree_credit	**<= 15 mois**	0	0	.	.	.
Age	**<= 25 ans**	1	0.6288	0.2454	6.5675	0.0104
Age	**> 25 ans**	0	0	.	.	.
Epargne	**< 500 euros**	1	0.6415	0.2366	7.3501	0.0067
Epargne	**pas épargne ou > 500 euros**	0	0	.	.	.
Garanties	**Avec garant**	1	-1.7210	0.5598	9.4522	0.0021
Garanties	**Sans garant**	0	0	.	.	.
Autres_credits	**Aucun crédit extérieur**	1	-0.5359	0.2439	4.8276	0.0280
Autres_credits	**Crédits extérieurs**	0	0	.	.	.

Variable	Modalité	Nb points
Age	> 25 ans	0
Age	≤ 25 ans	8
Autres_credits	Aucun crédit extérieur	0
Autres_credits	Crédits extérieurs	7
Comptes	Pas de compte	0
Comptes	CC ≥ 200 euros	13
Comptes	CC [0-200 euros[	19
Comptes	CC < 0 euros	25
Duree_credit	≤ 15 mois	0
Duree_credit	16-36 mois	13
Duree_credit	> 36 mois	18
Epargne	pas épargne ou > 500 euros	0
Epargne	< 500 euros	8
Garanties	Avec garant	0
Garanties	Sans garant	21
Historique_credit	Jamais aucun crédit	0
Historique_credit	Crédits sans retard	6
Historique_credit	Crédits en impayé	13

Exemples de notations

- Note d'un jeune de moins de 25 ans, qui demande pour la première fois un crédit dans l'établissement et qui n'en a pas ailleurs, sans impayé, avec un compte dont le solde moyen est légèrement positif (mais < 200 €), avec un peu d'épargne (< 500 €), sans garant, qui demande un crédit sur 36 mois :
  - 8 + 0 + 19 + 13 + 8 + 21 + 0 = 69 points
- Note d'un demandeur de plus de 25 ans, avec des crédits à la concurrence, sans impayé, avec un compte dont le solde moyen est > 200 €, avec plus de 500 € d'épargne, sans garant, qui demande un crédit sur 12 mois :
  - 0 + 7 + 13 + 0 + 0 + 21 + 0 = 41 points
- On constate la facilité de l'implémentation et du calcul du score mais on n'a pas encore tout à fait un outil d'aide à la décision

Découpage de la note de score

- On peut calculer les déciles du nombre de points et leurs taux d'impayés correspondants :

Analysis Variable : nbpoints			
Rang pour la variable nbpoints	N Obs	Minimum	Maximum
0	104	6.0000000	29.0000000
1	95	33.0000000	37.0000000
2	107	39.0000000	42.0000000
3	120	43.0000000	48.0000000
4	98	49.0000000	54.0000000
5	93	55.0000000	60.0000000
6	81	61.0000000	65.0000000
7	104	66.0000000	69.0000000
8	92	70.0000000	74.0000000
9	106	75.0000000	95.0000000

Seuils de taux

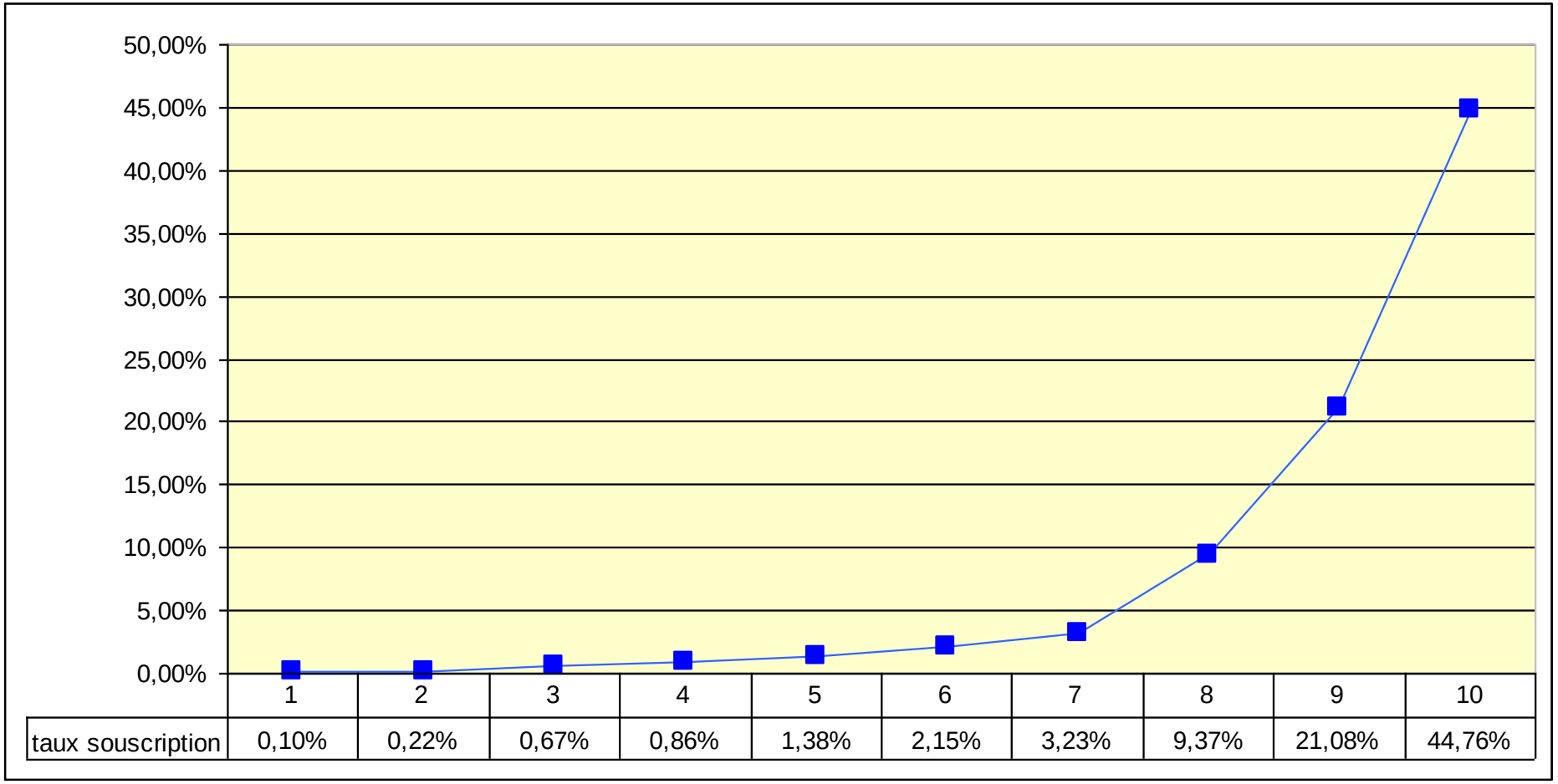
Table de dnbpoints par Cible			
dnbpoints(Rang pour la variable nbpoints)	Cible		
FREQUENCE Pct en ligne	OK	KO	Total
0	99 95.19	5 4.81	104
1	89 93.68	6 6.32	95
2	100 93.46	7 6.54	107
3	101 84.17	19 15.83	120
4	71 72.45	27 27.55	98
5	60 64.52	33 35.48	93
6	48 59.26	33 40.74	81
7	60 57.69	44 42.31	104
8	38 41.30	54 58.70	92
9	34 32.08	72 67.92	106
Total	700	300	1000

Taux d'impayés par tranches de score

Table de nbpoints par Cible			
nbpoints	**Cible**		
FREQUENCE **Pourcentage** **Pct en ligne**	**OK**	**KO**	**Total**
risque faible [0 , 48] points	389 38.90 91.31	37 3.70 8.69	426 42.60
risque moyen [49 , 69] points	239 23.90 63.56	137 13.70 36.44	376 37.60
risque fort ≥ 70 points	72 7.20 36.36	126 12.60 63.64	198 19.80
Total	700 70.00	300 30.00	1000 100.00

- Tranche de risque faible :
  - 8,69% d'impayés
  - octroi du crédit avec un minimum de formalités
- Tranche de risque moyen :
  - 36,44% d'impayés
  - octroi du crédit selon la procédure standard
- Tranche de risque élevé :
  - 63,64% d'impayés
  - octroi du crédit interdit sauf par l'échelon hiérarchique supérieur (directeur d'agence)

Les résultats du modèle retenu



	1	2	3	4	5	6	7	8	9	10
taux souscription	0,10%	0,22%	0,67%	0,86%	1,38%	2,15%	3,23%	9,37%	21,08%	44,76%

La mesure du pouvoir discriminant

Validation des modèles

- Étape très importante car des modèles peuvent :
  - mal se généraliser dans l'espace (autre échantillon) ou le temps (échantillon postérieur)
    - sur-apprentissage
  - être peu efficaces (trop de faux positifs)
  - être incompréhensibles ou inacceptables par les utilisateurs
    - souvent en raison des variables utilisées
- Principaux outils de comparaison des performances
  - matrices de confusion, courbes ROC, de lift, et indices associés
    - courbe ROC : proportion y de vrais positifs en fonction de la proportion x de faux positifs, qui doit être la plus forte possible
    - courbe de lift : proportion y de vrais positifs en fonction de la proportion x d'individus sélectionnés

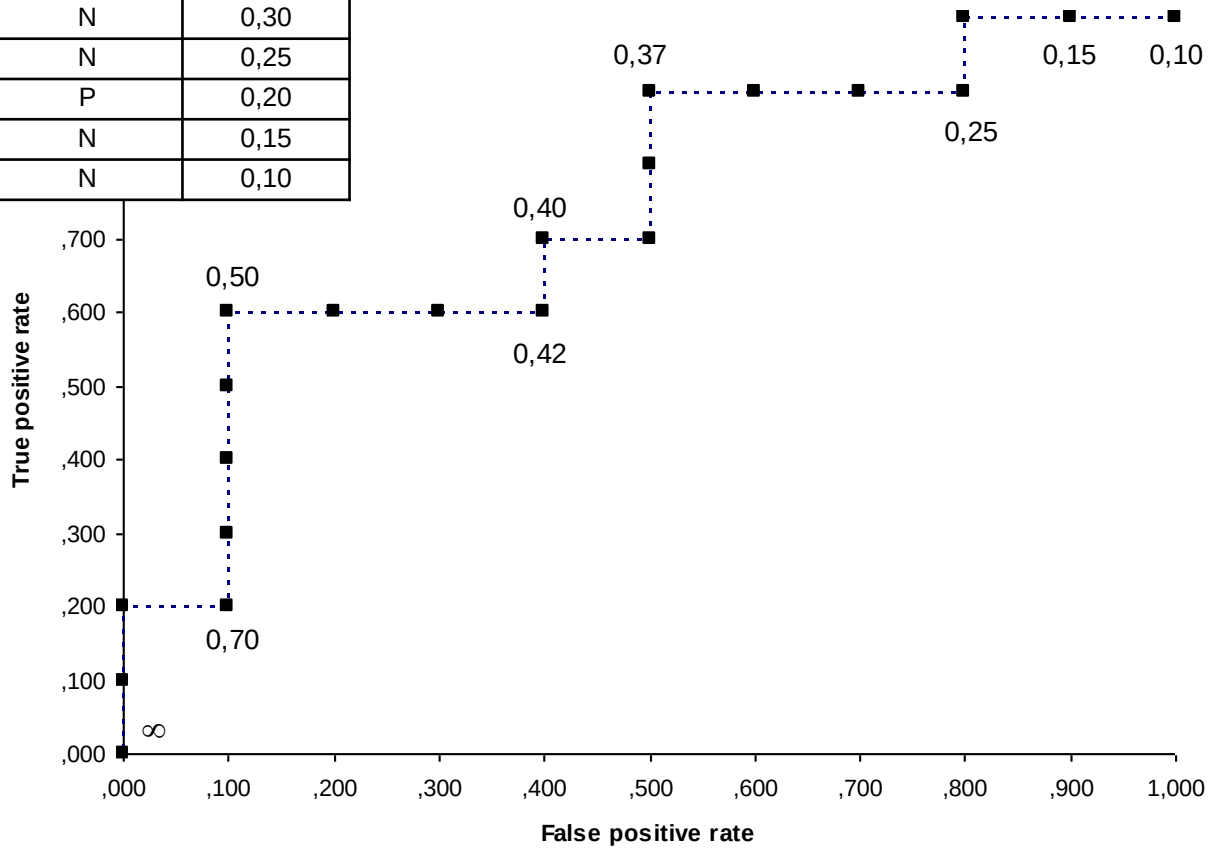
Sensibilité et spécificité

- 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 = $\alpha(s)$ = Prob(score ≥ s / A) = probabilité de bien détecter un positif
  - spécificité = $\beta(s)$ = Prob(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 - $\beta(s)$ = Prob(score ≥ s / B)
  - faux positifs : négatifs considérés comme positifs à cause du score
- Le meilleur modèle : permet de détecter le plus possible de vrais positifs avec le moins possible de faux positifs

La courbe ROC

#	Classe	Score	#	Classe	Score
1	P	0,90	11	P	0,40
2	P	0,80	12	N	0,39
3	N	0,70	13	P	0,38
4	P	0,65	14	P	0,37
5	P	0,60	15	N	0,35
6	P	0,55	16	N	0,30
7	P	0,50	17	N	0,25
8	N	0,45	18	P	0,20
9	N	0,44	19	N	0,15
10	N	0,42	20	N	0,10

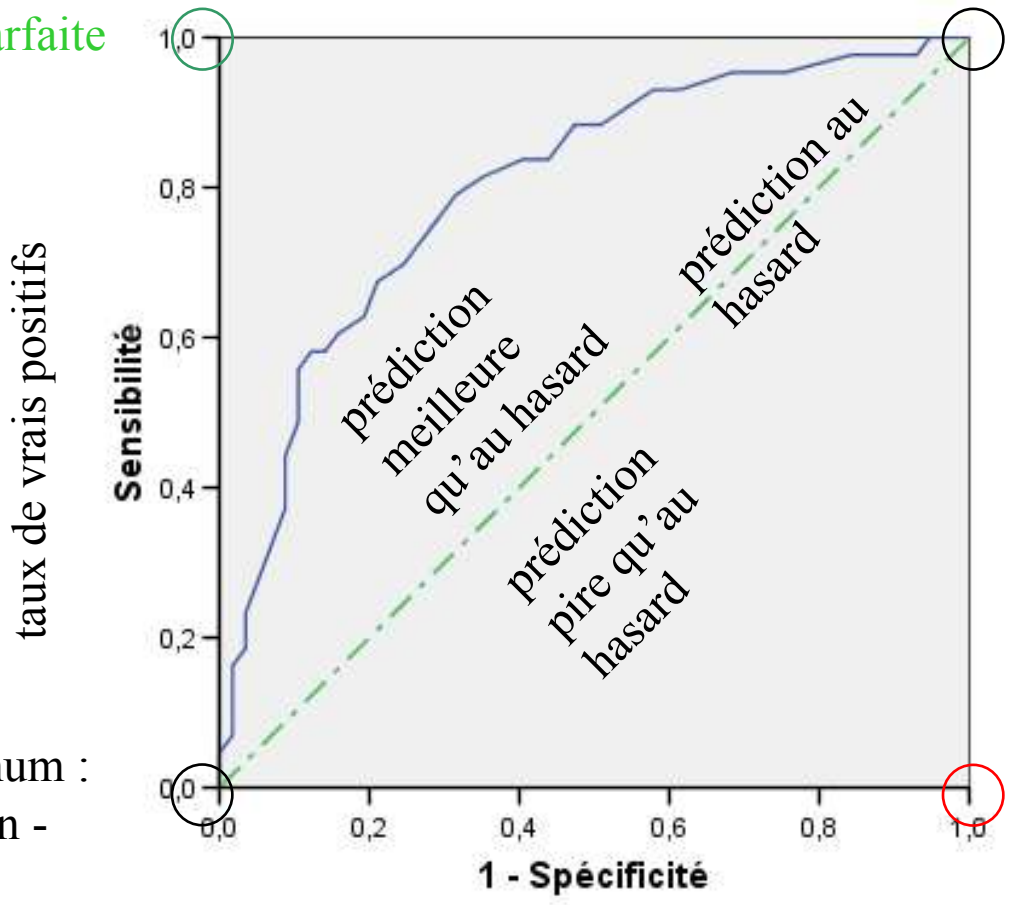
- La courbe ROC
  - sur l'axe Y : sensibilité = $\alpha(s)$
  - sur l'axe X : 1 - spécificité = 1 - $\beta(s)$
  - proportion y de vrais positifs en fonction de la proportion x de faux positifs, lorsque l'on fait varier le seuil s du score



Interprétation de la courbe ROC

prédiction parfaite

seuil s minimum :
tous classés en +



seuil s maximum :
tous classés en -

prédiction nulle

taux de faux positifs

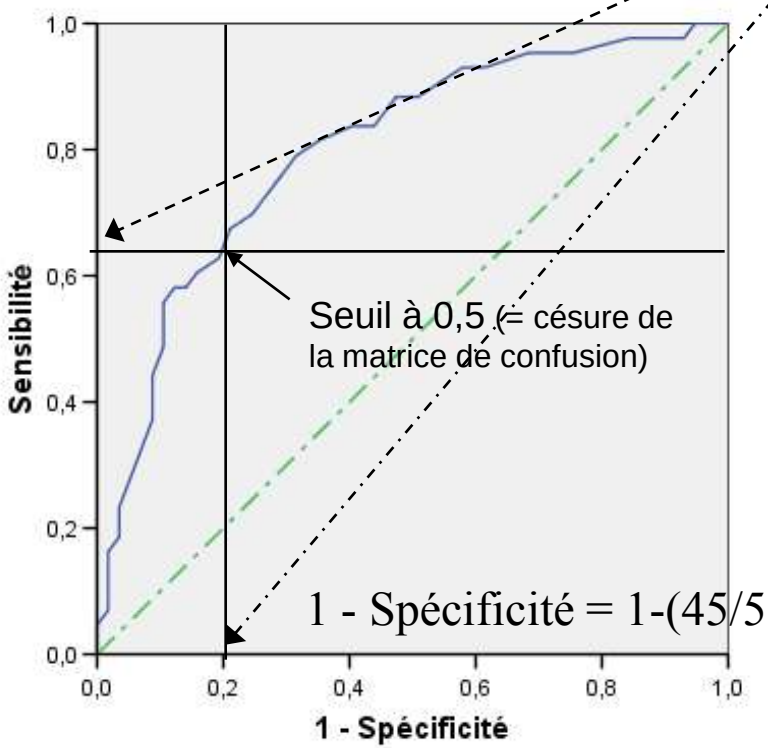
Courbe ROC et matrice de confusion

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



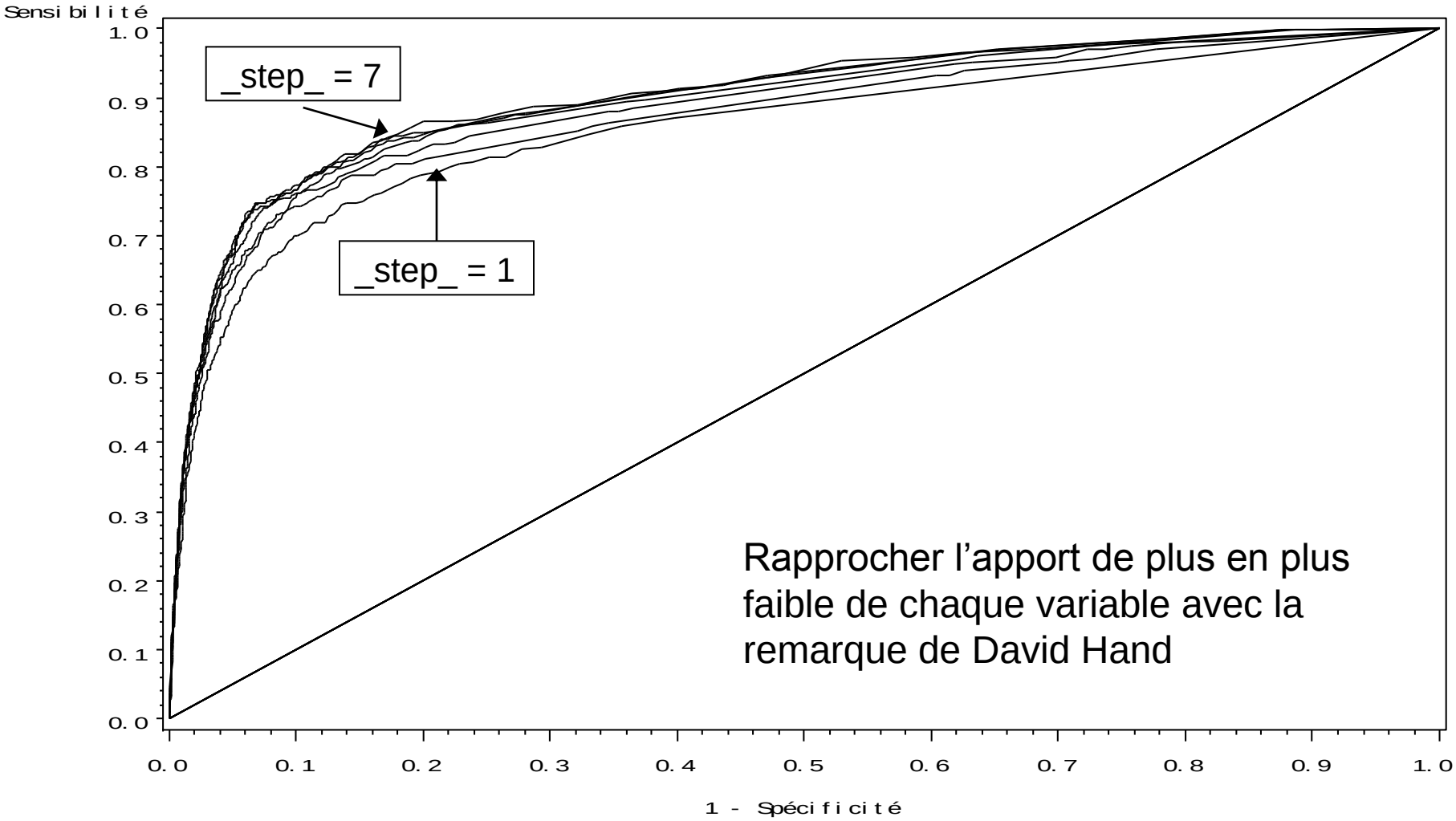
Ensemble des matrices de confusion

Table de classification									
Niveau de prob.	Correct		Incorrect		Pourcentages				
	Événement	Non-événement	Événement	Non-événement	Correct	Sensibilité	Spécificité	POS fausse	NEG fausse
0.000	43	0	57	0	43.0	100.0	0.0	57.0	.
0.100	42	6	51	1	48.0	97.7	10.5	54.8	14.3
0.200	39	24	33	4	63.0	90.7	42.1	45.8	14.3
0.300	36	32	25	7	68.0	83.7	56.1	41.0	17.9
0.400	32	41	16	11	73.0	74.4	71.9	33.3	21.2
0.500	27	45	12	16	**72.0**	**62.8**	**78.9**	**30.8**	**26.2**
0.600	25	50	7	18	75.0	58.1	87.7	21.9	26.5
0.700	19	51	6	24	70.0	44.2	89.5	24.0	32.0
					62.0	16.3	96.5	22.2	39.6
					58.0	2.3	100.0	0.0	42.4
					57.0	0.0	100.0	.	43.0

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

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

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 nombre d'observations dans A (resp. B)
  - on s'intéresse aux paires formées d'un x dans A et d'un y dans B
  - parmi ces n_1n_2 paires : on a concordance si score(x) > score(y) ; discordance si score(x) < score(y)
  - soient nc = nombre de paires concordantes ; nd = nombre de paires discordantes ; n_1n_2 - nc - nd = nombre d'ex aequo
  - aire sous la courbe ROC $\approx$ (nc + 0,5[n_1n_2 - nc - nd]) / n_1n_2
- 3e méthode équivalente : par le test de Mann-Whitney
  - U = $n_1n_2(1 – AUC)$ ou n_1n_2AUC
- Le modèle est d'autant meilleur que l'AUC s'approche de 1
- AUC = 0,5 $\Rightarrow$ modèle pas meilleur qu'une notation aléatoire

Conclusion

Les 8 principes de base de la modélisation

- La préparation des données est la phase la plus longue, peut-être la plus laborieuse 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 nombre de variables explicatives et surtout éviter leur colinéarité
- Perdre parfois de l'information pour en gagner
  - découpage des variables continues en classes
- On modélise mieux des populations homogènes
  - intérêt d'une classification préalable à la modélisation
- La performance d'un modèle dépend plus de la qualité des données et du type de problème que de la méthode

Quelques liens

- Site de la Société Française de Statistique : www.sfds.asso.fr
- Site de Gilbert Saporta (contenu riche, avec de nombreux cours) : http://cedric.cnam.fr/~saporta/
- Site de Philippe Besse (très complet sur les statistiques et le data mining) : www.math.univ-toulouse.fr/~besse/
- Site du livre *The Elements of Statistical Learning* de Hastie, Tibshirani et Friedman : http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- Site de Statsoft (statistiques et data mining) : www.statsoft.com/textbook/stathome.html
- StatNotes Online Textbook (statistiques) : www2.chass.ncsu.edu/garson/pa765/statnote.htm
- Statistique avec R : http://zoonek2.free.fr/UNIX/48_R/all.html
- Données réelles : http://www.umass.edu/statdata/statdata/index.htm
- Site d'Olivier Decourt (spécialiste de SAS) : www.od-datamining.com/
- Blog d'Arthur Charpentier : http://freakonometrics.blog.free.fr/