Machine Learning Supervisé avec R

Florian Pothin

Classe : IAS-M2-DA-2

23/10/2024

Plan

Qu'est-ce que le data mining ?

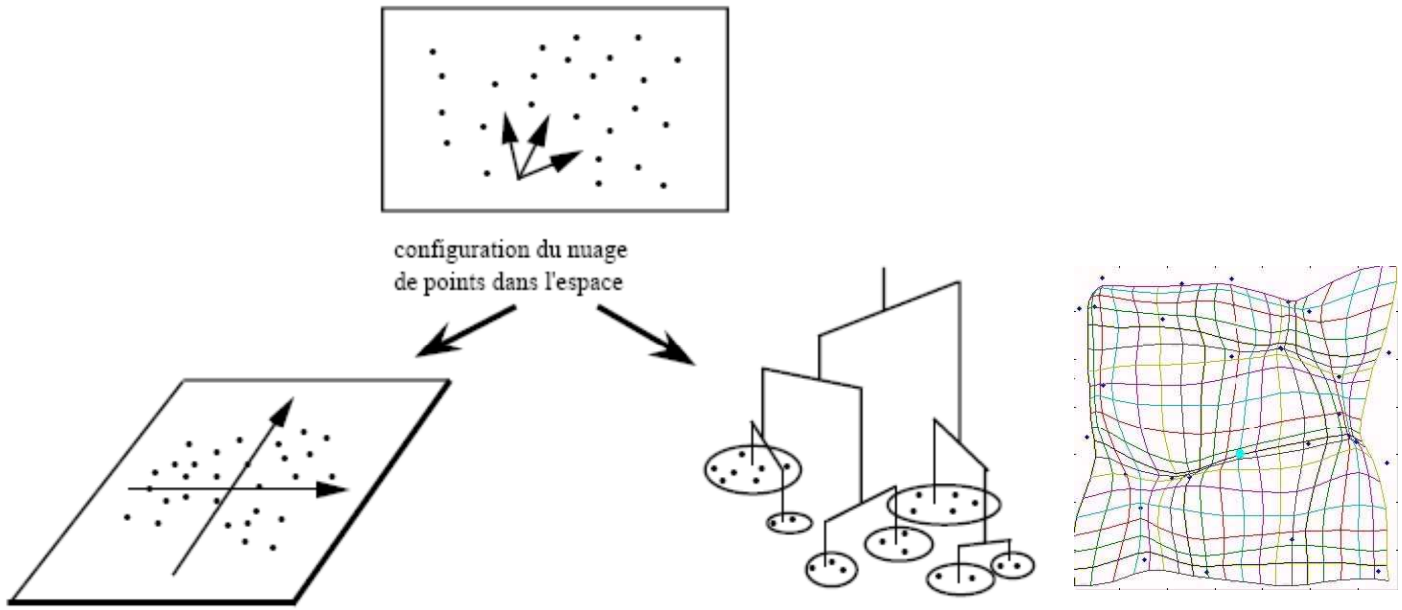
La fouille de données

- Le **data mining** est l'ensemble des :
  - méthodes scientifiques
  - … destinées à l'exploration et l'analyse
  - … de (souvent) grandes bases de données informatiques
  - … en vue de détecter dans ces données des profils-type, des comportements récurrents, des règles, des liens, des tendances inconnues (non fixées *a priori*), des structures particulières restituant de façon concise l'essentiel de l'information utile
  - … pour l'aide à la décision
- On parle d'extraire l'information de la donnée
- Selon le MIT (2001), c'est l'une des 10 technologies émergentes qui « changeront le monde » au XXI^{e} siècle
- Selon Hal Varian (2009), Chief Economist de Google : "I keep saying that the sexy job in the next 10 years will be statisticians."

Les 2 types de méthodes de data mining

- Les méthodes descriptives (recherche de « patterns ») :
  - visent à **mettre en évidence des informations présentes** mais cachées par le volume des données (c'est le cas des *segmentations* de clientèle et des *règles d'associations* de produits sur les tickets de caisse)
  - réduisent, résument, synthétisent les données
  - il n'y a pas de variable à expliquer
- Les méthodes prédictives (modélisation) :
  - visent à **extrapoler de nouvelles informations** à partir des informations présentes (c'est le cas du *scoring*)
  - expliquent les données
  - il y a une variable à expliquer

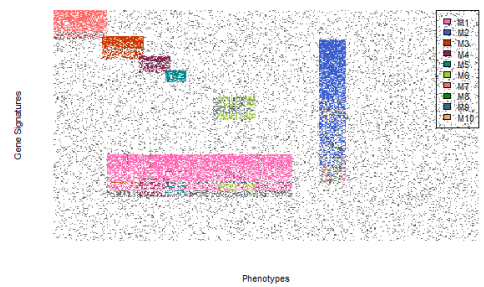
Les 2 principales familles de méthodes descriptives



Source : Lebart-Morineau-Piron, *Statistique exploratoire multidimensionnelle*, page 10

Qu'est-ce que la classification ?

- Regrouper des objets en groupes, ou classes, ou familles, ou segments, ou *clusters*, de sorte que :
  - 2 objets d'un même groupe se ressemblent le plus possible
  - 2 objets de groupes distincts diffèrent le plus possible
  - le nombre des groupes est parfois fixé
  - les groupes ne sont pas prédéfinis mais déterminés au cours de l'opération
- Méthode descriptive :
  - pas de variable à expliquer privilégiée
  - décrire de façon simple une réalité complexe en la résumant
- Utilisation en marketing, médecine, sciences humaines…
  - segmentation de clientèle marketing
- Les objets à classer sont :
  - des individus
  - des variables
  - les deux à la fois (biclustering)



Complexité du problème !

- Le nombre de partitions (classes non recouvrantes) de *n* objets est le nombre de Bell : $B_n = \frac{1}{e}\sum_{k=1}^{\infty}\frac{k^n}{k!}$
- Exemple : pour *n* = 4 objets, on a B_n = 15, avec
  - 1 partition à 1 classe (abcd)
  - 7 partitions à 2 classes (ab,cd), (ac,bd), (ad,bc), (a,bcd), (b,acd), (c,bad), (d,abc)
  - 6 partitions à 3 classes (a,b,cd), (a,c,bd), (a,d,bc), (b,c,ad), (b,d,ac), (c,d,ab)
  - 1 partition à 4 classes (a,b,c,d)
- Exemple : pour *n* = 30 objets, on a $B_{30} = 8{,}47.10^{23}$
- $B_n > \exp(n) \Rightarrow$ Nécessité de définir des critères de bonne classification et d'avoir des algorithmes performants

Classement et prédiction

- Ce sont des méthodes prédictives
  - on parle aussi d'apprentissage supervisé (réseaux de neurones)
- **Classement** : la variable à expliquer (ou « cible », « réponse », « dépendante ») est *qualitative*
  - on parle aussi de **classification** (en anglais) ou **discrimination**
- **Prédiction** : la variable à expliquer est *quantitative*
  - on parle aussi de **régression**
  - exemple : le prix d'un appartement (en fonction de sa superficie, de l'étage et du quartier)
- **Scoring** : classement appliqué à une problématique d'entreprise (variable à expliquer souvent binaire)
  - chaque individu est affecté à une classe (« risqué » ou « non risqué », par exemple) en fonction de ses caractéristiques

Tableau des méthodes descriptives

type	famille	sous-famille	méthode
méthodes descriptives	modèles géométriques	analyse factorielle (projection sur un espace de dimension inférieure)	analyse en composantes principales ACP (variables quantitatives)
			analyse factorielle des correspondances AFC (2 variables qualitatives) analyse des correspondances multiples ACM (+ de 2 var. qualitatives) analyse factorielle de données mixtes (var. qualitatives et quantitatives)
		analyse typologique (regroupement en classes homogènes)	méthodes de partitionnement (centres mobiles, *k*-means, nuées dynamiques)
			méthodes hiérarchiques (ascendantes, descendantes)
		analyse typologique + réduction dimens.	classification neuronale (cartes de Kohonen)
	modèles combinatoires		classification relationnelle (variables qualitatives)
	modèles à base de	détection de liens	détection des règles d'associations

En grisé : méthodes « classiques »

Tableau des méthodes prédictives

type	famille	sous-famille	méthode
méthodes prédictives	modèles à base de règles logiques	arbres de décision	arbres de décision (variable à expliquer quantitative ou qualitative)
	modèles à base de fonctions mathématiques	réseaux de neurones	réseaux à apprentissage supervisé : perceptron multicouches, réseau à fonction de base radiale
		modèles paramétriques ou semi-paramétriques	régression linéaire, ANOVA, MANOVA, ANCOVA, MANCOVA, modèle linéaire général GLM, régression PLS, ridge ou lasso, SVR (variable à expliquer continue)
			analyse discriminante linéaire, régression logistique, régression logistique PLS, ridge ou lasso, SVM (variable à expliquer qualitative)
			modèle log-linéaire, régression de Poisson (variable à expliquer discrète = comptage)
			modèle linéaire généralisé, modèle additif généralisé (variable à expliquer continue, discrète ou qualitative)
	prédiction sans modèle		*k*-plus proches voisins (*k*-NN)

En grisé : méthodes « classiques »

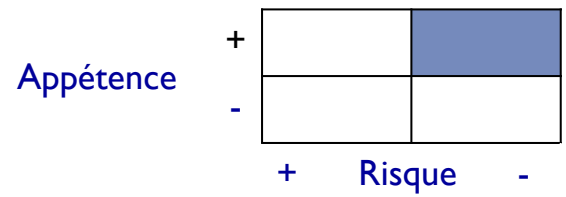
De la statistique à la data science

- Statistique (avant 1950) :
  - quelques centaines d'individus
  - quelques variables recueillies avec un protocole spécial (échantillonnage, plan d'expérience…)
  - fortes hypothèses sur les lois statistiques suivies (linéarité, normalité, homoscédasticité)
  - le modèle prime sur la donnée : il est issu de la théorie et confronté aux données
  - utilisation en laboratoire
- Analyse des données (1960-1980) :
  - quelques dizaines de milliers d'individus
  - quelques dizaines de variables
  - construction des tableaux « Individus x Variables »
  - importance du calcul et de la représentation visuelle
- Data Mining (1990-2010) :
  - plusieurs millions d'individus
  - plusieurs centaines de variables
  - données recueillies avant l'étude, à d'autres fins
  - données imparfaites, avec des erreurs de saisie, des valeurs manquantes…
  - pour l'aide à la décision
  - nécessité de calculs rapides, parfois en temps réel
  - on ne recherche pas toujours l'optimum théorique, mais le plus compréhensible pour des non statisticiens
  - faibles hypothèses sur les lois statistiques
  - la donnée prime sur le modèle : le modèle est issu des données et on en tire éventuellement des éléments théoriques
  - utilisation en entreprise
- Data Science (depuis 2010) :
  - plusieurs centaines de millions d'individus
  - plusieurs milliers de variables, de tous types (images, vidéos, textes, sons…)
  - Big Data recueillies dans les entreprises, les réseaux sociaux, les objets connectés, Internet
  - données parfois très bruitées
  - importance du Machine Learning, du Deep Learning, du text mining, des graphes…
  - pour l'aide à la décision, des nouveaux services, l'intelligence artificielle

A quoi sert le data mining ?

Quelques types de scores

- Score d'appétence
  - prédire l'achat d'un produit ou service
- Score de (comportement) risque
  - prédire les impayés ou la fraude
- Score de pré-acceptation
  - croisement des deux précédents
- 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 data mining dans la banque

- Naissance du score de risque en 1941 (David Durand)
- 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 II

Le data mining 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 data mining dans la téléphonie

- Deux événements :
  - fin du monopole de France Télécom dans la téléphonie fixe
  - arrivée à saturation du marché de la téléphonie mobile
- D'où les sujets dominants dans la téléphonie :
  - score d'attrition (*churn* = changement d'opérateur)
  - optimisation des campagnes marketing
    - et aussi le *text mining* (pour analyser les lettres de réclamation)
- Problème du *churn* :
  - coût d'acquisition moyen en téléphonie mobile : 250 euros
  - plus d'un million d'utilisateurs changent chaque d'année d'opérateur en France
  - les lois facilitant le changement d'opérateur
  - la portabilité du numéro facilite le churn

Le data mining dans le commerce

- Vente Par Correspondance
  - utilise depuis longtemps des scores d'appétence
  - pour optimiser ses ciblages et en réduire les coûts
  - des centaines de millions de documents envoyés par an
- e-commerce
  - personnalisation des pages du site web de l'entreprise, en fonction du profil de chaque internaute
  - optimisation de la navigation sur un site web
- Grande distribution
  - analyse du ticket de caisse
  - détermination des meilleures implantations (géomarketing)

Autres exemples

- De l'infiniment petit (génomique) à l'infiniment grand (astrophysique pour le classement en étoile ou galaxie)
- Du plus quotidien (reconnaissance de l'écriture manuscrite sur les enveloppes) au moins quotidien (aide au pilotage aéronautique)
- Du plus ouvert (e-commerce) au plus sécuritaire (détection de la fraude dans la téléphonie mobile ou les cartes bancaires)
- Du plus industriel (contrôle qualité pour la recherche des facteurs expliquant les défauts de la production) au plus théorique (sciences humaines, biologie…)
- Du plus alimentaire (agronomie et agroalimentaire) au plus divertissant (prévisions d'audience TV)

Qu'est-ce que le Big Data ?

Le Big Data

- Le Big Data recouvre l'ensemble des problématiques associées à la collecte et à l'exploitation de très grands ensembles de données, de natures et de formats très variés (textes, photos, clics, signaux de capteurs, d'objets connectés…), et en évolution très rapide, voire continue
- Le Big Data envahit de nombreux domaines d'activités : santé, industrie, transport, finance, banque et assurance, grande distribution, politiques publiques, sécurité, recherche scientifique…
- Les enjeux économiques sont majeurs :
  - McKinsey (*Big Data, the next frontier for innovation, competition and productivity*, 2011) estime que le Big Data permettrait d'économiser chaque année 300 milliards de dollars aux politiques de santé aux USA et d'engendrer 600 milliards de dollars de consommation en utilisant les données de localisation des consommateurs
  - selon l'Institut Montaigne (*Big data et objets connectés*, 2015), les répercussions des objets connectés et du Big Data sur l'économie française pourraient atteindre les 3,6 % de PIB à échéance 2020 (et environ 7 % en 2025)
- Les impacts sont très importants dans la vie des personnes et des entreprises
- Les défis technologiques sont formidables

Les 3 « V » du Big Data

- Volume
  - L'ordre de grandeur de ces volumes est le pétaoctet (10^{15} octets)
  - L'accroissement du volume vient de l'augmentation :
    - du nombre d'individus observés (plus nombreux ou à un niveau plus fin)
    - de la fréquence d'observation et d'enregistrement des données (mensuel → quotidien, voire horaire)
    - du nombre de caractéristiques observées
  - Cet accroissement vient aussi de l'observation de données nouvelles (Internet, objets connectés, géolocalisation…)
- Variété
  - Ces données sont de natures et de formes très diverses : numériques, logs web, textes, sons, images…
  - Cette variété rend difficile l'utilisation des bases de données usuelles et requiert une variété de méthodes (text mining, web mining…)
- Vitesse, ou Vélocité
  - Vélocité des données qui proviennent de sources où elles sont mises à jour rapidement, parfois en continu (streaming data)
  - Vitesse des traitements appliqués à ces données, parfois en temps réel
  - Dans certains cas, les modèles eux-mêmes sont mis à jour rapidement

Les outils du Big Data

- Outils informatiques de traitement des données massives
  - Besoin de distribuer le stockage et les calculs pour un coût raisonnable : MapReduce, Hadoop, Spark
  - Besoin d'accéder à une puissance informatique importante à la demande : le cloud computing
  - Besoin de gérer des données plus évolutives et moins structurées que les données habituelles : bases sans schéma (NoSQL)
  - Nombreuses solutions open source : Hadoop, Cassandra, Spark, R…
  - Enjeu : conjuguer puissance et sécurité des traitements
- Méthodes statistiques d'analyse des données en grande dimension
  - Machine Learning et Deep Learning
  - Méthodes d'échantillonnage, d'optimisation
  - Estimateurs régularisés (Lasso, ridge…)
  - Apprentissage incrémental (pour les data streams)
  - Text Mining (pour l'analyse des textes en langage naturel)
  - Théorie des graphes (pour les réseaux sociaux)
  - Enjeu : extraire l'information utile d'une masse énorme de données

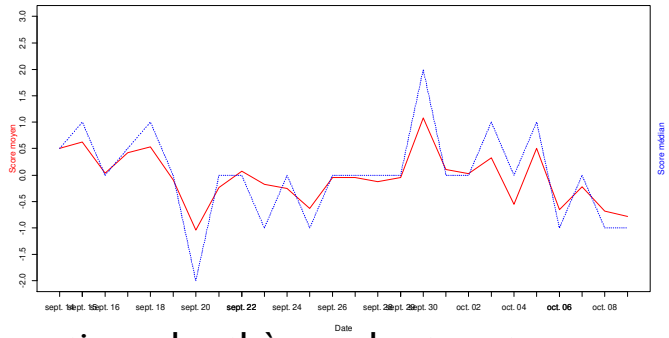
A quoi sert le Big Data ?

Le Big Data dans le marketing

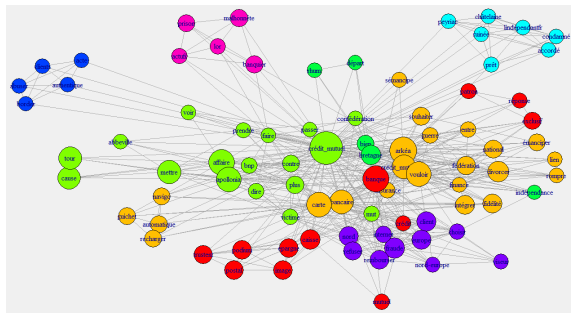
- L'analyse des réseaux sociaux, des forums et des moteurs de recherche permet de découvrir les centres d'intérêt et les préférences des internautes, et donc leur comportement possible face à une proposition de produit ou de service
- C'est particulièrement utile pour les entreprises qui font du B to B to C, ont des contacts avec des distributeurs et non leurs clients finaux, sur lesquels elles ont peu d'informations directes
- L'analyse des réseaux sociaux n'est pas seulement utile à la vente et elle peut aider à la conception de nouveaux produits, par l'analyse de la perception positive ou négative de certaines caractéristiques des produits, et la comparaison avec la concurrence
- Méthodes de traitement du langage naturel : analyse de sentiments…
- Des packages R existent pour traiter les données de Twitter et Facebook

Analyse textuelle automatique des tweets

- Calcul quotidien d'un score d'opinion sur la marque ou l'entreprise
  - On peut croiser ce score avec les nouvelles extraites sur Internet



- Détection automatique des thèmes des tweets



La publicité digitale 1/2

- Des plates-formes virtuelles automatisées (*ad exchange*) mettent en relation directe les acheteurs (annonceurs, ou leurs agences) et vendeurs (éditeurs, ou leurs régies) de publicité sur Internet dans un système d'enchères en temps réel (Real Time Bidding)
- Quand un internaute arrive sur une page Web sur laquelle s'affiche une *bannière* publicitaire, ce qui s'appelle une *impression*, la page envoie une requête (*ad call*) à la plate-forme *ad exchange*
  - cette requête contient des informations telles que l'identifiant de la bannière, l'adresse IP de l'internaute et un identifiant éventuel, la page d'où il vient (*referrer*), son historique de navigation (grâce à des cookies), le navigateur et le système d'exploitation (*user agent*)
- La plate-forme *ad exchange* envoie un appel à enchérir (*bid request*) à l'ensemble des acheteurs potentiels, qui répondent en proposant une enchère (*bid*)
- L'*ad exchange* attribue la bannière au plus offrant et sa publicité s'affiche instantanément sur le site du vendeur (lequel peut filtrer les annonceurs)
- Ce processus dure quelques dizaines de millisecondes, le temps du chargement de la page Web par l'internaute

La publicité digitale 2/2

- Des algorithmes permettent aux annonceurs de déterminer :
  - s'il faut cibler l'internaute
  - le produit et le montant d'enchère qu'ils ont intérêt à proposer
  - et le contenu précis à afficher en fonction des caractéristiques de l'internaute, d'un budget pour une campagne…
- Ils s'appuient sur le profil de l'internaute et son historique de navigation transmis par l'*ad exchange*, et sur des analyses qui montrent que les bannières ont plus ou moins d'effet sur ce type d'internaute quand elles sont placées sur certains sites
- Les annonceurs cherchent à placer leur publicité de façon optimale pour la conversion (achat, téléchargement…)
- Ils font appel à des entreprises de reciblage publicitaire (comme Criteo qui analyse plus de 230 To de données par jour) qui sont rémunérées au clic
- Ils savent repérer les enchaînements de sites visités qui mènent le mieux à la conversion souhaitée, et ils enchérissent quand ils repèrent qu'un internaute est en train d'exécuter cet enchaînement
- Initialement, les bannières publicitaires étaient intégrées statiquement dans les pages Web, mais elles ont ensuite été programmées dynamiquement par des *ad servers* pour s'adapter à l'internaute

Le Big Data dans la finance

- Risque boursier
  - une étude parue dans *Nature* (2013) démontre une corrélation entre les mots clés saisis sur Google et l'évolution des cours de bourse : Avant une chute des indices boursiers, les investisseurs sont préoccupés et recherchent sur Internet des informations les aidant à décider de conserver ou vendre leurs titres.
- Risque financier
  - ce que l'on dit d'une entreprise, son image chez ses partenaires, les analystes financiers ou le grand public, sa réputation, son image en termes de qualité, d'innovation, de respect social et environnemental… ces éléments peuvent concourir à sa santé financière à moyen/long terme et peuvent être intégrés dans les analyses
- Risque de fraude
  - les données de géolocalisation des détenteurs de smartphones peuvent être comparées aux informations relatives au terminal de paiement pour s'assurer qu'elles sont cohérentes

Google Trends et le CAC 40

- Une étude parue dans *Nature* (2013) démontre une corrélation entre les mots clés saisis sur Google et l'évolution des cours de bourse. Avant une chute des indices boursiers, les investisseurs sont préoccupés et recherchent sur Internet des informations les aidant à décider de conserver ou vendre leurs titres

- Google Trends permet de connaître la fréquence à laquelle un terme a été cherché sur Google, avec visualisation par région, des recherches associées et de certains titres dans l'actualité

Le Big Data dans l'assurance

- Aviva a mis au point une application pour smartphone (Aviva Drive) qui analyse le style de conduite des conducteurs afin de leur proposer des tarifs appropriés (http://www.aviva.co.uk/drive/)
- Un projet similaire avait été imaginé en 2006 mais abandonné en 2008 en raison de la difficulté d'installer des « boîtes noires » dans les véhicules
- Cette application analyse pendant 300 km le nombre de kilomètres parcourus, le temps, le type de route…
- Un changement radical de comportement pourra faire suspecter une fraude
- Chez Direct Assurance (AXA) le contrat YouDrive s'appuie sur une DriveBox branchée sur le véhicule, qui prend des mesures (accélérations, freinages, vitesse dans les virages et vitesse par rapport au trafic) permettant de calculer un score mensuel influant la tarification
- Les données du GPS peuvent aussi être exploitées
- « Allianz Conduite connectée » permet aussi d'analyser la conduite pour adapter la tarification et propose l'appel automatique d'assistance en cas d'accident
- Intérêt pour l'assureur et l'assuré (et la société) : diminution du risque de panne et d'accident

Le Big Data dans l'industrie

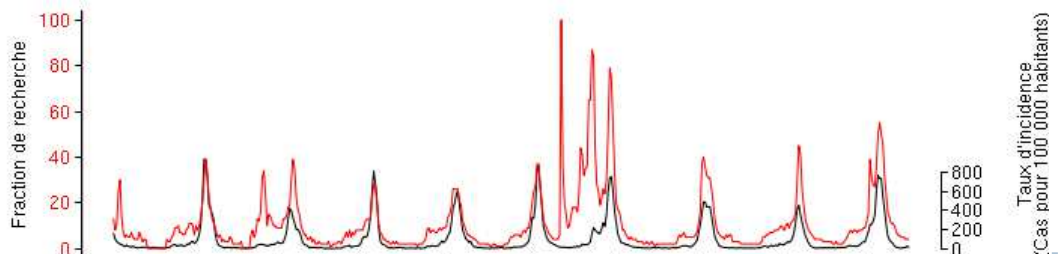
- Les capteurs (température, pression, vibration, usure…) placés sur les composants de l'appareil productif permettent de remonter en temps réel et à distance de nombreuses informations qui, analysées et modélisées, peuvent fournir une probabilité de défaillance, de rupture d'une pièce, et permettre un arbitrage entre :
  - des opérations de maintenance inutilement lourdes et fréquentes, entraînant des dépenses inutiles
  - des opérations de maintenance insuffisantes et laissant se produire des défaillances coûteuses, voire dangereuses
- Autres avantages :
  - optimisation de la chaîne d'approvisionnement (supply chain)
  - amélioration de la conception des futurs appareils
  - les contraintes mesurées sont les contraintes réelles et non celles prévues à la conception
- Nombreux exemples d'application, dont :
  - automobiles
  - ponts (capteurs de tension, corrosion, température, vitesse du vent…)
- Compteurs connectés : prédiction en temps réel de la consommation électrique, mais aussi des dysfonctionnements, et facturation plus économique et plus rapide (Linky)

Le Big Data dans la santé 1/2

- Diagnostic médical à distance : détection de risques de crise cardiaque
- Des applications pour smartphones savent analyser les données transmises par des capteurs (rythme cardiaque, pression sanguine…)
- Collecte anonyme de données de patients permettant de détecter des épidémies ou d'évaluer les effets secondaires d'un traitement
- Le Système National des Données de Santé (SNDS) regroupe les données de l'Assurance Maladie (base SNIIRAM), les données des hôpitaux (base PMSI) et d'autres données à venir, notamment sur les causes de décès
- Croisement possible de données médicales avec des données environnementales pour étudier par exemple l'effet de l'exposition aux pesticides
- Croisement de données génomiques avec des données cliniques, d'imagerie médicale ou de données sociaux-économiques pour expliquer l'apparition d'une maladie ou la réponse à un traitement

Le Big Data dans la santé 2/2

- En analysant les mots clés sur son moteur de recherche, Google a pu établir une corrélation entre certaines requêtes et l'apparition d'une épidémie de grippe. Cette corrélation a été corroborée par les organismes de veille sanitaire et a fait l'objet d'une publication dans *Nature* (2009).
  - Voir : http://www.google.org/flutrends/intl/en_us/about/how.html et http://websenti.u707.jussieu.fr/sentiweb/?page=google
- Cet exemple illustre le V de la vitesse, avec des mises à jour de données quotidiennes et non hebdomadaires comme dans les suivis traditionnels : permet une détection plus rapide de l'épidémie



Quelques principes du data mining

Les 7 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
  - ou utiliser une méthode de pénalisation ou d'agrégation
- 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 souvent plus de la qualité des données et du type de problème que de la méthode de modélisation

Validation croisée

- Dans la k-validation croisée, où n est la taille de l'échantillon et $k \in [2,n]$, on scinde l'échantillon complet X en k parties disjointes X_k de tailles égales, on ajuste un modèle M_k sur chaque complémentaire de partie $X–X_k$, et on mesure l'erreur de prédiction pour chaque observation à l'aide du seul modèle dont l'échantillon d'apprentissage ne contient pas l'observation

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
M1	T	A	A	A	A	A	A	A	A	A
M2	A	T	A	A	A	A	A	A	A	A
M3	A	A	T	A	A	A	A	A	A	A
...	...	...	...	...	...	...	...	...	...	...
M10	A	A	A	A	A	A	A	A	A	T
↓										
Prédiction	par M1	par M2	par M3	par M4	par M5	par M6	par M7	par M8	par M9	par M10

- En particulier, dans la n-validation croisée (« leave-one-out »), on divise l'échantillon en n parties, on ajuste n modèles chacun sur $n–1$ observations, et on mesure l'erreur de prédiction de chaque observation à l'aide du modèle ajusté sur les $n–1$ autres observations
- En pratique, un bon compromis entre le biais et la variance de l'estimateur de l'erreur par k-validation croisée se situe généralement entre $k = 5$ et $k = 15$ (préférer k petit si n est grand, à cause des temps de calcul) et la 10-validation croisée est la plus fréquente

Qualités attendues d'une technique prédictive

- La précision
  - erreur faible, aire sous la courbe ROC élevée…
- 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
- 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

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**	modèle à mesures répétées				

Comparaison des méthodes

	Précision	Robustesse	Concision	Lisibilité	Peu de données	Valeurs manquantes	Valeurs extrêmes	Variables corrélées	Rapidité de calcul	Off-the-shelf	Données complexes
Régression linéaire	+	+	+	+	+	–	–	–	+	–	–
Régression régularisée	+	+	+	+	+	–	–	+	+	–	=
Analyse discriminante linéaire	+	+	+	+	+	–	–	–	+	–	–
Analyse DISQUAL	+	+	+	+	+	=	=	+	+	–	–
Régression logistique	+	+	+	+	+	=	=	–	=	–	–
Arbres de décision	=	–	=	+	–	+	+	+	+	+	=
Bagging (d'arbres)	+	=	–	–	–	+	+	+	–	+	=
Forêts aléatoires	++	+	–	–	–	+	+	+	–	++	+
Extra-Trees	++	+	–	–	–	+	+	+	+	+	+
Boosting (d'arbres)	++	=	–	–	–	+	+	+	–	+	+
Réseaux de neurones	+	–	–	–	–	–	=	=	–	–	++
SVM à noyau non linéaire	+	=	–	–	–	–	=	=	–	–	+

La régression logistique

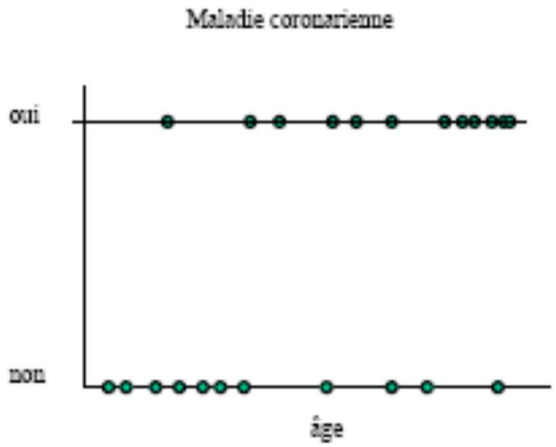
La régression logistique binaire

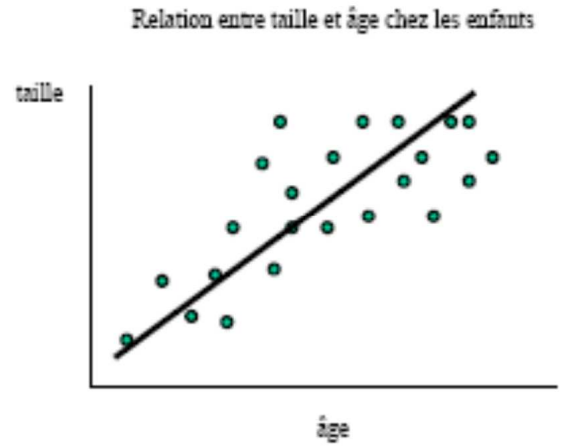
- Y variable à expliquer binaire Y = 0 / 1
- X_j p variables explicatives continues, binaires ou qualitatives
  - p = 1 régression logistique simple
  - p > 1 régression logistique multiple
- Généralisation : régression logistique polytomique
  - la variable à expliquer Y est qualitative à k modalités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

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

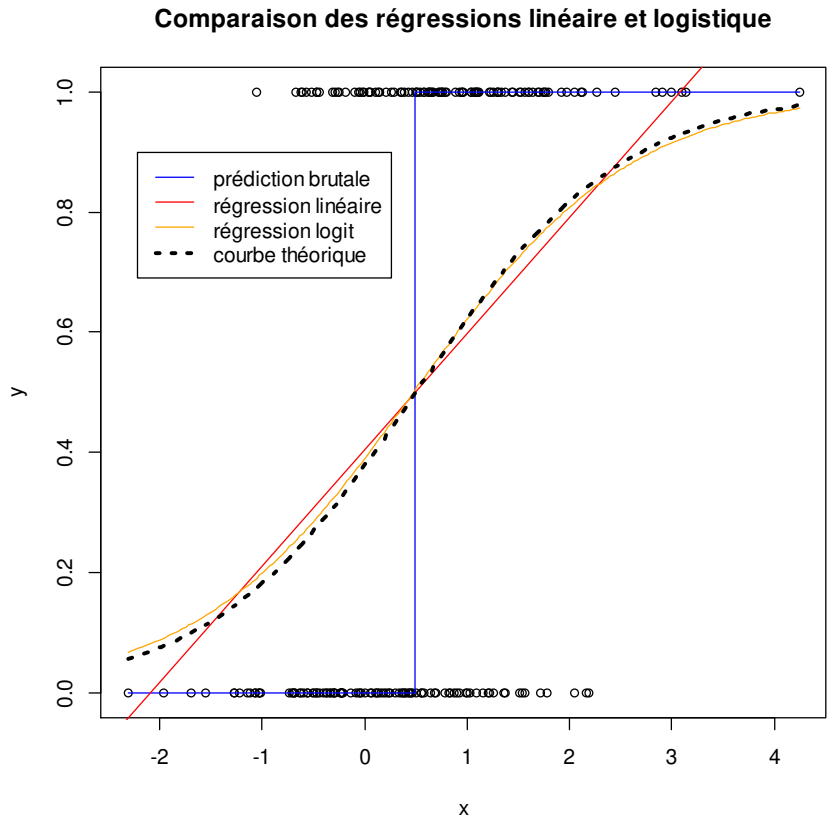
  sous la forme $E(Y/X=x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
- Difficulté ! X_i continues => terme de droite non borné alors que $Prob(Y=1/X=x) \in [0,1]$ => il faut le transformer
  - en régression linéaire : E(Y/X=x) n'est pas bornée

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 $Y=1$

On suppose que :

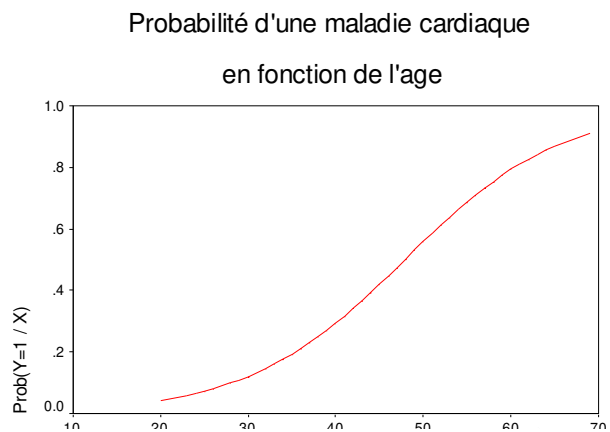
$P(Y=0) = P(Y=1)$

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, en épidémiologie, en économie…



Âge et 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

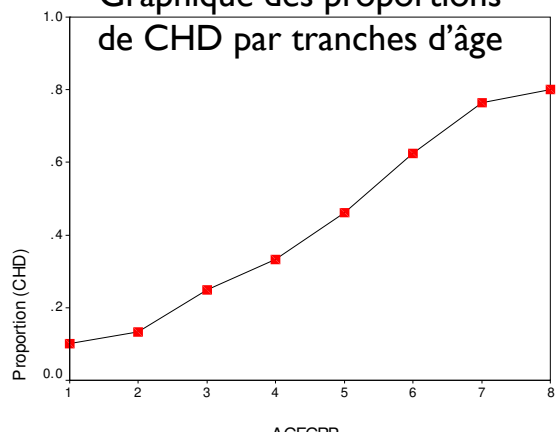
La régression logistique binaire

- 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 π(x) = Prob(Y=1/X=x) sous la forme :

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

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

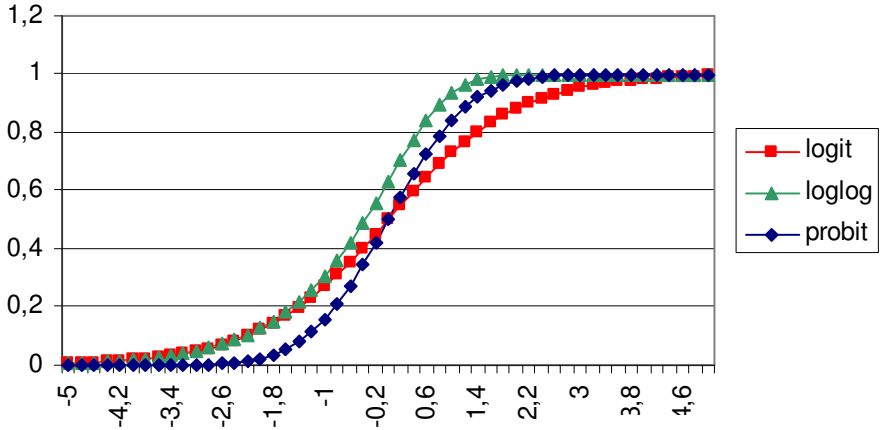
Fonction de lien : Logit(π(x))

- 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

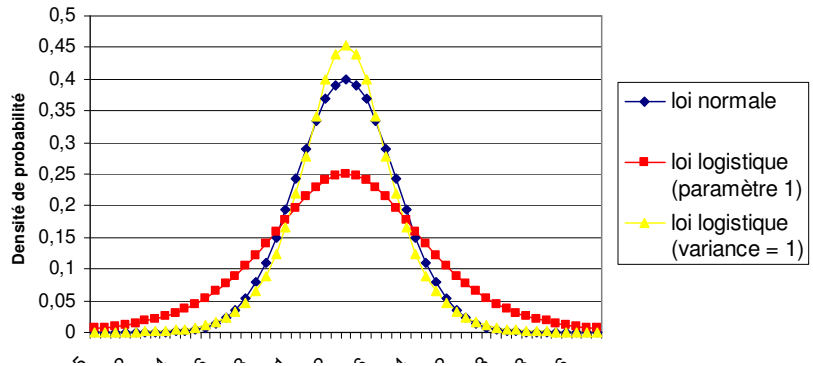
Modèle	**Fonction de lien**	**Fonction de transfert**
Logit	Log (μ/ [1 – μ]) fonction inverse de la fonction de répartition de la loi logistique de paramètre 1	$\frac{\exp(t)}{1+\exp(t)} = \int_{-\infty}^{t} \frac{\exp(z)}{(1+\exp(z))^2} dz$
Probit (normit)	fonction inverse de la fonction de répartition de la 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



NB : c'est la loi logistique de paramètre 1 qui est utilisée dans le logit, mais sa variance est $\frac{\pi^2}{3} \neq 1$ et son aplatissement ne peut directement être comparé à celui de la loi normale réduite

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



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 = fonction de densité mais vue comme fonction des coefficients et non des observations

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

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

- 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 (exemple : Newton-Raphson)

Recherche du maximum de vraisemblance

- L'objectif est de trouver les coefficients qui maximisent $Prob(Y=1/X=x^i)$ lorsque $y^i = 1$, et le minimisent lorsque $y^i = 0$
- Cas d'une seule variable:
  - La constante β_0 permet de décaler à droite ou à gauche la courbe en S en fonction de la position des 0 et des 1, et la pente β_1 permet de redresser plus ou moins la courbe logistique et est d'autant plus importante que les 0 et les 1 sont mieux séparés par la variable $\Rightarrow$ la variable aura plus de poids dans le modèle ((importance du pouvoir discriminant $\approx$ importance du coefficient)
- Il faut que la pente soit telle que, dans une région des x^i, la proportion de $y^i = 1$ soit le plus proche possible de l'espérance conditionnelle $Prob(Y=1/X=x^i)$: il faut donc ajuster les coefficients pour qu'au voisinage d'un x^i, l'ordonnée de la courbe soit au niveau, compris entre 0 et 1, de la proportion de $y^i = 1$
- Valeurs 0 et 1 de y^i peu séparées : pente faible
- Valeurs 0 et 1 de y^i très séparées : pente forte
  - voire infinie (marche d'escalier) en cas de séparation complète !

Recherche du maximum de vraisemblance

- Soit $\hat{\beta}$ le vecteur de coefficients maximisant la log-vraisemblance $LogL(\beta)$
- En dérivant, on a $LogL'(\hat{\beta}) = 0 \approx LogL'(\beta_k) + (\hat{\beta} - \beta_k)\, LogL''(\beta_k)$
- D'où $\hat{\beta} \approx \beta_k - LogL'(\beta_k)/LogL''(\beta_k)$
- Cela suggère l'algorithme itératif $\beta_{k+1} = \beta_k - \frac{LogL'(\beta_k)}{LogL''(\beta_k)}$
- Le numérateur $LogL'(\beta_k)$ est le gradient $\partial LogL(\beta)/\partial\beta$ qui s'écrit $\sum_{i=1}^{n} x_i(y_i - p(x_i, \beta))$ ou $X^t(y - p)$ en écriture vectorielle
- Le dénominateur $LogL''(\beta_k)$ est la matrice hessienne des dérivées secondes qui s'écrit $\frac{\partial^2 LogL(\beta)}{\partial\beta\partial\beta^t} = -\sum_{i=1}^{n} x_i x_i^t p(x_i, \beta)(1 - p(x_i, \beta))$ ou $-X^t\Omega X$ en écriture vectorielle, où Ω est la matrice diagonale dont le i^{ème} terme est $p(x_i, \beta)(1 - p(x_i, \beta))$

Maximum de vraisemblance dans R

```
maxiter <- 5
beta <- lm(Y~0+X)$coefficients
for (s in 1:maxiter){
pi <- 1/(1+exp(-X%*%beta))
gradient <- t(X)%*%(Y-pi)
omega <- matrix(0,nrow(X),nrow(X))
diag(omega) <- (pi*(1-pi))
hessian <- -t(X)%*%omega%*%X
if (sum(abs(solve(hessian)%*%gradient)) < (p*0.000001)) break
beta <- beta - solve(hessian)%*%gradient
}
sd <- sqrt(diag(solve(-hessian)))
# p-value associée au test de Student
modele <- data.frame(beta, sd, beta/sd, 2*(1-
pnorm(abs(beta/sd))))
colnames(modele) <- c("Estimate", "Std. Error", "z value",
"Pr(>|z|)")
```

Cas de la régression logistique simple

- On voit que l'erreur-type des coefficients est calculé ainsi :

```
sd <- sqrt(diag(solve(-hessian)))
```

- En effet, 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 inverse de la matrice hessienne :

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

intervient dans la statistique de Wald (voir + loin)

- À noter que l'inversion de la matrice hessienne est impossible en cas de séparation complète des classes

Exemple de régression logistique avec `glm`

```
> set.seed(123)
> n <- 1000
> p <- 5
> A <- matrix (runif(n*p), n, p)
> A <- data.frame(A, y = sample(c(0,1), n, replace = T))
> Y <- A["y"]
> X <- as.matrix(cbind(1, A[,!names(A) %in% "y"]))
> # régression logistique avec glm
> system.time(L <- glm(y ~., data=A, family=binomial(link="logit")))
utilisateur     système      écoulé
       0.01        0.00        0.02
> summary(L)
Call:
glm(formula = y ~ ., family = binomial(link = "logit"), data = A)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.423  -1.190   1.001   1.148   1.367

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.22069    0.25674   0.860   0.3900
X1           0.03367    0.22243   0.151   0.8797
X2           0.44478    0.22414   1.984   0.0472 *
X3          -0.19706    0.21917  -0.899   0.3686
X4          -0.14983    0.21714  -0.690   0.4902
X5          -0.45945    0.22624  -2.031   0.0423 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 1385.4  on 999  degrees of freedom
Residual deviance: 1376.3  on 994  degrees of freedom
AIC: 1388.3

Number of Fisher Scoring iterations: 4
```

Exemple de régression logistique I

- On obtient exactement les mêmes résultats avec notre propre codage, mais avec un temps de calcul plus grand

```
> time <- proc.time()
> maxiter <- 5
> beta <- lm(Y~0+X)$coefficients
> for (s in 1:maxiter){
+ pi <- 1/(1+exp(-X%*%beta))
+ gradient <- t(X)%*%(Y-pi)
+ omega <- matrix(0,nrow(X),nrow(X))
+ diag(omega) <- (pi*(1-pi))
+ hessian <- -t(X)%*%omega%*%X
+ if (sum(abs(solve(hessian)%*%gradient)) < (p*0.000001)) break
+ beta <- beta - solve(hessian)%*%gradient
+ }
> sd <- sqrt(diag(solve(-hessian)))
> modele <- data.frame(beta, sd, beta/sd, 2*(1-pnorm(abs(beta/sd))))
> colnames(modele) <- c("Estimate", "Std. Error", "z value", "Pr(>|z|)")
> modele
      Estimate Std. Error    z value   Pr(>|z|)
1   0.22069229  0.2567389  0.8595983 0.39001049
X1  0.03366634  0.2224279  0.1513585 0.87969296
X2  0.44478224  0.2241363  1.9844280 0.04720815
X3 -0.19706193  0.2191677 -0.8991375 0.36857944
X4 -0.14982850  0.2171412 -0.6900049 0.49019108
X5 -0.45945200  0.2262433 -2.0307874 0.04227657
> cat("\n", "Convergence en ", s-1 ," itérations","\n")

 Convergence en  3  itérations
> proc.time() - time
utilisateur     système      écoulé
```

Exemple de régression logistique II

- Remarque : Il suffit d'encapsuler le code R dans une fonction pour diviser par 3 le temps de calcul !

```
> f <- function(x,y){
+ maxiter <- 5
+ beta <- lm(y~0+x)$coefficients
+ for (s in 1:maxiter){
+ pi <- 1/(1+exp(-x%*%beta))
+ gradient <- t(x)%*%(y-pi)
+ omega <- matrix(0,nrow(x),nrow(x))
+ diag(omega) <- (pi*(1-pi))
+ hessian <- -t(x)%*%omega%*%x
+ if (sum(abs(solve(hessian)%*%gradient)) < (p*0.000001)) break
+ beta <- beta - solve(hessian)%*%gradient
+ }
+ sd <- sqrt(diag(solve(-hessian)))
+ modele <- data.frame(beta, sd, beta/sd, 2*(1-pnorm(abs(beta/sd))))
+ colnames(modele) <- c("Estimate", "Std. Error", "z value", "Pr(>|z|)")
+ print(modele)
+ cat("\n", "Convergence en ", s-1 ," itérations","\n")
+ }
> system.time(f(X,Y))
      Estimate Std. Error    z value   Pr(>|z|)
1   0.22069229  0.2567389  0.8595983 0.39001049
X1  0.03366634  0.2224279  0.1513585 0.87969296
X2  0.44478224  0.2241363  1.9844280 0.04720815
X3 -0.19706193  0.2191677 -0.8991375 0.36857944
X4 -0.14982850  0.2171412 -0.6900049 0.49019108
X5 -0.45945200  0.2262433 -2.0307874 0.04227657

 Convergence en  3  itérations
utilisateur     système      écoulé
       0.09        0.00        0.09
```

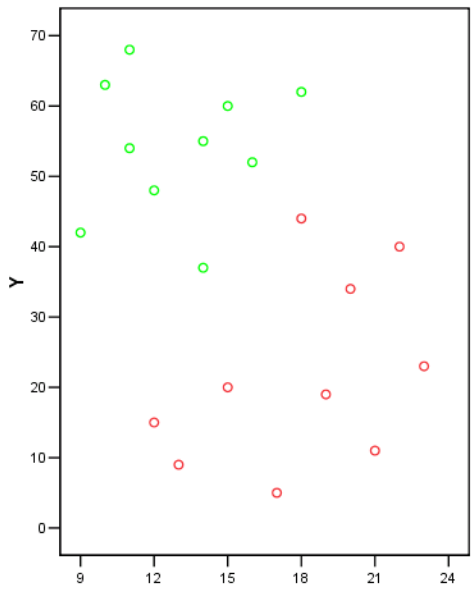
Séparation complète des classes

- On a séparation complète (ou parfaite) quand un prédicteur ou une combinaison linéaire de prédicteurs sépare parfaitement les classes à prédire
- Pas de convergence de l'algorithme de recherche du maximum de vraisemblance : les coefficients trouvés ont une très grande erreur-type
- Messages de :
  - SAS :
```
Complete separation of data points detected. WARNING:
The maximum likelihood estimate does not exist. WARNING:
The LOGISTIC procedure continues in spite of the above
warning. Results shown are based on the last maximum
likelihood iteration. Validity of the model fit is
questionable.
```
  - SPSS :
```
The parameter covariance matrix cannot be computed.
Remaining statistics will be omitted. Estimation terminated
at iteration number 20 because a perfect fit is detected.
This solution is not unique.
```
  - R :
```
Warning messages:
1: glm.fit: l'algorithme n'a pas convergé
2: glm.fit: des probabilités ont été ajustées numériquement
à 0 ou 1
```

Séparation complète : solutions

- Voir si la variable à expliquer n'est pas une version transformée d'un prédicteur (= variable explicative)
- Exclure le prédicteur concerné du modèle (alors qu'il est très prédictif !)
- Regrouper des modalités du prédicteur concerné pour faire disparaître la séparation complète
- Ne rien faire car seul le (ou les) prédicteur concerné a un coefficient non fiable et les autres coefficients nous intéressent
  - on peut calculer un autre intervalle de confiance que celui de Wald, comme avec l'option CLPARM=PL de SAS qui permet au moins d'obtenir une borne supérieure ou une borne inférieure de l'intervalle de confiance, mais cela ne corrige pas les coefficients et les prédictions du modèle
- Utiliser la régression logistique exacte (R : package `elrm`) si l'échantillon n'est pas trop important, et son estimateur médian non biaisé à la place de l'estimateur du maximum de vraisemblance
- Utiliser la régression logistique pénalisée de Firth (1993) "Bias reduction of maximum likelihood estimates", *Biometrika*, 80,1 (R : package `logistf`) dont la solution pour diminuer le biais de l'estimateur du maximum de vraisemblance converge toujours, même dans le cas de la séparation complète
- Supprimer des observations provoquant la séparation complète ou modifier leur valeur de variable à expliquer : le côté arbitraire est embêtant
- Utiliser une méthode bayésienne quand on a une connaissance *a priori* sur les prédicteurs et la loi de probabilité de leurs coefficients (on peut utiliser une loi non informative)
- Considérer que la séparation complète n'existe pas que dans l'échantillon de modélisation mais aussi dans la population tout entière, et prédire la variable à expliquer à l'aide du seul prédicteur séparant complètement

Séparation complète des classes



Non convergence vers une solution

La solution atteinte au terme des 20 itérations a des erreurs-types énormes pour les coefficients

Historique des itérations[a,b,c,d]

Itération		-2log-vraisemblance	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.

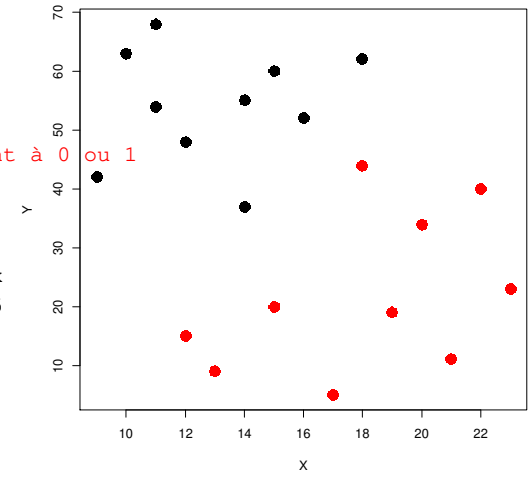
Variables dans l'équation

		B	E.S.	Wald	ddl	Signif.	Exp
Etape 1[a]	X	13,184	2237,865	,000	1	,995	5318
	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 complète des classes

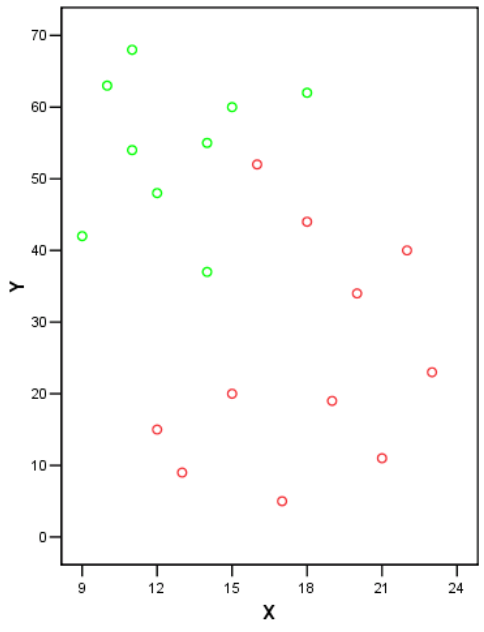
```
> X  <- c(9,10,11,11,12,14,14,15,16,18,12,13,15,17,19,18,20,21,22,23)
> Y  <- c(42,63,54,68,48,37,55,60,52,62,15,9,20,5,19,44,34,11,40,23)
> C1 <- c(0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1)
> plot(X, Y, col=factor(C1))
> modele1 <- glm(C1~X+Y, family=binomial(link = "logit"))
Warning messages:
1: glm.fit: l'algorithme n'a pas convergé
2: glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1
> summary(modele1)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.671e-05  -2.110e-08   0.000e+00   2.110e-08   1.900e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -131.786 186755.901  -0.001    0.999
X              17.332  19127.792   0.001    0.999
Y              -3.584   3781.318  -0.001    0.999

Null deviance: 2.7726e+01  on 19  degrees of freedom
Residual deviance: 1.4181e-09  on 17  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25
```



Non convergence vers une solution

La solution atteinte au terme des 25 itérations a des erreurs-types énormes pour les coefficients

Séparation incomplète des classes



Un seul individu a changé de classe $\Rightarrow$ plus de séparation complète $\Rightarrow$ convergence en 10 itérations vers une solution avec des erreurs-types limitées

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-vrais emblance 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.

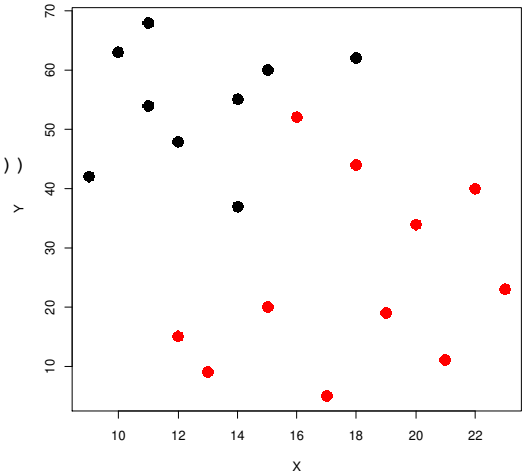
Variables dans l'équation

		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.

Séparation incomplète des classes

```
> X  <- c(9,10,11,11,12,14,14,15,16,18,12,13,15,17,19,18,20,21,22,23)
> Y  <- c(42,63,54,68,48,37,55,60,52,62,15,9,20,5,19,44,34,11,40,23)
> C2 <- c(0,0,0,0,0,0,0,0,1,0,1,1,1,1,1,1,1,1,1,1)
> plot(X, Y, col=factor(C2))
> modele2 <- glm(C2~X+Y, family=binomial(link = "logit"))
> summary(modele2)

Call:
glm(formula = C2 ~ X + Y, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.12171  -0.03615   0.00044   0.03398   1.62025

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.8414    13.9489  -0.777    0.437
X             1.4169     1.3789   1.028    0.304
Y            -0.2467     0.1894  -1.302    0.193

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.5256  on 19  degrees of freedom
Residual deviance:  5.0144  on 17  degrees of freedom
AIC: 11.014

Number of Fisher Scoring iterations: 9
```



Un seul individu a changé de classe => plus de séparation complète => convergence en 9 itérations vers une solution avec des erreurs-types limitées

Vraisemblance et déviance d'un modèle

- Soit $L(\beta_0)$ = vraisemblance du modèle réduit à la constante
- Soit $L(\beta_{max})$ = vraisemblance du modèle saturé (avec toutes les variables explicatives et toutes les interactions pour en avoir autant que d'observations distinctes) = vraisemblance maximale = 1 pour un modèle binaire
- Soit $L(\beta_k)$ = vraisemblance du modèle avec k variables
- On définit la déviance :
  $D(\beta_k) = -2\,[\text{Log } L(\beta_k) - \text{Log } L(\beta_{max})] = -2 \text{ Log } L(\beta_k)$
- D'après la formule de la vraisemblance $\prod_{i=1}^{n} \pi(x^i)^{y^i}(1-\pi(x^i))^{1-y^i}$, la déviance est la somme des carrés des résidus des déviances individuelles $\pm\sqrt{-2\left[y^i Log\left(\pi(x^i)\right) + (1-y^i)Log\left(1-\pi(x^i)\right)\right]}$
  - le signe de l'expression étant positif si la valeur observée y^i est plus grande que la valeur prédite $\pi(x^i)$, négatif sinon
- C'est la somme des carrés résiduels en régression linéaire
  - $L(\beta_k)$ petit $\in [0,1] \Rightarrow$ -2 Log $L(\beta_k) \in [0,+\infty[$ avec un terme « 2 » pour avoir l'analogie entre déviance et $\Sigma(\text{erreurs})^2$
- But de la régression logistique : maximiser la vraisemblance $L(\beta_k) \Leftrightarrow$ minimiser la déviance $D(\beta_k)$

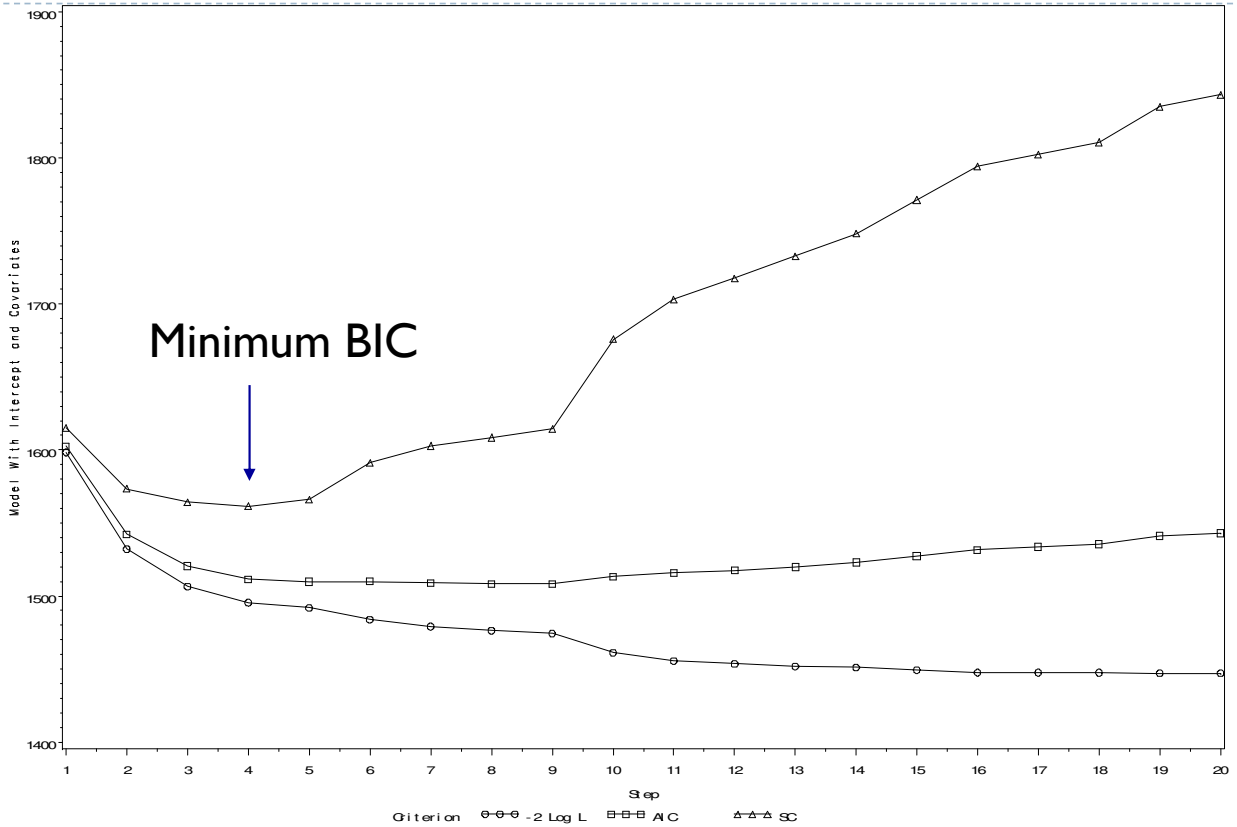
Comparaison de modèles

- Pour savoir s'il convient d'ajouter q degrés de liberté (variables quantitatives ou modalités de variables qualitatives) 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 q 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 n'accepte pas les q nouveaux degrés de liberté (en fait on ne rejette pas l'hypothèse qu'ils sont 0)
- Méthode utilisée en régression pas à pas

Critères AIC et BIC

- Critère d'Akaike AIC = – 2 Log $L(\beta_k)$ + 2(k+1)
  - k = nombre de degrés de liberté = nombre de paramètres à estimer
  - parfois préférable quand n est petit
  - mesure la perte d'information (au sens de la théorie de l'information) que représente le modèle par rapport au « vrai » modèle
  - peut être corrigé par l'ajout d'une pénalité 2k(k+1)/(n–k–1) si k est grand
- Critère de Schwartz SC = BIC = – 2 Log $L(\beta_k)$ + (k+1).log n
  - n = nombre total d'individus
  - pénalise les modèles complexes
  - le modèle qui minimise le critère BIC est celui qui maximise la probabilité *a posteriori* du modèle conditionnellement aux données observées : c'est le modèle qui a la plus forte probabilité d'avoir généré les données observées
- Ces 2 critères permettent de comparer 2 modèles
  - ils doivent être le plus bas possible

Utilisation de l'AIC et du BIC



Le χ^2 de Wald

- Statistique de Wald = $(\beta_i$ / écart-type$(\beta_i))^2$
- suit un χ^2 à 1 degré de liberté sous l'hypothèse nulle H_0 : le coefficient $\beta_i = 0$
- puisque β_i / écart-type(β_i) suit une loi normale centrée-réduite sous H_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 > 3,84 = $1{,}96^2$ au seuil de 95% (sorties de SAS et SPSS)
  - ou $|\ \beta_i$ / écart-type$(\beta_i)\ |$ > 1,96 (sorties de R)

- Méthode utilisée en régression pas à pas
- Éviter le χ^2 de Wald si peu d'observations ou si les coefficients β_i sont grands : Hauck et Donner (1977) et Jennings (1986) ont mis en évidence dans ce cas le manque de puissance du test de Wald, qui peut donc ne pas rejeter l'hypothèse H_0 de la nullité du coefficient, alors que celui-ci est significativement différent de zéro
- Pour les variables qualitatives à plus de 2 modalités, la significativité du résultat de ce test dépend du choix de la modalité de référence

Logit : odds-ratio d'un régresseur X_i

- Mesure l'évolution du rapport des probabilités 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)
- Si la cote $\pi(x)/[1 - \pi(x)]$ est infinie alors le coefficient β_i aussi (cas de séparation complète)
- Formule générale (mais vraie seulement pour la régression logistique *logit*)

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

- 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 d'un odds-ratio

- Attention : odds-ratio $\neq$ du risque relatif $\pi(x+1)/\pi(x)$
  - sauf si $\pi(x)$ est petit (détection de phénomène rare)
- Un seul odds-ratio pour X binaire
  - exemple : comparer les hommes (x=1) et les femmes (x=0)
- Un seul odds-ratio pour X continue
  - remarque : il n'est pas toujours pertinent de comparer l'âge 61 et 60, 60 et 59… avec le même odds-ratio, car l'évolution de la morbidité n'est pas forcément la même tout au long de l'existence
  - de plus, on risque de se heurter à un manque de robustesse du modèle par manque de données (voir exemple CHD ci-dessus)
- Les variables qualitatives ont autant d'odds-ratios que de modalités moins 1, l'une des modalités étant prise pour référence et son coefficient étant généralement posé égal à 0
  - c'est la convention la plus fréquente et la plus commode, mais ce coefficient peut aussi être l'opposé de la somme de tous les autres coefficients
- Un odds-ratio < 1 (un coefficient B < 0) indique une influence négative de la variable explicative sur la variable à prédire, et un odds-ratio > 1 (un coefficient B > 0) indique une influence positive

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		

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 probabilité est faible), les fréquences observées et attendues sont significativement différentes et le modèle ne s'ajuste pas bien aux données

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

Résumé des tests et vérifications

- Test du χ^2 sur la déviance – 2 Log L(β_k)
- AIC et BIC
- Test du χ^2 sur indicateur de Wald (> 3,84)
  - 1 $\notin$ IC à 95 % de l'odds-ratio = exp($a_i \pm 1{,}96\sigma(a_i)$)
- Cohérence des coefficients et notamment de leurs signes
- Test de Hosmer et Lemeshow sur la comparaison des proportions observées et théoriques
- Multicolinéarité (tolérance, VIF)
- Matrice de confusion, tests de concordance, courbe ROC, aire sous la courbe ROC
- Moins de 20 degrés de liberté (variables ou modalités) sont souvent retenus

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 du logit mais non du 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 variables explicatives ont mêmes distributions dans E et E'), mais la fréquence de l'événement sera plus faible dans les acceptés de E', puisque leur probabilité $P(Y=1/X=x, X \in E') < P(Y=1/X=x, X \in E)$

Avantages de la régression logistique

- Permet de traiter les variables explicatives discrètes, qualitatives ou continues
- Permet de traiter une variable à expliquer ordinale ou nominale
- Hypothèses plus générales que l'analyse discriminante (pas de multinormalité ni d'homoscédasticité)
- Odds-ratios facilement interprétables (pour le modèle *logit*)
- Peut prendre en compte les interactions entre variables
- Modélise directement une probabilité
- Fournit des intervalles de confiance sur les coefficients
- Nombreux tests statistiques disponibles
- Possibilité de sélection pas à pas des variables
- Modèles produits lisibles et aisément programmables

Limites de la régression logistique

- Suppose la non-colinéarité des variables explicatives
  - sauf à utiliser une pénalité du type Lasso ou ridge
- 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)

Modèle linéaire généralisé

- 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
- La distribution de Y/X=x peut être :
  - normale (continue : régression) $g(\mu) = \mu$
    - c'est le cas du modèle linéaire général
  - gamma (continue positive) $g(\mu) = -1/\mu$
  - Bernoulli (discrète : oui/non) $g(\mu)=\log(\mu/1-\mu)\ldots$ (logit, probit, log-log)
  - de Poisson (discrète : comptage) $g(\mu) = \log(\mu)$
    - Y = nombre de sinistres (assurance) ou effectif d'un tableau de contingence (modèle log-linéaire)
  - multinomiale, etc.

Modèle additif généralisé

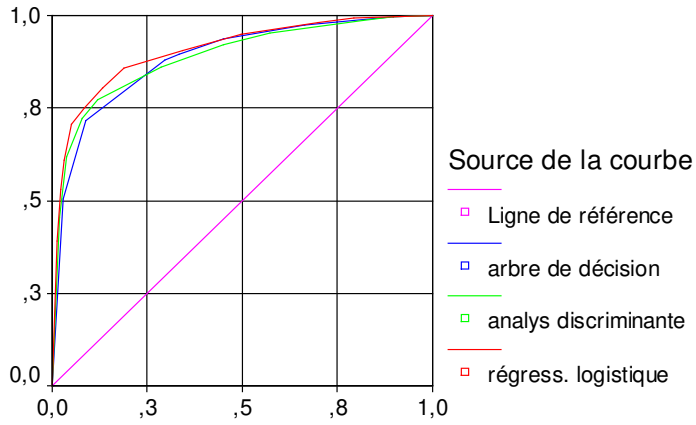
- On écrit $g(E(Y/X=x)) = \beta_0 + \Sigma_i f_i(x_i)$
- g : fonction de lien (g^{-1} : fonction de transfert)
- f_i : fonction quelconque (non-paramétrique : on n'a plus un simple paramètre comme le coefficient β_i) de x_i
  - par exemple : f_i = fonction spline
- Mais le modèle reste additif (c'est Σ_i qui combine les f_i)
- La distribution de Y peut être normale, de Poisson ou binomiale
  - ex : modèle logistique additif généralisé si $g(\mu) = \log(\mu/1-\mu)$
- Modélisation puissante mais malaisée à programmer, et attention au sur-apprentissage et à l'interprétabilité des résultats
  - surtout réservée à l'exploration des données et l'analyse des relations entre la variable à expliquer et les variables explicatives
- Source : Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability, volume 43, Chapman & Hall
- Packages R : `gam`, `mgcv` et `VGAM`

Mesures de performance

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 $\geq$ 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 $\geq$ s / B)
  - faux positifs : négatifs considérés comme positifs à cause de leur score
- Le meilleur modèle : permet de détecter le plus possible de vrais positifs avec le moins possible de faux positifs

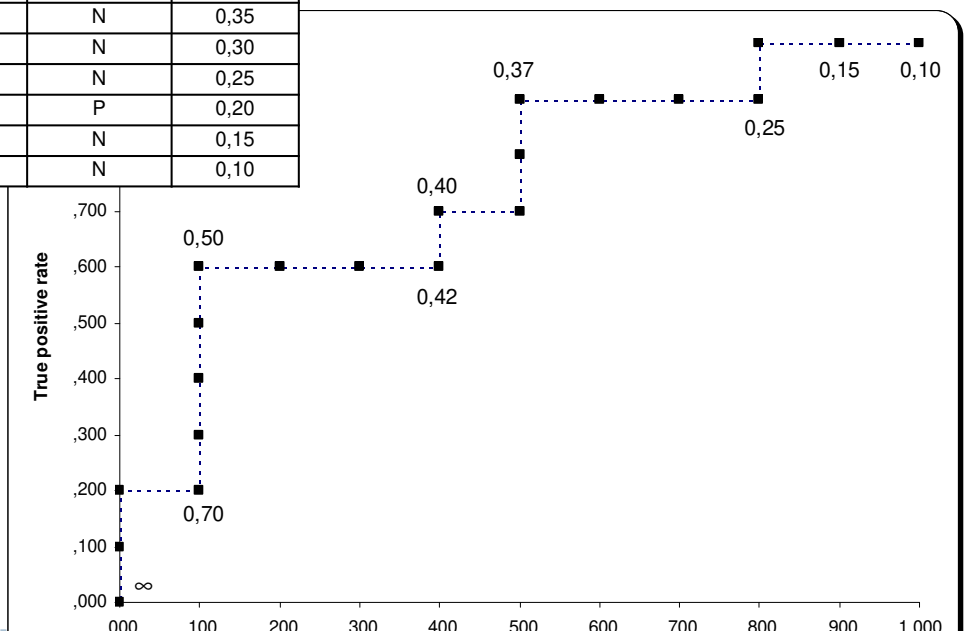
Courbe ROC



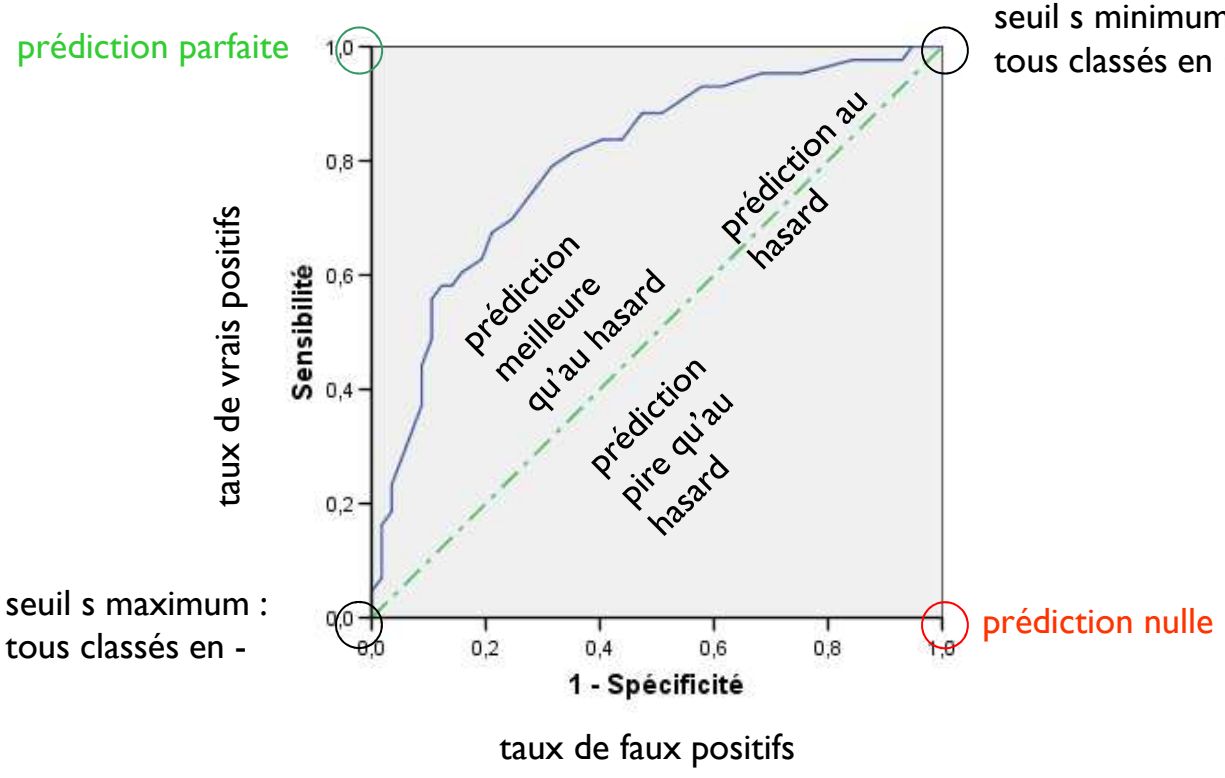
- La courbe ROC
  - sur l'axe Y : sensibilité = $\alpha(s)$
  - sur l'axe X : 1 - spécificité = 1 - $\beta(s)$
  - proportion y de vrais positifs en fonction de la proportion x de faux positifs, lorsque l'on fait varier le seuil s du score
- Aire AUC sous la courbe ROC = probabilité que score(x) > score(y), 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
  - 3e méthode équivalente : par le test de Mann-Whitney
- Le modèle est d'autant meilleur que l'AUC s'approche de 1
- AUC = 0,5 $\Rightarrow$ modèle pas meilleur qu'une notation aléatoire

Exemple de 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



Interprétation de la courbe ROC



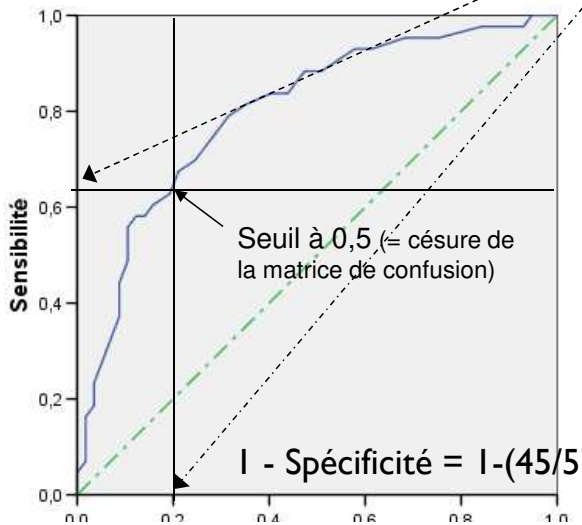
Matrice de confusion et courbe ROC

Tableau de classement[a]

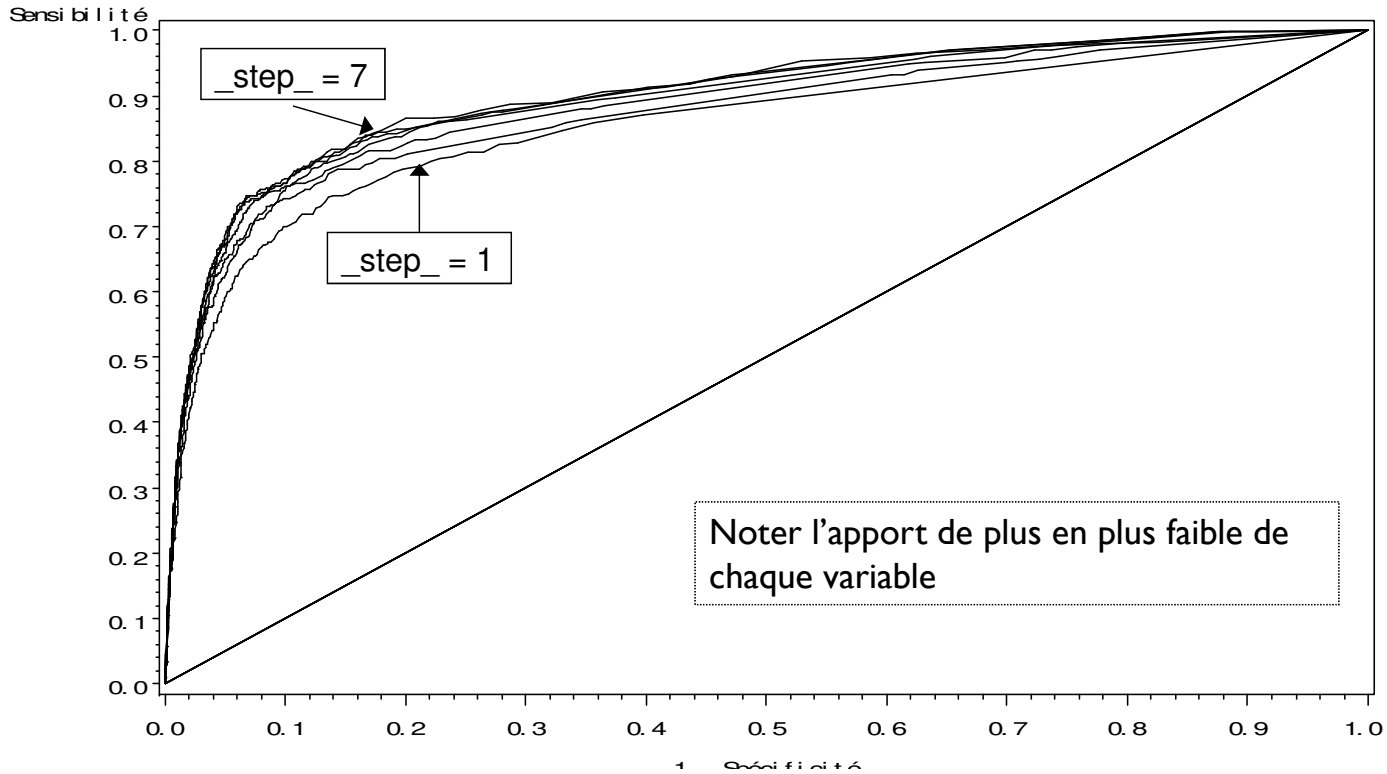
Observé		Prévu		
		CHD		Pourcentage correct
		0	1	
CHD	0	45	12	78,9
	1	16	27	62,8
Pourcentage global				72,0

a. La valeur de césure est ,500

Sensibilité = 27/43 = 0,63

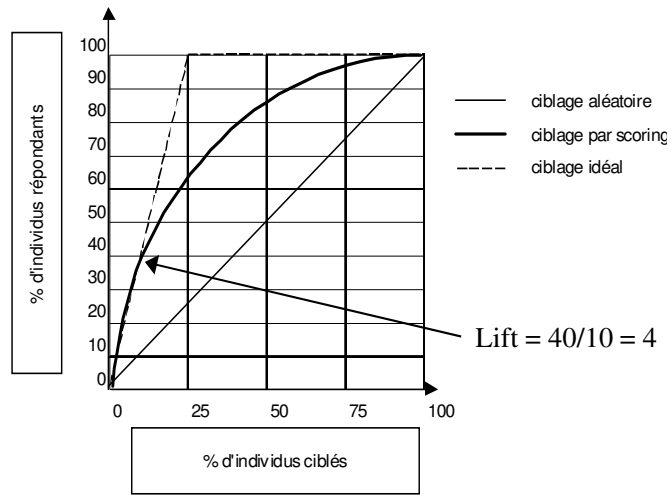


Courbes ROC avec entrée progressive des variables du modèle



Courbe de lift

- La courbe de lift :
  - sur l'axe Y : sensibilité = $\alpha(s)$ = Prob(score $\geq$ s / A)
  - sur l'axe X : Prob(score $\geq$ s)
  - proportion de vrais positifs en fonction de la proportion d'individus sélectionnés, en faisant varier le seuil s du score
  - même ordonnée que la courbe ROC, mais une abscisse généralement plus grande
  - > la courbe de lift est en général sous la courbe ROC
- Très utilisée en marketing
- Équivalence des critères
  - $AUC_1 > AUC_2 \Leftrightarrow AUL_1 > AUL_2$



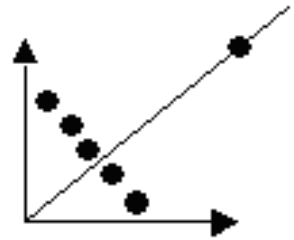
Lien entre courbe de lift et ROC

- Relation entre l'aire AUL sous la courbe de lift et l'aire AUC :
  - $AUC - AUL = p(AUC - 0{,}5) \Leftrightarrow AUL = p/2 + (1 - p)AUC$
  - où p = Proba(A) = probabilité *a priori* de l'événement dans la population (= 0,25 dans l'exemple précédent)
- Cas particuliers :
  - $AUC = 1 \Rightarrow AUL = p/2 + (1 - p) = 1 - p/2$ (modèle parfait !)
  - $AUC = 0{,}5 \Rightarrow AUL = p/2 + 1/2 - p/2 = 0{,}5$
  - p petit $\Rightarrow$ AUC et AUL sont proches
  - $AUC_1 > AUC_2 \Leftrightarrow AUL_1 > AUL_2$
- Ces indicateurs sont des critères universels de comparaison de modèles
- Indice Gini $= \dfrac{\text{surface entre la courbe de lift réelle et la diagonale}}{\text{surface entre la courbe de lift parfaite et la diagonale}}$

$$= (AUL - 0{,}5) / (1 - p/2 - 0{,}5) = 2.AUC - 1$$

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 : $x = 1, 2, 3\ldots$ et $y = e^1, e^2, e^3\ldots$

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)

moins puissant

(*) 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 $\Rightarrow$ on détecte à tort des différences significatives (au seuil de 95 % : dans 27 % des cas pour 4 moyennes égales)

Tableau ANOVA et statistique F

Source de variation	Somme des carrés (SC)	Degrés de liberté (*dl*)	Carré moyen (CM)	*F*
Totale	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2$	*n - 1*	*SC/dl*	
Inter-classe	$\sum_{i=1}^{k} n_i(\overline{Y}_i-\overline{Y})^2$	*k - 1*	*SC/dl*	$\frac{CM_{interclasse}}{CM_{intraclasse}}$
Intra-classe	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$	*n - k*	*SC/dl*	

CM_{inter}/CM_{intra} = F à comparer au F d'une loi de Fisher de ddl (k-1,n-k)

$\eta^2 = SC_{interclasse} / SC_{totale}$ = proportion de la variance expliquée

Principe du test ANOVA

- On appelle « analyse de la variance » ce qui est en fait un test d'égalité de la moyenne, en raison de la façon de réaliser ce test, qui consiste à décomposer la variance de la variable continue Y en 2 parties :
  - ce qui peut être attribué aux différences entre groupes (variance inter-classe)
  - ce qui peut être attribué aux variations à l'intérieur des classes (variance intra-classe)
- Si CM_{inter}/CM_{intra} = est grand, c'est-à-dire si les variations intra-classes sont faibles par rapport à l'effet des différences entre classes, on peut rejeter H_0 (égalité des moyennes)
- Cela se produit quand CM_{inter}/CM_{intra} dépasse la valeur critique de la loi de Fisher au niveau α avec k–1 et n–k degrés de liberté

Statistique de Mann-Whitney

- Utilisée pour k = 2 groupes, d'effectifs n_1 et n_2
  - quand les hypothèses de normalité et d'égalité des variances ne sont pas satisfaites
- Soit R_i = somme des rangs des observations du groupe i
- La statistique du test comparée à une valeur théorique est :

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

- Avec les observations des 2 groupes G_1 et G_2 :

  G_1 : 3 5 6 10 14

  G_2 : 8 12 16 18

- On obtient les rangs

  G_1 : 1 2 3 5 7 G_2 : 4 6 8 9

- D'où $R_1 = 18, R_2 = 27, U = \min(20+15-18, 20+10-27) = 3$
  - il n'y a que 3 paires où un élément de G_2 est < à un élément de G_1

U_1 = nombre de fois où une valeur du groupe 1 < une valeur du groupe 2

Test non-paramétrique de Wilcoxon-Mann-Whitney

- Statistique de la somme des rangs de Wilcoxon $S = R_i$
  - où i est soit le 1er groupe, soit le plus petit groupe
- Les groupes sont d'autant plus significativement différents :
  - que le U de Mann-Whitney est petit
  - que le S de Wilcoxon est très grand ou très petit
- À chacune de ces statistiques est associé un test dont l'hypothèse nulle H_0 est que les rangs du groupe 1 ne diffèrent pas des rangs du groupe 2
  - les tests sont équivalents $\Rightarrow$ test de Wilcoxon-Mann-Whitney
- On peut :
  - comparer U et S à des valeurs lues en table
  - ou, si n_1 et $n_2 > 8$, utiliser la convergence sous H_0 vers une loi normale $N(\mu,\sigma)$ et calculer $Z = (U-\mu)/\sigma$ et $|Z|$

Test non-paramétrique de Kruskal-Wallis

- Utilisé pour $k \geq 2$ groupes
  - quand les hypothèses de normalité et d'égalité des variances ne sont pas satisfaites
- Soient N = nombre d'observations, n_i l'effectif du groupe i et R_i la somme des rangs des observations du groupe i
- La statistique du test est :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

- Correctif à apporter en cas d'égalités de rangs
- Si les effectifs sont grands ou si $k > 6$, H tend sous H_0 vers χ^2 à $k-1$ degrés de liberté
  - sinon, regarder les valeurs critiques dans une table

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
  - indépendamment du nombre de modalités et de l'effectif
  - 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)

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
$\chi^2 = 3$			
Probabilité du χ^2 = 0,08326454			
V de Cramer = 0,14142136			

	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
$\chi^2 = 30$			
Probabilité du χ^2 = $4{,}3205.10^{-8}$			
V de Cramer = 0,14142136			

- Quand la taille de la population augmente, le moindre écart finit par devenir significatif aux seuils usuels

L'indicateur gamma de Goodman et Kruskal

- La statistique γ de Goodman et Kruskal est une mesure de liaison non paramétrique de deux variables ordinales U et V
- Pour chaque paire (i,j) d'observations, on note S+ (resp. S–) le nombre de paires pour lesquelles [U(i)–U(j)][V(i)–V(j)] > 0 (resp. < 0) (les ex aequo ne sont pas pris en compte) $\Rightarrow \gamma = (S+ - S-)/(S+ + S-)$
- On peut donc utiliser la fonction suivante de R :

```
> goodman <- function(x,y){
+     Rx <- outer(x,x,function(u,v) sign(u-v))
+     Ry <- outer(y,y,function(u,v) sign(u-v))
+     S1 <- Rx*Ry
+     return(sum(S1)/sum(abs(S1)))
+   }
> goodman(as.numeric(credit$Comptes), as.numeric(credit$Cible))
[1] -0.5546812
```

- On peut aussi utiliser le package `vcdExtra` qui donne un intervalle de confiance

```
> library(vcdExtra)
> GKgamma(table(credit$Comptes, credit$Cible), level = 0.95)
gamma        : -0.555  # gamma < 0 car le risque diminue quand le solde du compte augmente
std. error   : 0.039
CI           : -0.631 -0.478
# comparaison avec le V de Cramer
> cramer.v(table(credit$Comptes, credit$Cible))
[1] 0.3517399
```

Weight of evidence

- On utilise souvent en scoring la notion de « weight of evidence », définie pour chaque modalité d'une variable explicative :
  - g_i = la proportion de tous les clients sans impayé qui sont dans la modalité i
  - q_i = la proportion de tous les clients avec impayés qui sont dans la modalité i
- On a WOE_i = logarithme népérien du rapport g_i/q_i
- Le WOE est > 0 (resp. < 0) pour les modalités moins (resp. plus) risquées que la moyenne

```
> woe <- function(X,Y){
+ tab <- table(X,Y)
+ woe <- log((tab[,1]/sum(tab[,1])) /(tab[,2]/sum(tab[,2])))
+ return(woe)
+ }
> woe(credit$Comptes,credit$Cible)
CC [0-200 euros[     CC < 0 euros   CC > 200 euros    Pas de compte
      -0.4013918       -0.8180987        0.4054651        1.1762632
```

- On modélise parfois en recodant les variables discrètes ou qualitatives par leur WOE, ce qui permet de passer à une variable continue

La valeur d'information

- À partir du WOE on peut définir la valeur d'information d'une modalité en prenant en compte l'écart absolu et pas seulement relatif (WOE) des proportions
  - VI = $(g_i - q_i) \times WOEi$
- On définit la valeur d'information d'une variable comme la somme des VI de ses modalités (éventuellement multipliée par 100)

```
> IV <- function(X,Y){
+ tab <- table(X,Y)
+ IV <- 100*sum(((tab[,1]/sum(tab[,1])) - (tab[,2]/sum(tab[,2]))) *
log((tab[,1]/sum(tab[,1])) /(tab[,2]/sum(tab[,2]))))
+ return(IV)
+ }
> IV(credit$Comptes,credit$Cible)
[1] 66.60115
```

- On utilise parfois la grille d'interprétation suivante du pouvoir discriminant de la variable
  - VI ≤ 2 : variable non discriminante
  - VI entre 2 et 10 : faible pouvoir discriminant
  - VI entre 10 et 30 : pouvoir discriminant moyen
  - VI > 30 : pouvoir discriminant élevé

Sélection des variables : bootstrap I

- Une fonction ajuste un modèle logit sur un ensemble d'observations de l'échantillon et indique la significativité à 95 % de chaque coefficient

```
> fonction <- function(data, i){
+   d <- data[i,]
+   logit <- glm(Cible ~ ., data=d, family=binomial(link = "logit"))
+   return((summary(logit)$coefficients[,4] < 0.05))
+ }
```

- On effectue un tirage de R échantillons bootstrap en calculant la fonction précédente sur chaque échantillon

```
> library(boot)
> set.seed(123)
> resultat <- boot(data=credit, statistic=fonction, R=100)
```

- La composante `resultat$t0` = résultat de la fonction sur l'échantillon entier, et la matrice `resultat$t` contient une colonne par coefficient et une ligne par échantillon bootstrap, indiquant si le coefficient est significatif dans cet échantillon

```
> apply(resultat$t,2,sum)
 [1]  15  36  69 100  91  10  18  38  87 100  48  72  94   5   8   7   6  44  76
28  18  86  99  11  18  50   9  26  48  95  12  53  20  16  75  64  35  23  15
8  41  33   9  82  43  27  48  14   4  11  11   9  17  25
```

Sélection des variables : bootstrap II

- On affiche la liste des coefficients avec leur significativité dans la population entière et le nombre d'échantillons bootstrap où ils sont significatifs

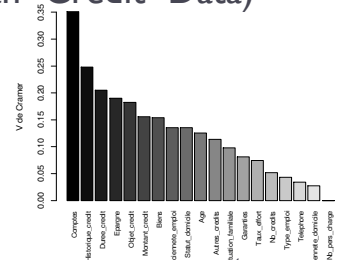
```
> as.matrix(cbind(resultat$t0, apply(resultat$t,2,sum)))
                        [,1] [,2]
(Intercept)                0   15
ComptesA12                 0   36
ComptesA13                 1   69
ComptesA14                 1  100
Duree_credit               1   91
Historique_creditA31       0   10
Historique_creditA32       0   18
Historique_creditA33       0   38
Historique_creditA34       1   87
…
```

- Noter que les modalités de référence n'apparaissent pas

Préparation des données
German Credit Data

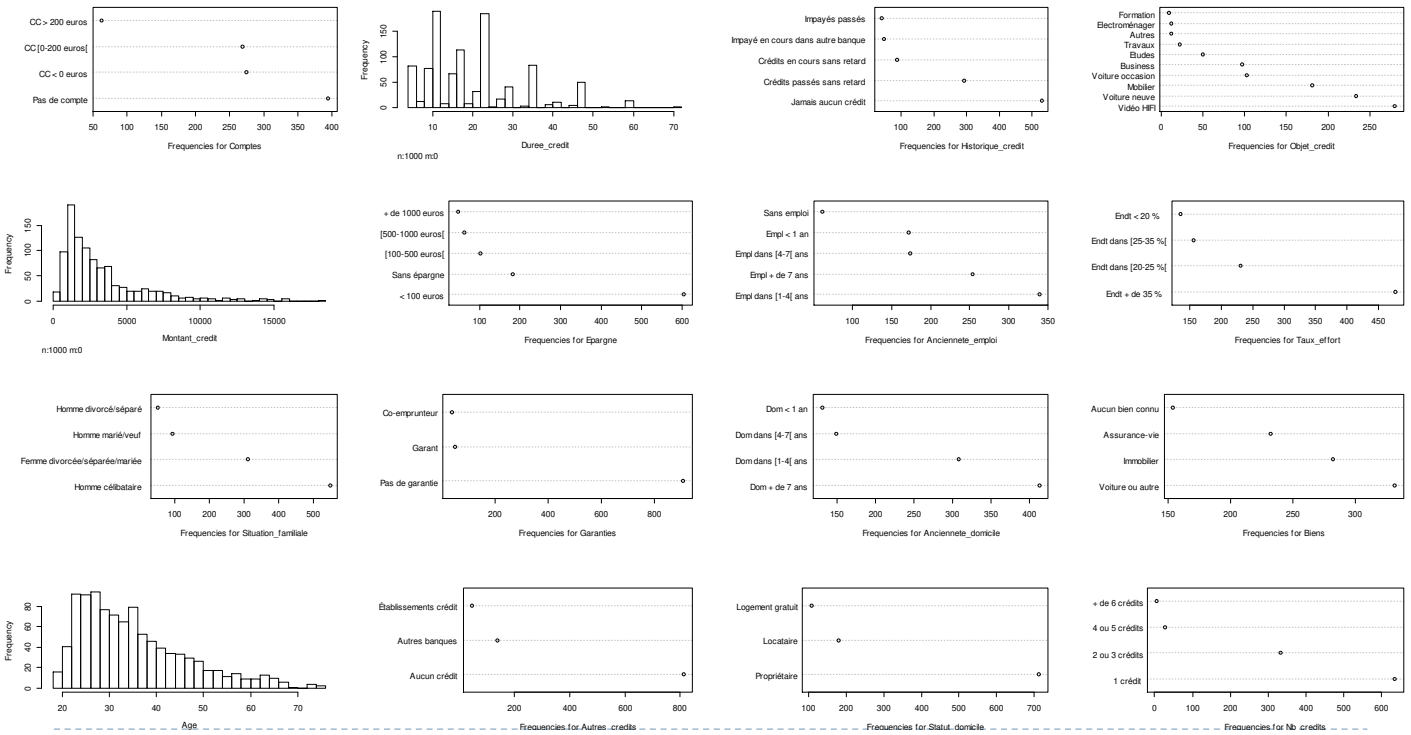
German credit data : description

- Demandes de crédits à la consommation décrites par des prédicteurs et bien remboursés ou non
- Librement disponible, par exemple ici :
  - https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
- 1000 dossiers de crédit
  - 700 bons
  - 300 mauvais (impayés)
- 19 prédicteurs quantitatifs et qualitatifs
  - durée du crédit, montant du crédit, âge, solde moyen sur compte courant, épargne, nombre de crédits déjà détenus, taux d'endettement, ancienneté au domicile, nombre de personnes à charge, objet du crédit, historique de remboursement du demandeur, autres crédits détenus (hors de la banque), biens de valeur détenus par le demandeur, garanties, situation familiale, ancienneté dans l'emploi, statut au domicile, type d'emploi et téléphone
- Hétérogénéité du pouvoir discriminant des prédicteurs



Statistiques descriptives

```
> library(Hmisc)
> hist.data.frame(credit[,1:16])
```



Variables continues croisées avec la cible

- Statistiques descriptives (fonction `summary`) par groupes (0/1 : sans/avec impayé)

```
> by(credit[,c("Age","Duree_credit","Montant_credit")], list(Cible=credit$Cible), summary)
Cible: 0
      Age         Duree_credit   Montant_credit
 Min.   :19.00   Min.   : 4.00   Min.   :  250
 1st Qu.:27.00   1st Qu.:12.00   1st Qu.: 1376
 Median :34.00   Median :18.00   Median : 2244
 Mean   :36.22   Mean   :19.21   Mean   : 2985
 3rd Qu.:42.25   3rd Qu.:24.00   3rd Qu.: 3635
 Max.   :75.00   Max.   :60.00   Max.   :15857
------------------------------------------------------------------------------------------------
Cible: 1
      Age         Duree_credit   Montant_credit
 Min.   :19.00   Min.   : 6.00   Min.   :  433
 1st Qu.:25.00   1st Qu.:12.00   1st Qu.: 1352
 Median :31.00   Median :24.00   Median : 2574
 Mean   :33.96   Mean   :24.86   Mean   : 3938
 3rd Qu.:40.00   3rd Qu.:36.00   3rd Qu.: 5142
 Max.   :74.00   Max.   :72.00   Max.   :18424
```

- Test de Kruskal-Wallis

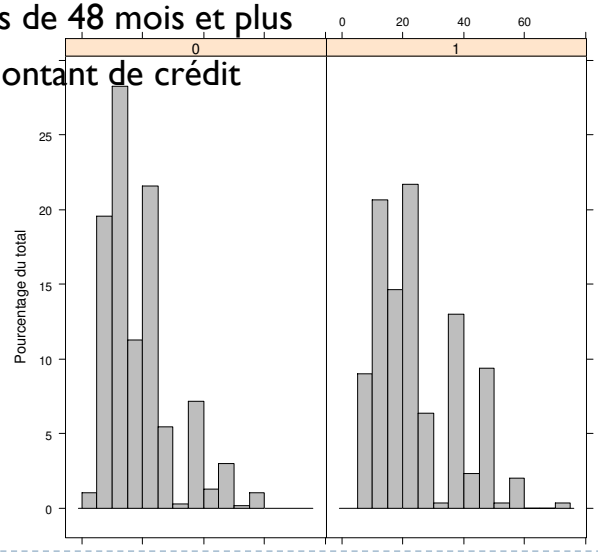
```
> kruskal.test(credit$Age~credit$Cible)$statistic
Kruskal-Wallis chi-squared
                  12.57424
> kruskal.test(credit$Duree_credit~credit$Cible)$statistic
Kruskal-Wallis chi-squared
                  42.26386
> kruskal.test(credit$Montant_credit~credit$Cible)$statistic
Kruskal-Wallis chi-squared
                   7.57588
```

Variables continues croisées avec la cible

- Histogrammes par groupes

```
> library(lattice)
> histogram(~Duree_credit | Cible , data = credit, type="percent", col="grey",
breaks=10)
```

- La durée du crédit présente des pics prévisibles à 12, 24, 36, 48 et 60 mois. On constate assez nettement la plus forte proportion de crédits plus longs parmi ceux qui ont des impayés, particulièrement les crédits de 48 mois et plus
- Mêmes analyses possibles pour l'âge et le montant de crédit



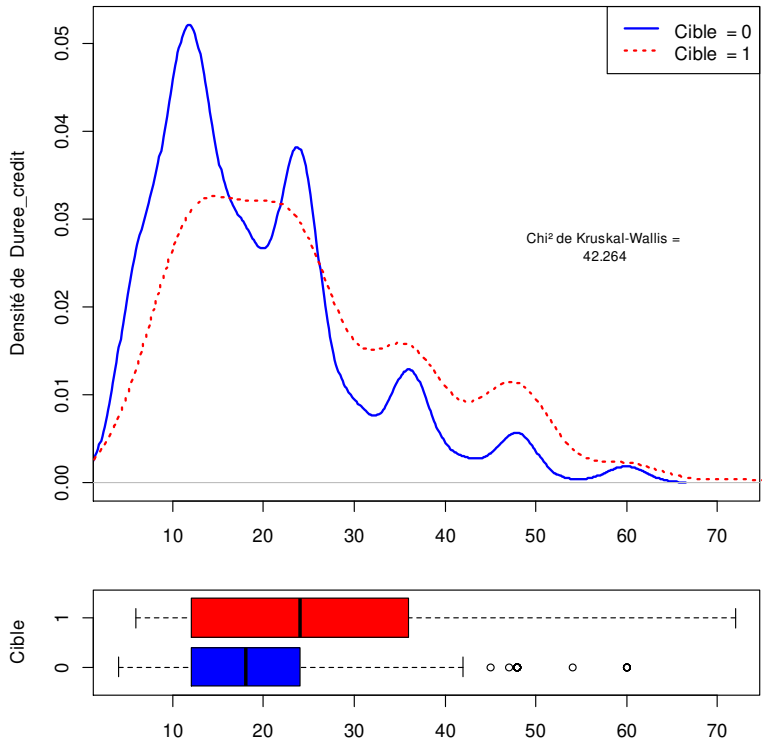
Variables continues croisées avec la cible

- Superposition des fonctions de densité et des boîtes à moustaches des bons et mauvais dossiers

```
> varQuanti = function(base,y,x)
+ {
+ old <- par(no.readonly = TRUE)
+ layout(matrix(c(1, 2)), heights=c(3, 1))
+ par(mar = c(2, 4, 2, 1))
+ base0 <- base[base[,y]==0,]
+ base1 <- base[base[,y]==1,]
+ xlim1 <- range(c(base0[,x],base1[,x]))
+ ylim1 <- c(0,max(max(density(base0[,x])$y),max(density(base1[,x])$y)))
+ plot(density(base0[,x]),main=" ",col="blue",ylab=paste("Densité de ",x),
+ xlim = xlim1, ylim = ylim1 ,lwd=2)
+ lines(density(base1[,x]),col="red",lty=3,lwd=2,
+ xlim = xlim1, ylim = ylim1,xlab = '', ylab = '',main=' ')
+ legend("topright",c(paste(y," = 0"),paste(y," = 1")), lty=c(1,3), col=c("blue","red"),lwd=2)
+ texte <- c("Chi² de Kruskal-Wallis = \n\n",
round(kruskal.test(base[,x]~base[,y])$statistic,digits=3))
+ text(xlim1[2]*0.8, ylim1[2]*0.5, texte,cex=0.75)
+ plot(base[,x]~base[,y], horizontal = TRUE, xlab= y, col=c("blue","red"))
+ par(old)
+ }
> varQuanti(credit,"Cible","Duree_credit")
```

Variables continues croisées avec la cible

- Superposition des densités et des boîtes à moustaches des bons et mauvais dossiers



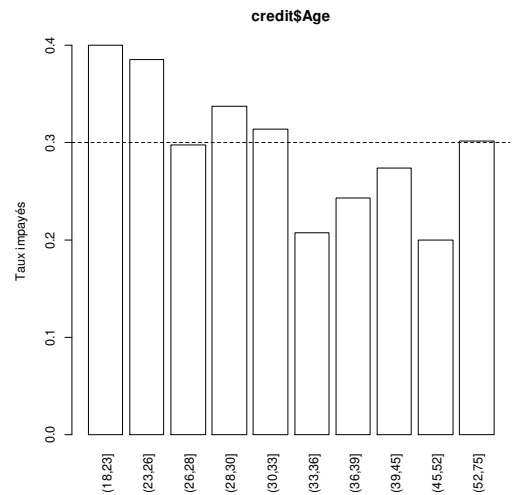
La fonction `density` ne restitue pas la fonction de densité empirique, mais le produit de convolution de cette densité empirique avec la densité de probabilité d'une loi qui est par défaut la loi normale (convolution avec un noyau gaussien)

Plus précisément, la distribution empirique et la distribution théorique sont discrétisées sur 512 points et une convolution discrète est effectuée

Taux d'impayés par quantiles

- Calcul et affichage du taux d'impayés par quantiles d'une fonction continue

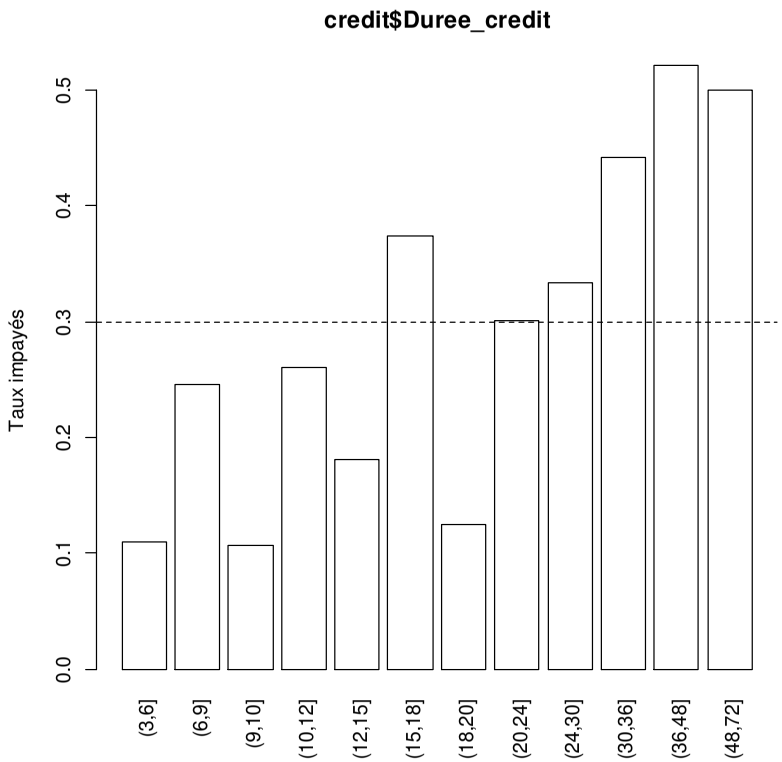
```
> ti = function(x, pas=0.1)
+ {
+ q <- unique(quantile(x, seq(0, 1, by=pas))) # fonction « unique » pour supprimer les doublons
+ qx <- cut(x, q, include.lowest=TRUE)
+ tab <- table(qx, credit$Cible)
+ print(prop.table(tab,1)) # affichage % en ligne
+ barplot(prop.table(tab,1)[,2],las=3,main=deparse(substitute(x)),ylab="Taux impayés",density=0,horiz=F)
+ abline(h=prop.table(table(credit$Cible))[2],lty=2)
+ }
> ti(credit$Age)

qx                0         1
  (18,23] 0.6000000 0.4000000
  (23,26] 0.6148148 0.3851852
  (26,28] 0.7021277 0.2978723
  (28,30] 0.6623377 0.3376623
  (30,33] 0.6857143 0.3142857
  (33,36] 0.7927928 0.2072072
  (36,39] 0.7567568 0.2432432
  (39,45] 0.7256637 0.2743363
  (45,52] 0.8000000 0.2000000
  (52,75] 0.6979167 0.3020833
```



Taux d'impayés par qua

- On détecte des seuils à 25 ou 26 ans pou
  pour la durée de crédit, et 3970 euros pou
- Les variables sont discrétisées en conséqu

```
> ti(credit$Duree_credit,pas=0.05)
> ti(credit$Montant_credit,pas=0.05
```



Discrétisations des variables continues

- Les variables sont discrétisées et les taux d'impayés par classes calculés

```
> credit$Age <- cut(credit$Age,c(0,25,Inf),right=TRUE) #intervalle semi-fermé à droite
> tab <- table(credit$Age,credit$Cible)
> prop.table(tab,1)

                  0         1
  (0,25]   0.5789474 0.4210526
  (25,Inf] 0.7283951 0.2716049
> credit$Duree_credit <- cut(credit$Duree_credit,c(0,15,36,Inf),right=TRUE)
> tab <- table(credit$Duree_credit,credit$Cible)
> prop.table(tab,1)

                  0         1
  (0,15]   0.7935035 0.2064965
  (15,36]  0.6556017 0.3443983
  (36,Inf] 0.4827586 0.5172414
> credit$Montant_credit <- cut(credit$Montant_credit,c(0,4000,Inf),right=TRUE)
> tab <- table(credit$Montant_credit,credit$Cible)
> prop.table(tab,1)

                     0         1
  (0,4e+03]   0.7413793 0.2586207
  (4e+03,Inf] 0.5731707 0.4268293
```

Liaison des variables explicatives avec la variable à expliquer I

- Calcul de la valeur d'information
- Calcul du V de Cramer = racine carrée de (chi^2 / chi^2_{max}) = racine carrée de (chi^2 / effectif)

```
> cramer  <- data.frame(NA,ncol(credit),4)
> effectif <- dim(credit)[1]
> for (i in (1:ncol(credit)))
+ {   cramer[i,1] <- names(credit[i])
+                 cramer[i,2] <-
sqrt(chisq.test(table(credit[,i],credit$Cible))$statistic/effectif)
+                 cramer[i,3] <- chisq.test(table(credit[,i],credit$Cible))$p.value
+                 cramer[i,4] <- IV(credit[,i],credit$Cible)}
> colnames(cramer) <- c("variable", "V de Cramer", "p-value chi2", "Valeur
d'information")
> vcramer <- cramer [order(cramer[,4], decreasing=T),]
```

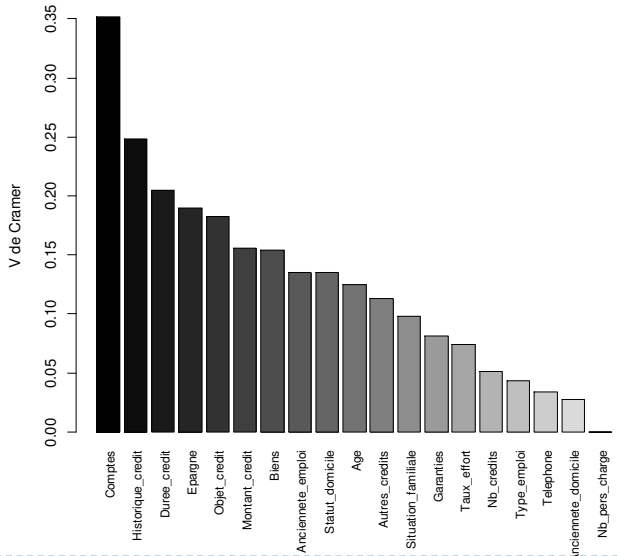
Liaison des variables explicatives avec la variable à expliquer II

```
> vcramer
                variable V de Cramer  p-value chi2 Valeur d'information
20                 Cible  0.99761905 1.941344e-218                   Inf
1                Comptes  0.35173988  1.218902e-26          66.601150335
3     Historique_credit   0.24837753  1.279187e-12          29.323354739
2          Duree_credit   0.20498754  7.507527e-10          19.894498845
6                Epargne  0.18999718  2.761214e-07          19.600955690
4           Objet_credit  0.18263747  1.157491e-04          16.919506567
5        Montant_credit   0.15555196  8.699412e-07          11.266919848
12                Biens   0.15401153  2.858442e-05          11.263826241
7     Anciennete_emploi   0.13552961  1.045452e-03           8.643363103
15      Statut_domicile   0.13490679  1.116747e-04           8.329343362
13                  Age   0.12515654  7.564407e-05           7.316642365
14       Autres_credits   0.11331014  1.629318e-03           5.761454196
9   Situation_familiale   0.09800619  2.223801e-02           4.467067763
10             Garanties  0.08151912  3.605595e-02           3.201932202
8            Taux_effort  0.07400535  1.400333e-01           2.632209005
16            Nb_credits  0.05168364  4.451441e-01           1.326652424
17           Type_emploi  0.04341838  5.965816e-01           0.876276571
19             Telephone  0.03424264  2.788762e-01           0.637760503
11 Anciennete_domicile    0.02737328  8.615521e-01           0.358877319
18       Nb_pers_charge   0.00000000  1.000000e+00           0.004339223
```

Représentation des V de Cramer

- On affiche le graphique des V de Cramer de chaque variable explicative avec la variable à expliquer

```
> old <- par(no.readonly = TRUE)
> par(mar = c(8, 4, 4, 0))
> barplot(as.numeric(vcramer[-1,2]),col=gray(0:nrow(vcramer)/nrow(vcramer)),
+ names.arg=vcramer[-1,1], ylab='V de Cramer', ylim=c(0,0.35),cex.names = 0.8, las=3)
> par(old)
```



Représentation des valeurs d'information

- On affiche le graphique des VI de chaque variable explicative avec la variable à expliquer

```
> old <- par(no.readonly = TRUE)
> par(mar = c(8, 4, 4, 0))
> barplot(as.numeric(vcramer[-1,4]),col=gray(0:nrow(vcramer)/nrow(vcramer)),
+ names.arg=vcramer[-1,1], ylab="Valeur d'information", ylim=c(0,70), cex.names = 0.8,
las=3)
> par(old)
```

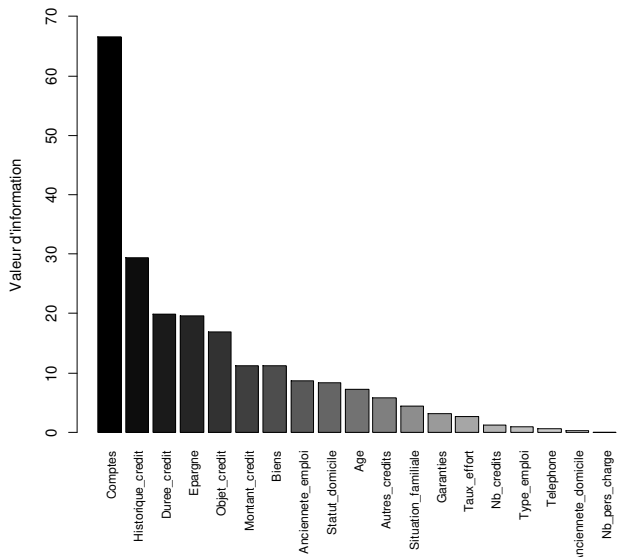


Tableau croisé comme SAS

- Le package `gmodels` permet de reproduire les tableaux croisés de SAS et SPSS

```
> CrossTable(credit$Comptes, credit$Cible, prop.chisq=F, chisq = T, format="SAS")
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
Total Observations in Table:  1000
                 | credit$Cible
  credit$Comptes |         0 |         1 | Row Total |
-----------------|-----------|-----------|-----------|
CC [0-200 euros[ |       164 |       105 |       269 |
                 |     0.610 |     0.390 |     0.269 |
                 |     0.234 |     0.350 |           |
                 |     0.164 |     0.105 |           |
-----------------|-----------|-----------|-----------|
    CC < 0 euros |       139 |       135 |       274 |
                 |     0.507 |     0.493 |     0.274 |
                 |     0.199 |     0.450 |           |
                 |     0.139 |     0.135 |           |
-----------------|-----------|-----------|-----------|
  CC > 200 euros |        49 |        14 |        63 |
                 |     0.778 |     0.222 |     0.063 |
                 |     0.070 |     0.047 |           |
                 |     0.049 |     0.014 |           |
-----------------|-----------|-----------|-----------|
   Pas de compte |       348 |        46 |       394 |
                 |     0.883 |     0.117 |     0.394 |
                 |     0.497 |     0.153 |           |
                 |     0.348 |     0.046 |           |
-----------------|-----------|-----------|-----------|
    Column Total |       700 |       300 |      1000 |
                 |     0.700 |     0.300 |           |
-----------------|-----------|-----------|-----------|
```

Taux d'impayés et effectifs par modalités

- L'examen des tableaux qui suivent permet le regroupement approprié des modalités
  - dont les taux d'impayés sont proches
  - de préférence si leurs significations sont proches
  - surtout si certaines de ces classes ont de petits effectifs

```
> ct <- function(x)
+ { cat("\n",names(credit)[x],"\n")
+ cbind(prop.table(table(credit[,x],credit$Cible),1), table(credit[,x])) }
> for (i in (1:ncol(credit))) { print(ct(i)) }

 Comptes
                          0         1
CC [0-200 euros[ 0.6096654 0.3903346 269
CC < 0 euros     0.5072993 0.4927007 274
CC > 200 euros   0.7777778 0.2222222  63
Pas de compte    0.8832487 0.1167513 394
```

- Un solde moyen sur compte courant négatif accroît le risque d'impayés
- Un solde supérieur à 200 euros diminue le taux d'impayés de plus de moitié
- Les taux d'impayés par modalités présentent une grande amplitude qui augure un fort pouvoir discriminant de cette variable

Taux d'impayés de l'historique de crédit

```
 Historique_credit
                                          0         1
Crédits en cours sans retard      0.6818182 0.3181818  88
Crédits passés sans retard        0.8293515 0.1706485 293
Impayé en cours dans autre banque 0.4285714 0.5714286  49
Impayés passés                    0.3750000 0.6250000  40
Jamais aucun crédit               0.6811321 0.3188679 530
```

- Taux d'impayés logiquement liés à l'historique de remboursement du demandeur
- Nette gradation depuis le demandeur qui a déjà eu des crédits qu'il a bien remboursés jusqu'à celui qui a eu des impayés dans la banque ou en a encore dans d'autres banques
- Ne figurent pas ceux qui ont encore des impayés dans la banque elle-même, car leur demande de nouveau crédit est automatiquement rejetée
- Deux classes intermédiaires ont des taux d'impayés très proches l'un de l'autre : ceux qui ont déjà un crédit en cours de remboursement, sans retard actuellement (31,82 % d'impayés), et ceux qui n'ont jamais eu de crédit (31,89 % d'impayés). Plus précisément, ces derniers n'ont jamais eu de crédit dans la banque, et s'ils en ont eu dans d'autres banques, ils les ont tous bien remboursés puisqu'ils ne sont pas fichés. Des taux d'impayés proches, un sens métier proche (des clients dont on ne connaît pas complètement le comportement de débiteur), une des classes qui représente moins de 9 % des dossiers : trois raisons de regrouper ces deux classes

Taux d'impayés de l'objet du crédit

```
 Objet_credit
                            0         1
Autres             0.5833333 0.4166667  12
Business           0.6494845 0.3505155  97
Electroménager     0.6666667 0.3333333  12
Etudes             0.5600000 0.4400000  50
Formation          0.8888889 0.1111111   9
Mobilier           0.6795580 0.3204420 181
Travaux            0.6363636 0.3636364  22
Vidéo HIFI         0.7785714 0.2214286 280
Voiture neuve      0.6196581 0.3803419 234
Voiture occasion   0.8349515 0.1650485 103
```

- Les modalités de l'objet de crédit sont trop nombreuses, et parfois trop petites
- Il faut donc opérer des regroupements : il est logique, et conforme aux taux d'impayés, de regrouper les objets « mobiliers », « électro-ménager » et « travaux »
- Nous regroupons aussi l'objet « formation » avec « études », en dépit d'un taux d'impayés plus bas pour la « formation », qui ne signifie rien car il repose sur un seul impayé. Nous leur adjoignons aussi la modalité « Autres » dont le taux d'impayés est proche. On peut aussi envisager de regrouper cette modalité avec « business »
- Taux d'impayés du simple au double entre les voitures d'occasion et les voitures neuves
- On peut noter que l'objet est toujours défini, et que les crédits accordés sont donc des crédits « affectés », et non des crédits personnels non affectés ou des crédits revolving

Taux d'impayés du montant d'épargne

```
 Epargne
                            0         1
[100-500 euros[  0.6699029 0.3300971 103
[500-1000 euros[ 0.8253968 0.1746032  63
+ de 1000 euros  0.8750000 0.1250000  48
< 100 euros      0.6401327 0.3598673 603
Sans épargne     0.8251366 0.1748634 183
```

- On a un taux d'impayés faible pour ceux qui n'ont pas de produit d'épargne (ce qui est différent d'avoir un produit d'épargne avec un encours nul), ce qui signifie que leur épargne est dans un autre établissement bancaire
- Le taux d'impayés augmente ensuite fortement, en prenant des valeurs proches en deçà de 100 euros et entre 100 et 500 euros, modalités que l'on rapproche donc
- On peut hésiter à regrouper les deux tranches « [500 – 1000 euros [» et « ≥ 1000 euros », car leurs taux d'impayés sont quelque peu différents, mais la petite taille de ces deux tranches pousse à les regrouper

Taux d'impayés de l'ancienneté à l'emploi

```
 Anciennete_emploi
                              0         1
Empl + de 7 ans       0.7470356 0.2529644 253
Empl < 1 an           0.5930233 0.4069767 172
Empl dans [1-4[ ans   0.6932153 0.3067847 339
Empl dans [4-7[ ans   0.7758621 0.2241379 174
Sans emploi           0.6290323 0.3709677  62
```

- L'ancienneté à l'emploi suit une tendance générale logique : un demandeur plus ancien présente moins de risque
- Mais certains points sont un peu étonnants et viennent peut-être d'un nombre d'impayés trop faible pour assurer des taux d'impayés parfaitement fiables
- Ils peuvent peut-être aussi s'expliquer autrement. Ainsi, une personne sans emploi est légèrement moins risquée qu'une personne travaillant depuis moins d'un an. Cela peut venir du fait que la catégorie « sans emploi » regroupe non seulement des demandeurs d'emploi, mais aussi des personnes n'ayant financièrement pas besoin de travailler, voire des retraités (non identifiés par ailleurs dans le jeu de données)
- Quant à la catégorie des personnes dans leur emploi depuis plus de sept ans, elle est un peu plus risquée que celle qui y est depuis quatre à sept ans. Cela peut s'expliquer par la présence de salariés âgés, parfois plus souvent victimes d'une perte d'emploi
- Quoi qu'il en soit, pour éviter d'éventuelles incohérences et aussi pour s'assurer des modalités suffisamment importantes, on regroupera ces modalités

Taux d'impayés du taux d'effort

```
 Taux_effort
                                0         1
Endt + de 35 %          0.6659664 0.3340336 476
Endt < 20 %             0.7500000 0.2500000 136
Endt dans [20-25 %[     0.7316017 0.2683983 231
Endt dans [25-35 %[     0.7133758 0.2866242 157
```

- La vue du tableau sur le taux d'effort (taux d'endettement) corrobore un fait bien connu des spécialistes du *credit scoring* : cette variable qui semble populaire auprès de certains analystes de crédit est en réalité peu prédictive du risque d'impayés, du moins après l'exclusion probable avant scoring des demandeurs les plus endettés
- Les taux d'impayés sont si proches qu'il est impossible que cette variable puisse être d'un quelconque intérêt, et nous l'ôterons de la sélection

Taux d'impayés de la situation de famille

```
 Situation_familiale
                                        0         1
Femme divorcée/séparée/mariée 0.6483871 0.3516129 310
Homme célibataire             0.7335766 0.2664234 548
Homme divorcé/séparé          0.6000000 0.4000000  50
Homme marié/veuf              0.7282609 0.2717391  92
```

- Faible écart entre les taux d'impayés des différentes modalités : les deux dont les taux d'impayés sont les plus proches seront regroupées, mais la variable aura du mal à jouer un rôle utile dans la prédiction. La faiblesse de cette variable est souvent constatée en scoring.
- Il est dommage que la modalité « femme divorcée/séparée » ait été d'emblée regroupée avec la modalité « femme mariée » dont le taux d'impayés est probablement inférieur
- On note l'absence un peu étonnante de la modalité « femme célibataire » dans l'échantillon : le modèle de score sera donc élaboré sans cette variable et se montrera incapable de noter, si elle survient, une demande de crédit formulée par une femme célibataire. Pour éviter cette situation imprévue, il vaut mieux ajouter une règle *a priori* concernant les femmes célibataires, par exemple en assimilant la modalité « femme célibataire » à l'une des modalités présentes : « femme divorcée/séparée/mariée » ou « homme célibataire »
- Face à une modalité absente de l'échantillon de modélisation, une autre façon de procéder consiste à assimiler cette modalité, non pas à la plus proche logiquement (si cela se peut déterminer) mais à la plus risquée

Taux d'impayés de la présence de garantie

```
 Garanties
                          0         1
Co-emprunteur    0.5609756 0.4390244  41
Garant           0.8076923 0.1923077  52
Pas de garantie  0.7001103 0.2998897 907
```

- L'existence d'un garant contribue à diminuer le taux d'impayés, contrairement à celle d'un co-emprunteur, mais cette situation est assez rare
- On peut tout de même tester cette variable, qui pourrait apporter un complément d'information utile pour les individus concernés. Il faut ici prendre garde au petit nombre de dossiers (52) pour lesquels un garant existe, qui doit faire considérer avec un peu de circonspection le taux d'impayés de 19,23 %
- Il est d'ailleurs connu que la présence d'un garant contribue généralement moins à faire baisser le risque d'impayés qu'à augmenter le taux de récupération en cas d'impayé. Dans la terminologie Bâle 2, nous dirons que le garant diminue plus la LGD (« loss given default ») que la PD (« probability of default »).

Taux d'impayés de l'ancienneté au domicile

```
 Anciennete_domicile
                                0         1
Dom + de 7 ans      0.6997579 0.3002421 413
Dom < 1 an          0.7230769 0.2769231 130
Dom dans [1-4[ ans  0.6850649 0.3149351 308
Dom dans [4-7[ ans  0.7114094 0.2885906 149
```

- L'ancienneté au domicile influe de façon étonnante sur le risque d'impayés, puisque les plus récents dans leur logement sont les moins risqués. De surcroît, aucune tendance ne se dégage
- On peut avoir des doutes sur la qualité du renseignement de cette variable, qui est généralement plus discriminante : est-elle seulement mise à jour lors d'un déménagement, ou ne le serait-elle pas également lors de toute fiabilisation de l'adresse ou mise à jour du numéro de téléphone ?

Taux d'impayés des biens détenus

```
 Biens
                           0         1
Assurance-vie     0.6939655 0.3060345 232
Aucun bien connu  0.5649351 0.4350649 154
Immobilier        0.7872340 0.2127660 282
Voiture ou autre  0.6927711 0.3072289 332
```

- Le lien entre le bien de plus forte valeur détenu et le risque d'impayés est logique : l'absence de bien connu double le taux d'impayés par rapport à la détention d'un bien immobilier. Un demandeur dans ce dernier cas dispose d'une meilleure assise financière, surtout s'il a fini de rembourser son bien immobilier. Et quand cela n'est pas le cas, il sera plus attentif que la moyenne au bon remboursement de ses échéances. De toute façon, si un crédit immobilier lui avait été consenti, c'est qu'il présentait une certaine fiabilité financière
- Les deux modalités intermédiaires, « assurance-vie » et « voiture ou autre » ont des taux d'impayés égaux et sont donc regroupées pour former une modalité « bien non immobilier »

Taux d'impayés de la présence de crédits à l'extérieur

```
 Autres_credits
                                0         1
Aucun crédit           0.7248157 0.2751843 814
Autres banques         0.5899281 0.4100719 139
Établissements crédit 0.5957447 0.4042553  47
```

- La détention de crédits dans d'autres établissements entraîne logiquement un plus grand risque d'impayés, car le client est plus endetté et l'on ne voit ni ne maîtrise ce qui se passe dans l'autre établissement
- Qu'il s'agisse d'un établissement bancaire classique ou d'un établissement spécialisé dans le crédit (notamment pour le crédit sur le lieu de vente) ne change en revanche rien au taux d'impayés, et l'on regroupe les deux premières modalités

Taux d'impayés du statut au domicile

```
 Statut_domicile
                           0         1
Locataire          0.6089385 0.3910615 179
Logement gratuit   0.5925926 0.4074074 108
Propriétaire       0.7391304 0.2608696 713
```

- Les propriétaires sont logiquement moins risqués et l'on peut regrouper les deux autres modalités, dont les taux d'impayés sont proches

Taux d'impayés du nombre de crédits

```
 Nb_credits
                          0         1
+ de 6 crédits 0.6666667 0.3333333   6
1 crédit       0.6840442 0.3159558 633
2 ou 3 crédits 0.7237237 0.2762763 333
4 ou 5 crédits 0.7857143 0.2142857  28
```

- Le nombre de crédits détenus dans la banque (en incluant celui en cours d'instruction) représente des effectifs trop restreints pour être modélisés à partir de quatre crédits
- Quant aux taux d'impayés dans le cas d'un crédit, et dans le cas de deux ou trois crédits, ils sont trop proches pour être distingués
- On pourrait s'étonner de voir un taux d'impayés diminuer un peu en présence de plusieurs crédits: on peut l'interpréter en supposant que la banque n'octroie plusieurs crédits qu'à des clients jugés fiables, et des difficultés rencontrées avec le premier crédit dissuadent d'en accorder d'autres
- Cette variable est peu discriminante

Taux d'impayés du type d'emploi

```
 Type_emploi
                                    0         1
Cadre                       0.6554054 0.3445946 148
Employé-Ouvrier qualifié    0.7047619 0.2952381 630
Non qualifié                0.7200000 0.2800000 200
Sans emploi                 0.6818182 0.3181818  22
```

- Le type d'emploi se révèle peu prédictif du risque d'impayés, et de surcroît pas d'une façon intuitive, puisque les cadres sont les plus risqués, ce qui n'est pas ce que l'on constate en général
- Cela tient peut-être aussi au fait que cette modalité contient aussi les professions libérales, sans doute plus risquées
- On peut s'étonner de l'absence des retraités, peut-être regroupés avec les autres « sans emploi », ce qui expliquerait que le taux d'impayés de cette modalité soit à peine supérieur à la moyenne

Taux d'impayés du nombre de personnes à charge

```
 Nb_pers_charge
                        0         1
+ de 3 pers 0.7032258 0.2967742 155
0-2 pers    0.6994083 0.3005917 845
```

- Un plus grand nombre de personnes à charge n'entraîne donc pas plus de difficultés pour rembourser son crédit
- Cette variable est régulièrement testée mais est rarement significative

Taux d'impayés du téléphone déclaré

```
 Telephone
                    0         1
Avec Tél 0.7202970 0.2797030 404
Sans Tél 0.6862416 0.3137584 596
```

- L'absence de téléphone, ou plus exactement l'absence de numéro de téléphone fourni, est généralement considéré comme un facteur de risque, mais il est ici très peu marqué
- Cette variable est parfois examinée, mais n'est pas toujours très utile sous cette forme binaire : il faudrait distinguer un numéro fixe, mobile et surtout professionnel
- Quand un client a fourni un numéro de téléphone professionnel, outre que cela signale qu'il possède un emploi, cela peut indiquer une intention plus manifeste de rembourser son crédit

Liaisons entre les variables explicatives

- Calcul des V de Cramer de toutes les paires de variables explicatives

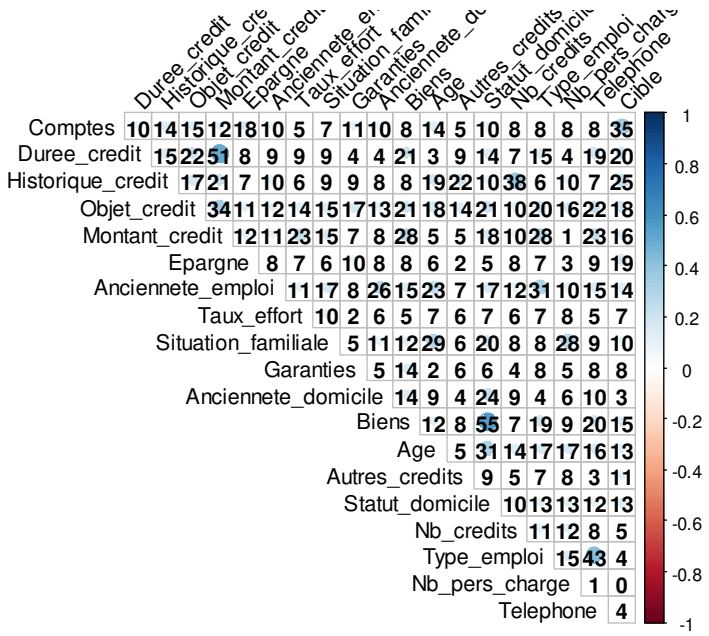
```
> library(questionr)
> cramer  <- matrix(NA,ncol(credit),ncol(credit))
> for (i in (1:ncol(credit)))
+ {     for (j in (1:ncol(credit)))
+ {
+ cramer[i,j] <- cramer.v(table(credit[,i],credit[,j]))
+ }
+ }
> colnames(cramer) <- colnames(credit)
> rownames(cramer) <- colnames(credit)
```

- Le package `corrplot` permet d'afficher des matrices de corrélation

```
> library(corrplot)
> corrplot(cramer, method="shade", shade.col=NA, tl.col="black", tl.srt=45)
> old <- par(no.readonly = TRUE)
> par(omi=c(0.4,0.4,0.4,0.4))
> corrplot(cramer, type="upper", tl.srt=45, tl.col="black", tl.cex=1, diag=F,
addCoef.col="black", addCoefasPercent=T)
> par(old)
```

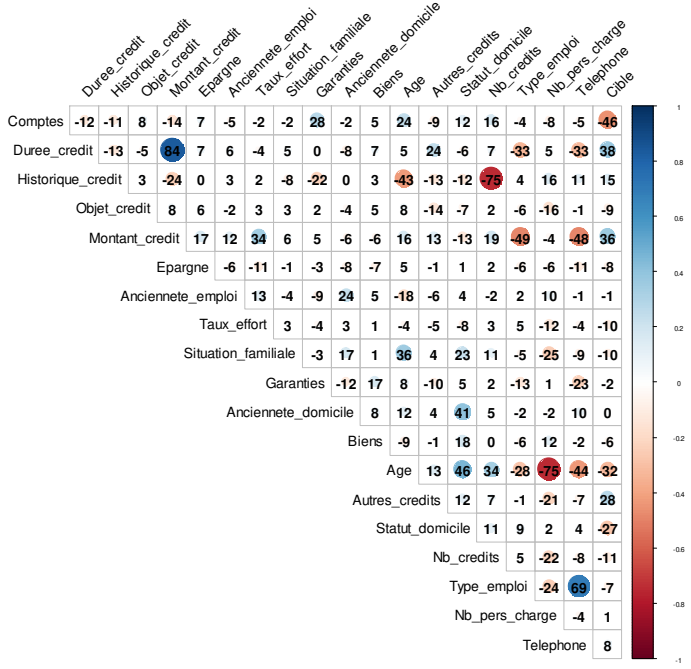
Graphique des V de Cramer entre les variables explicatives

- Options de `corrplot` pour améliorer la lisibilité :
  - `diag=F` : supprimer la diagonale
  - `type="upper"` : n'afficher que le triangle supérieur de la matrice
  - `addCoef.col="black"` : afficher les coefficients dans la matrice
  - `addCoefasPercent=T` : les afficher sous forme de % pour économiser un peu de place
  - `lt.col` et `tl.srt` : régler l'apparence des titres
- Les liaisons les plus fortes sont aisément repérables par leur couleur plus foncée :
  - biens et statut domicile
  - durée et montant de crédit
  - type d'emploi et téléphone



Graphique des Gamma de Goodman-Kruskal entre les variables explicatives

- On affiche le `corrplot` en remplaçant le V de Cramer par le Gamma de Goodman-Kruskal
- On remplace `cramer.v(table(credit[,i],credit[,j]))` par `GKgamma(table(credit[,i],credit[,j]))$gamma`
- Exemple de liaisons :
  - positives : 0,84 entre la durée et le montant de crédit car la durée augmente avec le montant
  - négatives : –0,46 entre la variable « comptes » et la variable cible, car le risque d'impayé diminue quand le solde sur le compte augmente
  - négatives : –0,32 entre l'âge et la cible car les impayés sont plus fréquents chez les jeunes



Liaisons les plus fortes

- Entre biens et statut du domicile :

```
> tab <- table(credit$Biens, credit$Statut_domicile)
> prop.table(tab,1) # affichage % en ligne
                    Locataire Logement gratuit Propriétaire
  Assurance-vie    0.198275862      0.008620690  0.793103448
  Aucun bien connu 0.116883117      0.675324675  0.207792208
  Immobilier       0.195035461      0.003546099  0.801418440
  Voiture ou autre 0.180722892      0.003012048  0.816265060
```

- Le bien de plus haute valeur n'est lié au statut du domicile que par une modalité pour chaque variable : la modalité « aucun bien connu » avec la modalité « logement gratuit ». Pour les autres modalités de la variable « Biens », la proportion de locataires, propriétaires et logés à titre gratuit est presque la même
- Entre durée et montant de crédit :

```
> tab <- table(credit$Duree_credit, credit$Montant_credit)
> prop.table(tab,1) # affichage % en ligne
             Montant_credit
Duree_credit  (0,4e+03] (4e+03,Inf]
    (0,15]    0.94663573  0.05336427
    (15,36]   0.68672199  0.31327801
    (36,Inf] 0.17241379  0.82758621
```

- L'intensité de la liaison vient de ce que 95 % des crédits d'au plus 15 mois sont d'un montant inférieur à 4 000 euros, tandis que 83 % des crédits de plus de 36 mois sont d'un montant supérieur à 4 000 euros
- Il est peu probable que les deux variables puissent figurer simultanément dans un modèle de régression, et la durée (V de Cramer = 0,20) est plus discriminante que le montant (V de Cramer = 0,16)

Choix entre la durée et le montant de crédit

- Le tableau croisé avec la cible montre que la durée offre un découpage plus équilibré de la population puisqu'aucune modalité ne dépasse la moitié des demandes de crédit
- De plus, l'une des modalités (≤ 15 mois) est nettement moins risquée que la moyenne, tandis qu'une autre (> 36 mois) est nettement plus risquée, ce qui permet de mieux répartir les dossiers entre ceux qui sont plus ou moins risqués

```
> tab <- table(credit$Duree_credit, credit$Cible)
> cbind(prop.table(tab,1), addmargins(tab,2))
                0         1   0   1 Sum
(0,15]   0.7935035 0.2064965 342  89 431
(15,36]  0.6556017 0.3443983 316 166 482
(36,Inf] 0.4827586 0.5172414  42  45  87
> tab <- table(credit$Montant_credit, credit$Cible)
> cbind(prop.table(tab,1), addmargins(tab,2))
                   0         1   0   1 Sum
(0,4e+03]   0.7413793 0.2586207 559 195 754
(4e+03,Inf] 0.5731707 0.4268293 141 105 246
```

- Le montant tire peut-être son pouvoir discriminant de son lien avec la durée, car le lien avec le risque d'impayés est plus évident pour la durée, dans la mesure où ces montants, bien plus bas que pour un crédit habitat, ne se traduisent pas par des mensualités d'ordres de grandeur différents

Liaison entre type d'emploi et téléphone

- La 3e paire de variables les plus fortement liées est « téléphone » et « type d'emploi », avec 16 % de travailleurs non qualifiés ayant fourni un numéro de téléphone, contre 86 % pour les cadres
- Nous n'aurons sans doute pas à choisir entre ces variables toutes deux peu discriminantes du risque d'impayés

```
> tab <- table(credit$Type_emploi,credit$Telephone)
> prop.table(tab,1)
```

```
                           Avec Tél  Sans Tél
Cadre                     0.8581081 0.1418919
Employé-Ouvrier qualifié 0.3809524 0.6190476
Non qualifié              0.1550000 0.8450000
Sans emploi               0.2727273 0.7272727
```

- Les autres paires de variables ont des V de Cramer < 0,40 : intensités de liaison sans doute acceptables dans un modèle de régression

Description de la base avant regroupements de modalités

```
> summary(credit)
             Comptes      Duree_credit                          Historique_credit          Objet_credit
 CC [0-200 euros[:269   (0,15]  :431   Crédits en cours sans retard     : 88     Vidéo HIFI      :280
 CC < 0 euros    :274   (15,36] :482   Crédits passés sans retard       :293     Voiture neuve   :234
 CC > 200 euros  : 63   (36,Inf]: 87   Impayé en cours dans autre banque: 49     Mobilier        :181
 Pas de compte   :394                  Impayés passés                   : 40     Voiture occasion:103
                                       Jamais aucun crédit              :530     Business        : 97
                                                                                 Etudes          : 50
                                                                                 (Other)         : 55
    Montant_credit             Epargne              Anciennete_emploi              Taux_effort
 (0,4e+03]  :754    [100-500 euros[ :103   Empl + de 7 ans    :253     Endt + de 35 %     :476
 (4e+03,Inf]:246    [500-1000 euros[: 63   Empl < 1 an        :172     Endt < 20 %        :136
                    + de 1000 euros : 48   Empl dans [1-4[ ans:339     Endt dans [20-25 %[:231
                    < 100 euros     :603   Empl dans [4-7[ ans:174     Endt dans [25-35 %[:157
                    Sans épargne    :183   Sans emploi        : 62

                    Situation_familiale           Garanties           Anciennete_domicile
 Femme divorcée/séparée/mariée:310      Co-emprunteur  : 41   Dom + de 7 ans    :413
 Homme célibataire             :548     Garant         : 52   Dom < 1 an        :130
 Homme divorcé/séparé          : 50     Pas de garantie:907   Dom dans [1-4[ ans:308
 Homme marié/veuf              : 92                           Dom dans [4-7[ ans:149

Biens           Age                   Autres_credits         Statut_domicile
 Assurance-vie   :232   (0,25]  :190   Aucun crédit         :814    Locataire       :179
 Aucun bien connu:154   (25,Inf]:810   Autres banques       :139    Logement gratuit:108
 Immobilier      :282                  Établissements crédit: 47    Propriétaire    :713
 Voiture ou autre:332

Nb_credits                   Type_emploi      Nb_pers_charge    Telephone   Cible
 + de 6 crédits:  6   Cadre                  :148   + de 3 pers:155    Avec Tél:404   0:700
 1 crédit      :633   Employé-Ouvrier qualifié:630   0-2 pers   :845    Sans Tél:596   1:300
 2 ou 3 crédits:333   Non qualifié           :200
 4 ou 5 crédits: 28   Sans emploi            : 22
```

Description de la base après regroupements de modalités

```
> summary(credit2)
            Comptes      Duree_credit                                Historique_credit           Objet_credit
 CC [0-200 euros[:269   (0,15]  :431   Crédits en impayé                    : 89     Etudes-business :168
 CC < 0 euros    :274   (15,36] :482   Crédits passés sans retard          :293     Intérieur       :495
 CC > 200 euros  : 63   (36,Inf]: 87   Pas de crédits ou en cours sans retard:618     Voiture neuve   :234
 Pas de compte   :394                                                               Voiture occasion:103
    Montant_credit                       Epargne               Anciennete_emploi             Taux_effort
 (0,4e+03]  :754    < 500 euros              :706   E [1-4[ ans          :339     Endt + de 35 %     :476
 (4e+03,Inf]:246    Pas épargne ou > 500 euros:294   E GE 4 ans           :427     Endt < 20 %        :136
                                                     Sans emploi ou < 1 an:234     Endt dans [20-25 %[:231
                                                                                   Endt dans [25-35 %[:157
                  Situation_familiale       Garanties          Anciennete_domicile            Biens
 Femme divorcée/séparée/mariée:310      Avec garant: 52   Dom + de 7 ans    :413      Aucun bien    :154
 Homme célibataire/marié/veuf :640      Sans garant:948   Dom < 1 an        :130      Immobilier    :282
 Homme divorcé/séparé         : 50                        Dom dans [1-4[ ans:308      Non immobilier:564
                                                          Dom dans [4-7[ ans:149
      Age                    Autres_credits         Statut_domicile          Nb_credits
 (0,25]  :190   Aucun crédit extérieur:814    Non Propriétaire:287    + de 6 crédits:  6
 (25,Inf]:810   Crédits extérieurs    :186    Propriétaire    :713    1 crédit      :633
                                                                     2 ou 3 crédits:333
                                                                     4 ou 5 crédits: 28
                 Type_emploi      Nb_pers_charge Telephone  Cible
 Cadre                  :148   + de 3 pers:155    A191:596   0:700
 Employé-Ouvrier qualifié:630   0-2 pers   :845    A192:404   1:300
 Non qualifié           :200
 Sans emploi            : 22
```

Échantillonnage

- Le package `sampling` permet d'effectuer un échantillonnage stratifié (srswor : simple random sampling without replacement)
- Nous stratifions sur la variable à expliquer pour constituer un échantillon d'apprentissage et un échantillon de validation

```
> library(sampling)
> set.seed(123)
> id <- strata(credit, stratanames="Cible", size=c(sum(credit$Cible==0)*2/3,sum(credit$Cible==1)*2/3),
method="srswor", description=T)$ID_unit
Stratum 1
Population total and number of selected units: 700 466.6667
Stratum 2
Population total and number of selected units: 300 200
Number of strata  2
Total number of selected units 666.6667

> train  <- credit2[id,]
> valid  <- credit2[-id,]
> table(train$Cible)/nrow(train)

        0         1
0.6996997 0.3003003
> table(valid$Cible)/nrow(valid)

        0         1
0.7005988 0.2994012
```

Régression logistique avec R

- La fonction de base est `glm` (package `stats`)

```
> logit <- glm(Cible~Comptes+Historique_credit+Duree_credit+Age+Epargne+Garanties+Autres_credits, data=train, family=binomial(link = "logit"))
> summary(logit)
Call:
glm(formula = Cible ~ Comptes + Historique_credit + Duree_credit + Age + Epargne + Garanties + Autres_credits, family = binomial(link = "logit"),
data = train)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8813  -0.7450  -0.4171   0.8348   2.7460


Coefficients:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -0.5798     0.5908  -0.981  0.32639
ComptesCC < 0 euros                                0.1584     0.2313   0.685  0.49340
ComptesCC > 200 euros                             -1.3485     0.4688  -2.876  0.00402 **
ComptesPas de compte                              -1.5137     0.2683  -5.641 1.69e-08 ***
Historique_creditCrédits passés sans retard       -1.5873     0.3652  -4.346 1.39e-05 ***
Historique_creditPas de crédits ou en cours sans retard -0.8970     0.3254  -2.757  0.00583 **
Duree_credit(15,36]                                0.5767     0.2104   2.741  0.00612 **
Duree_credit(36,Inf]                               1.5116     0.3596   4.203 2.63e-05 ***
Age(25,Inf]                                       -0.4508     0.2317  -1.946  0.05171 .
EpargnePas épargne ou > 500 euros                 -1.1000     0.2521  -4.364 1.28e-05 ***
GarantiesSans garant                               1.3195     0.4888   2.700  0.00694 **
Autres_creditsCrédits extérieurs                   0.5588     0.2441   2.289  0.02205 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 814.01  on 665  degrees of freedom
Residual deviance: 636.11  on 654  degrees of freedom
AIC: 660.11

Number of Fisher Scoring iterations: 5
```

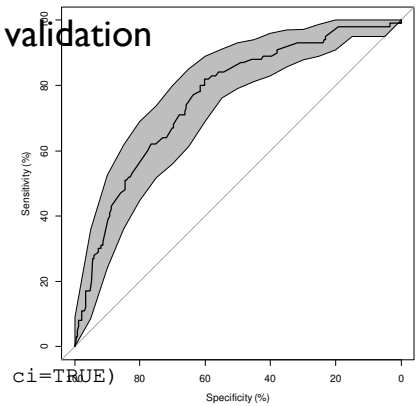
Courbe ROC du modèle logit obtenu

- Nombre de coefficients non significatifs au seuil de 5 %

```
> sum(summary(logit)$coefficients[,4] >= 0.05)
[1] 3
```

- Mesure de l'aire sous la courbe ROC sur l'échantillon de validation

```
> pred.logit <- predict(logit, newdata=valid, type="response")
> head(pred.logit)
         1          2          3          4          6          9
0.09621757 0.79481654 0.05667550 0.43594338 0.06629085 0.03835418
> library(pROC)
> auc(valid$Cible,pred.logit, quiet=TRUE)
Area under the curve: 0.7596
```



- Courbe ROC avec intervalle de confiance

```
> roc <- plot.roc(valid$Cible, pred.logit, main="", percent=TRUE, ci=TRUE)
> roc.se <- ci.se(roc, specificities=seq(0, 100, 5))
> plot(roc.se, type="shape", col="grey")
```

- Le package `pROC` permet aussi d'ajouter des intervalles de confiance calculés par simulations de Monte-Carlo
- La fonction `ci.se` calcule l'intervalle de confiance de la sensibilité pour les spécificités spécifiées, dans notre exemple les spécificités entre 0 et 1, avec un pas de 0,05. Ce calcul se fait à l'aide rééchantillonnages bootstrap de la courbe ROC, au nombre de 2000 sauf mention contraire

Calcul des weights of evidence (WoE) I

- Une fonction `woe` permet de calculer les WoE d'une variable explicative croisée avec la variable à expliquer

```
> woe <- function(X,Y){
+ tab <- table(X,Y)
+ woe <- log((tab[,1]/sum(tab[,1])) /(tab[,2]/sum(tab[,2])))
+ levels(X) <- woe
+ return(as.numeric(as.character(X)))
+ }
> Z <- woe(credit2$Comptes,credit2$Cible)
> table(Z,credit2$Cible)

Z                     0   1
  -0.818098705694941 139 135
  -0.401391782720529 164 105
  0.405465108108164   49  14
  1.17626322289818   348  46
```

- Application de la fonction `woe` à toutes les variables explicatives qui sont des facteurs

```
> colClasses <- sapply(credit2,class)
> varquali   <- intersect(which(colClasses %in% c("factor")),which(names(credit2) != "Cible"))
> varquanti  <- setdiff(names(credit2), names(credit2)[varquali])
> credit_woe <- sapply(varquali, function(i) woe(credit2[[i]], credit2$Cible))
> colnames(credit_woe) <- names(credit2)[varquali]
> credit_woe <- data.frame(credit_woe, Cible=credit2$Cible)
```