Module de Régression Régularisé. Régression logistique. Scoring

Mathilde Mougeot

ENSIIE

2017-2018

Régression Logistique

- Modèle **très utilisé** dans de nombreux domaines
- Domaine médical -premières applications-
- Banque, Assurance : Credit Scoring
- Module disponible dans tous les logiciels de stats

Plan

- Applications
- Modèle de régression logistique
- Interprétation du modèle
- Critères de performances
- Sélection de modèles

APPLICATIONS

Domaine Médical -Maladie cardiaque

chd : target variable ; 8 covariables

chd	Coronary heart disease binary response
sbp	systolic blood pressure (integer)
tobacco	cumulative tobacco (kg) (real)
ldl	low density lipoprotein cholesterol (real)
adiposity	(real)
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior (integer)
obesity	(real)
alcohol	current alcohol consumption (real)
age	age at onset (real)

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa.
data described in Rousseauw et al, 1983, South African Medical Journal.
Elements of Statistical Learning, Hastié, Tibshirani, Friedman.

`http://statweb.stanford.edu/~tibs/ElemStatLearn/`

Data -Maladie cardiaque

nř	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1
6	132	6.20	6.47	36.21	Present	62	30.77	14.14	45	0
7	142	4.05	3.38	16.20	Absent	59	20.81	2.62	38	0
8	114	4.08	4.59	14.60	Present	62	23.11	6.72	58	1
9	114	0.00	3.83	19.40	Present	49	24.86	2.49	29	0
10	132	0.00	5.80	30.96	Present	69	30.11	0.00	53	1
...	...	...	...	...	...					

Volumétrie des données :

$n = 462$, $p(chd = 1) = 34\%$.

Problématique

- Comprendre quels sont les facteurs liés au déclenchement de la maladie (chd), sur ce jeu de données
  - Significativité (oui/non : résultat d'un test)
  - Importance de l'effet
  - effet positif ou négatif

  $\rightarrow$ Prévention médicale

- Evaluer la qualité du modèle,
  la performance du modèle sur de nouvelles données.

- **Aide à la décision**
  Prédire, pour un nouveau patient, le risque de déclencher la maladie.
  $\rightarrow$ Meilleur suivi des patients à risque
- Recherche d'un Modèle parcimonieux
  Meilleure compréhension, suvi des facteurs
  Meilleur pouvoir prédictif

Domaine Bancaire

Incident	Présence d'un incident bancaire
revenu	(valeur numérique)
depnaiss	département de naissance (variable qualitative)
datenaiss	année de naissance
duree	durée du crédit en cours
montcred	montant du crédit en cours
situfam	situtation familiale
ancienn	nombre de mois d'ancienneté
cb	possession d'une carte bleue (1) ou non (0)
numero	numéro du client dans la base

Volumétrie des données

Données réelles $n = 50000$ clients, $p_{Incident=1} = 2\%$ $(2/1000)$

Problématique

- Trouver un Modèle parcimonieux (peu de variables)
  $\rightarrow$ minimiser le nombre de variables nécessaires au diagnostic
- Evaluer les risques d'incident pour un nouveau client
  $\rightarrow$ la banque décide d'accorder (ou non) le prêt
- Evaluer les performances
  $\rightarrow$ coûts d'exploitation du modèle

Application similaire : ciblage marketing, churn,...

Plan

- Applications
- Modèle de régression logistique
- Interprétation du modèle
- Critères de performances
- Sélection de modèles

MODELE DE REGRESSION LOGISTIQUE

BINAIRE

(ordinale)

Régression Logistique binaire

Les variables :

- Y variable cible binaire $\{0, 1\}$
- $X_1, \ldots X_d$ variables explicatives quantitatives ou binaires (indicatrices de modalités)
  - X_1 : Rég. logistique simple
  - $X_1, X_2, \ldots$: Rég. logistique multiple

Les données :

- Echantillon de taille (n, d) de données numériques (ex. table SAS, dataframe sous R)
- $\mathcal{D}_n = \{(x_i, y_i) \;\; 1 \leq i \leq n, x_i \in \mathbb{R}^d \; y_i \in \{0, 1\}\}$
- Notations : d contient la constante

Régression Logistique

Les variables :

- Y variable cible binaire $\{0, 1\}$ (variable dépendante)
- X variable explicative multivariée (variable indépendante)

L'objectif est de modéliser : $\mathbb{E}(Y/X = x)$

Pour une variable Y valant 0 ou 1, cette valeur est :

$$\begin{aligned}\mathbb{E}(Y/X = x) &= Prob(Y = 1/X = x) \\ &= \eta(x)\end{aligned}$$

Remarques :

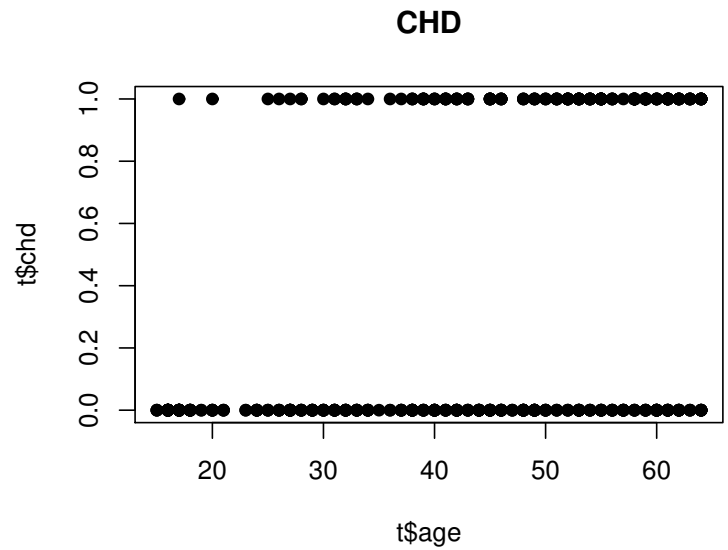
- $Prob(Y = 1/X = x)$: probabilité a posteriori
- cas où le modèle linéaire $\eta(x) = \beta_0 + \beta_1 X_1 + ...$ est peu adapté

Régression Logistique simple

Modèle simple (une variable) permettant d'expliquer :

- chd (Coronary Heart Disease, $\{0, 1\}$)
- en fonction de l'âge (valeur réelle)

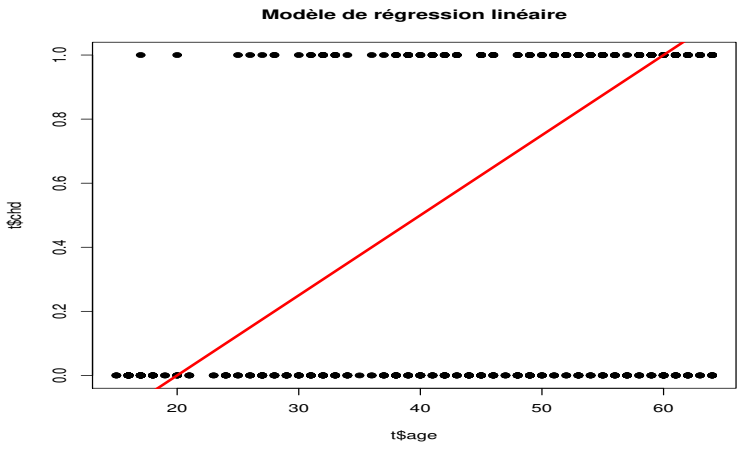
Illustration : données brutes

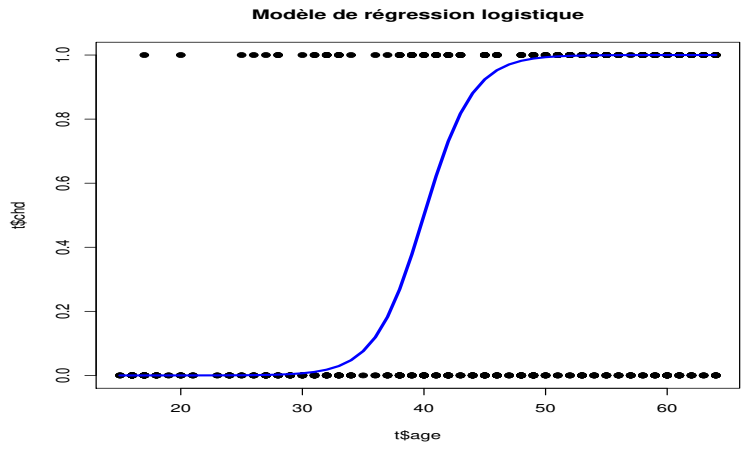


Régression Logistique simple

Modèle simple permettant d'expliquer :

- chd (Coronary Heart Disease, $\{0, 1\}$)
- en fonction de l'âge (valeur réelle)





$\rightarrow$ Modèle linéaire $\eta(x) = \beta_0 + \beta_1 X_1 + ...$ peu adapté (fig. gauche)

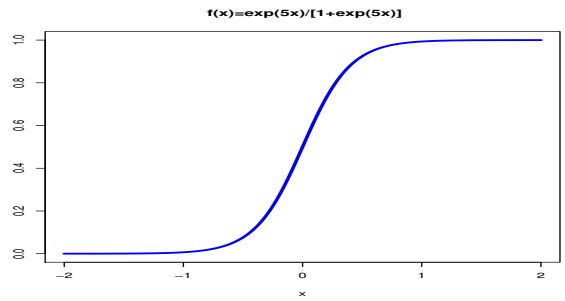
$\rightarrow$ Modèle de régression logistique (fig. droite)

Le modèle logistique

Fonction de Transfert : $\eta(x)$

posons $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$

$$\eta(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$



$$\eta(z) \ : \mathbb{R} \to [0,1]$$

- fonction de lien **Logit** : $\log(\frac{\eta(x)}{1-\eta(x)}) = \beta^T x$

GLM :Generalized Linear Models

Modèle Statistique sous-jacent

Régression logistique :

- $Y \in \{0, 1\}$, X variable multivariée.

- Modèle Statistique considéré :
  - $\mathcal{L}(X, Y) \equiv (\mathcal{P}_X, \eta)$
    avec $\eta(x) = \mathbb{E}(Y/X = x)$
    avec ici η connue et $\mathcal{P}_X$ non spécifiée.

Analyse Discriminante (pour mémo) :

- $\mathcal{L}(X, Y) \equiv (p, \mathcal{L}(X/Y))$ avec $\mathcal{L}$ loi normale.

Les Fonctions de liens

Différentes fonctions de transfert ont été proposées :

$$\mathbb{E}(Y/X = x) = Prob(Y = 1/X = x) = \eta(x)$$

C'est un choix de modélisation, un choix "Métier". Fonctions de liens :

- Modèle Logit : $\eta(z) = e^z/(1 + e^z) \leftrightarrow g(\eta) = \log(\frac{\eta}{1-\eta})$
- Modèle probit (normit) : $\eta(z) = \Phi(z) \leftrightarrow g(\eta) = \Phi^{-1}(\eta)$
  où Φ est la fonction de répartition de la loi Normale centrée, réduite
- "Modè Log-Log" : $g(\eta) = \log(-\log(1 - \eta))$
  (épidémiologie, toxicologie)

$\rightarrow$ **qui entraîne un changement de modèle statistique**

$$\mathcal{L}(X, Y) \equiv (\mathcal{P}_X, \eta) \text{ et } \eta = \mathbb{E}(Y/X = x) \text{ modifiée}$$

Estimation des paramètres du modèle logistique

- **Variables aléatoires :**
  - Y variable binaire à valeur dans $\{0, 1\}$
  - X variable à valeur réelle $X \in \mathbb{R}^d$ ($d = 1$ régression logistique simple)
- **Les données observées i.i.d.**
  - n échantillon de données
  - $\mathcal{D}_n = \{(x_i, y_i),\ \ 1 \leq i \leq n,\ \ y_i \in \{0, 1\}\}$
- **Le modèle pour une observation x_i :**
  - $\eta(x_i) = Prob(Y = 1/X = x_i)$
    $= \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$
- Les paramètres β sont estimés par maximum de vraisemblance (minimisation de la - log-vraisemblance)

Vraisemblance conditionnelle des données

Echantillon de données iid : $\mathcal{D}_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Probabilité d'observer $\mathcal{D}_n$:

$$\mathcal{L}(\beta, (x_1, y_1), \ldots, (x_n, y_n))$$

$$\begin{aligned}
\mathcal{L}_{\beta,\mathcal{D}_n} &= \prod_{i=1}^{n} Prob(Y = y_i / X = x_i) \\
&= \prod_{i=1}^{n} \eta(x_i)^{y_i} (1 - \eta(x_i))^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\beta^T x_i}}\right)^{1-y_i} \\
&= \mathcal{L}_{\mathcal{D}_n}(\beta)
\end{aligned}$$

Avec $y_i = 1$ ou $y_i = 0$
$\eta(x_i) = Prob(Y = 1/X = x_i)$ et $1 - \eta(x_i) = Prob(Y = 0/X = x_i)$

Log-Vraisemblance conditionnelle

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}_n}(\beta) &= \prod_{i=1}^{n} Prob(Y = y_i / X = x_i) \\
\ell_{\mathcal{D}_n}(\beta) &= \log(\mathcal{L}_{\mathcal{D}}(\beta)) \\
&= \log(\prod_{i=1}^{n} \eta(x_i)^{y_i}(1-\eta(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{n} y_i \log(\tfrac{\eta(x_i)}{1-\eta(x_i)}) + \log(1-\eta(x_i)) \\
&= \sum_{i=1}^{n} \{y_i \beta^T x_i - \log(1+e^{\beta^T x_i}\}
\end{aligned}$$

Rappel : on inclut dans "x_i" un terme constant égale à 1.

- On cherche $\hat{\beta}$ maximisant la Log-Vraisemblance $\ell_{\mathcal{D}}(\beta)$
- La Log-Vraisemblance est une fonction concave, il existe donc une solution unique $\hat{\beta}$

Maximiser la Log-Vraisemblance

Pour maximiser la Log-Vraisemblance :

- il faut annuler les d dérivées en $\beta = (\beta_1, \ldots, \beta_d)^T$
  $\forall j\ \beta_j,\ \frac{\delta \ell}{\delta \beta_j} = 0$
- Les (d) équations de score valent :
  $\frac{\delta \ell(\beta)}{\delta \beta_j} = \sum_{i=1}^{n} x_{i,j}(y_i - \eta(x_i, \beta)) = 0$ avec $\eta(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$
- En matriciel : $\frac{\delta \ell(\beta)}{\delta \beta} = \sum_{i=1}^{n} x_i(y_i - \eta(x_i, \beta)) = 0$

En particulier : pour $x_i = 1$ (terme constant), on a :
$\sum_{i=1}^{n} y_i/n = \mathbb{E}\ \eta(x_i = 1, \beta_1)$,
le nombre moyens d'observations dans la classe 1 est égale au nombre attendu, à son espérance.

Maximiser la Log-Vraisemblance conditionnelle

On cherche la valeur $\beta = (\beta_1, .., \beta_d)$ qui annulent les équations de score :

$\frac{\delta \ell(\beta)}{\delta \beta} = \sum_{i=1}^{n} x_{i,j}(y_i - \eta(x_i, \beta)) = 0$

avec $\eta(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

Pas d'expression analytique Directe
mais solution unique...

Maximiser la Log-Vraisemblance

On souhaite annuler les dérivées : $\boxed{\frac{\delta \ell}{\delta \beta} = \sum_{i=1}^{n} x_i(y_i - \eta(x_i, \beta)) = 0}$

Développement de Taylor d'une fonction $f(x)$ au premier ordre :

$$f(x) \sim f(x_0) + f^{'}(x_0)(x - x_0)$$

On souhaite résoudre l'équation affine $0 = f(x_0) + f^{'}(x_0)(x - x_0)$, on obtient alors un point x_1, et on construit par récurrence :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

L'algorithme de Newton-Raphson nécessite ici le calcul des dérivées seconde, de la matrice hessienne :
La mise à jour des valeurs successives de β est donnée par :

$$\beta^{new} = \beta^{old} - (\frac{\delta^2 \ell(\beta^{old})}{\delta \beta \delta \beta^T})^{-1} \frac{\delta \ell(\beta^{old})}{\delta \beta}$$

avec $\frac{\delta^2 \ell(\beta)}{\delta \beta \delta \beta^T} = -\sum_{i=1}^{n} x_i x_i^T \eta(x_i, \beta)(1 - \eta(x_i, \beta))$

Maximiser la Log-Vraisemblance

Notation matricielle :

- X : matrice $(n \times p)$
- Y : vecteur $(n \times 1)$
- η : vecteur $(n \times 1)$, élément i, $\eta(x_i, \beta^{old})$

$$\boxed{\frac{\delta \ell(\beta)}{\delta \beta} = X^T(Y - \eta)}$$

- W : matrice diagonale $(n \times n)$,
- $W(i,i) = \eta(x_i, \beta^{old})(1 - \eta(x_i, \beta^{old})$

$$\boxed{H = \frac{\delta^2 \ell(\beta)}{\delta \beta \delta \beta^T} = -X^T W X}$$

H : matrice Hessienne

Maximiser la Log-Vraisemblance

Méthode d'estimation : Newton-Raphson et IRLS

$$\begin{aligned}\beta^{new} &= \beta^{old} + (X^T W X)^{-1} X^T (Y - \eta) \\ &= (X^T W X)^{-1} X^T W (X\beta^{old} + W^{-1}(Y - \eta)) \\ &= (X^T W X)^{-1} X^T W z\end{aligned}$$

Avec $z = X\beta^{old} + W^{-1}(Y - \eta)$

IRLS : Iteratively Reweighted Least Square

$\beta^{new} \leftarrow ArgMin(z - X\beta)^T W (z - X\beta)$

$\beta = 0$ bonne valeur de départ. Convergence non garantie.

Etapes : Estimation/Prédiction/Décision

Au départ, n-Echantillon : $\mathcal{D}_n$. choix d'un modèle.

1. Estimation des paramètres β du modèle
   - Calcul de $\hat{\beta}$
   - Sélection des paramètres (méthode forward, backward, stepwise)

2. Utilisation du modèle : Prédiction
   - Nouvelle observation x_{new}, $x_{new} \notin \mathcal{D}_n$.
   - Estimation (calcul) de Probabilité : $\boxed{\hat{\eta}(x, \hat{\beta}) = \frac{e^{\hat{\beta}^T x_{new}}}{1+e^{\hat{\beta}^T x_{new}}}}$

3. Décision

   En fonction de la valeur d'un seuil S, $S \in [0, 1]$
   - $\hat{Y} = 1$ si $\hat{\eta}(x, \hat{\beta}) > S$
   - $\hat{Y} = 0$ sinon

   Par exemple, $S = 0.5$, mais d'autres valeurs peuvent être plus appropriées (voir plus loin).

Plan

- Applications
- Modèle de régression logistique
- Interprétation du modèle
- Critères de performances
- Sélection de modèles

Estimation et Intervalles de confiance

$\hat{\beta}$ est un estimateur du maximum de vraisemblance.
Il en possède toutes les propriétés.

- il est asymptotiquement sans biais
- il est de variance minimale
- il est asymptotiquement gaussien

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow_{n\rightarrow\infty} \mathcal{N}(0, \Sigma_0)$$

$\Sigma_0 \simeq -H_n^{-1}$ avec $H_n = -X^T WX/n$ ($H_n = H$ matrice Hessienne)

La connaissance de la loi asymptotique permet de calculer :

- Intervalles de confiance sur β
- Test de significativité ($\neq 0$) sur β

Significativité des coefficients

- Le modèle

$$\eta_\beta = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_d X_d}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_d X_d}}$$

- Test de Wald (au risque α) :
  - $H_0 : \beta_j = 0$
  - $H_1 : \beta_j \neq 0$
- Statistique de test (Wald) : $Z_j = \frac{\hat{\beta}_j}{S_{\beta_j}}$

  suit asymptotiquement une loi Normale (R)
- Décision
  - en fonction de la valeur de la p-value et du risque α

Remarque : attention à la collinéarité entre les variables (impact sur le calcul de S_{β_j})

Pertinence du Modèle global

Modèle logit :

$$\eta_\beta = \frac{e^{\beta_1+\beta_2 X_2 \ldots+\beta_d X_d}}{1+e^{\beta_1+\beta_2 X_2+\ldots+\beta_d X_d}}$$

- Classifieur sans variables explicatives. Modèle M_0, $\beta_2 = \ldots = \beta_d = 0$.
  - Estimation : $\hat{\beta}_1 = \ln \frac{\bar{y}}{1-\bar{y}} = \ln \frac{n_+}{n_-}$
  - Log-vraisemblance : $\ell_0(\hat{\beta}_1, \mathcal{D}_n) = \sum y_i \ln(\bar{y}) + (1-y_i)\ln(1-\bar{y}) = n_+ \ln(\bar{y}) + n_- \ln(1-\bar{y})$
  - **Deviance :** $D_0 = -2 \times \ell_0$
- Déviance du modèle complet étudié : $D_M = -2\ell(\hat{\beta}, \mathcal{D}_n)$
- sous l'hypothèse H_0 (tous les coefficients sont nuls)
  $(D_0 - D_M) \sim \chi^2(d-1)$

Modèle M_0

Estimation du paramètre β :

$$\mathcal{L}_{\mathcal{D}_n}(\beta) = \prod_{i=1}^{n} Prob(Y = y_i / X = x_i)$$

$$\ell_{\mathcal{D}_n}(\beta) = \log(\mathcal{L}_{\mathcal{D}}(\beta)) = \log(\prod_{i=1}^{n} \eta(x_i)^{y_i}(1-\eta(x_i))^{1-y_i})$$

$$= \sum_{i=1}^{n} y_i \log(\frac{\eta(x_i)}{1-\eta(x_i)}) + \log(1-\eta(x_i))$$

$$= \sum_{i=1}^{n} \{y_i \beta^T x_i - \log(1+e^{\beta^T x_i}\}$$

$$= \sum_{i=1}^{n} \{y_i \beta - \log(1+e^{\beta}\}$$

On cherche $\hat{\beta}$ qui annule la dérivée de $\ell_{\mathcal{D}_n}(\beta)$:

- $\hat{\beta} = \ln \frac{\bar{y}}{1-\bar{y}} = \ln \frac{n_+}{n_-}$
- Log-vraisemblance : $\ell_0(\hat{\beta}_1, \mathcal{D}_n) = \sum y_i \ln(\bar{y}) + (1-y_i)\ln(1-\bar{y})$
  $= n_+ \ln(\bar{y}) + n_- \ln(1-\bar{y})$

Domaine Médical -Maladie cardiaque

chd : target variable ; 8 covariables

chd	Coronary heart disease response
sbp	systolic blood pressure (integer)
tobacco	cumulative tobacco (kg) (real)
ldl	low density lipoprotein cholesterol (real)
adiposity	(real)
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior (integer)
obesity	(real)
alcohol	current alcohol consumption (real)
age	age at onset (real)

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa.
data described in Rousseauw et al, 1983, South African Medical Journal.
Elements of Statistical Learning, Hastié, Tibshirani, Friedman.

`http://statweb.stanford.edu/~tibs/ElemStatLearn/`

Regression logistique, application logicielle

- sous R

```
res=glm(chd~. ,family=binomial,data=tab);
summary(res)
```

- sous SAS

```
proc logistic data=tab;
class chd(desc);
model chd=age; run;
```

Interprétation du modèle

Sorties du logiciel R :

```
res=glm(chd...,family=binomial,data=tab);summary(res)
```

$n = 468$, $p = 7$, $Y \leftarrow chd$

Coefficients :	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-4.129	0.964	-4.283	1.84e-05	***
sbp	0.005	0.005	1.023	0.30643	
tobacco	0.079	0.026	3.034	0.00242	**
ldl	0.184	0.057	3.219	0.00129	**
famhistPresent	0.939	0.224	4.177	2.96e-05	***
obesity	-0.034	0.029	-1.187	0.23529	
alcohol	0.000	0.004	0.136	0.89171	
age	0.042	0.010	4.181	2.90e-05	***

Odd-ratio (rapport de côte)

Interprétation et influence des coefficients :

Odd-ratio

Il mesure l'évolution du rapport des probabilités d'apparition de l'événement $Y = 1$ contre $Y = 0$, lorsque X_j passe de x_j à $x_j + 1$.

- Pour une variable X réelle : $OR = \frac{\eta(x_j+1)/(1-\eta(x_j+1))}{\eta(x_j)/(1-\eta(_j x))} = e^{\beta_j}$
- Pour une variable X à valeur binaire $\{0, 1\}$:

  $OR = \frac{P(Y=1/X_j=1)/(1-P(Y=1/X_j=1)}{P(Y=1/X_j=0)/(1-P(Y=1/X_j=0)} = e^{\beta_j}$

$\rightarrow$ Un OR < 1 indique une influence négative de X_j sur Y.
$\rightarrow$ Un OR > 1 indique une influence positive de X_j sur Y.

Intervalle de confiance de OR au niveau 95% : $[e^{\hat{\beta}-1.96S_{X_j}}, e^{\hat{\beta}+1.96S_{X_j}}]$

Odd-ratio et intervalle de confiance

Résultats SAS :

Procédure LOGISTIC (variable CHD):
Estimations par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	3.5212	0.4160	71.6469	<.0001
age	1	-0.0641	0.00853	56.4428	<.0001

Estimations des rapports de cotes:

Effet	Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
age	0.938	0.922	0.954

Odd-ratio

Influence du codage sur les valeurs des coefficients

Résultats R :

```
Coefficients (CODAGE Y=1 CHD=OUI/1):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.521710   0.416031   8.465  < 2e-16 ***
age         -0.064108   0.008532  -7.513 5.76e-14 ***
exp(res2$coeff)
(Intercept)         age
 33.8422607   0.9379037


Coefficients (CODAGE Y=1 CHD=NON/0):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.521710   0.416031  -8.465  < 2e-16 ***
age          0.064108   0.008532   7.513 5.76e-14 ***

exp(res$coeff)
(Intercept)         age
 0.02954885  1.06620758
```

Sortie R Régression logistique sur données CHD

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp              0.0057607  0.0056326   1.023  0.30643
tobacco          0.0795256  0.0262150   3.034  0.00242 **
ldl              0.1847793  0.0574115   3.219  0.00129 **
famhistPresent   0.9391855  0.2248691   4.177 2.96e-05 ***
obesity         -0.0345434  0.0291053  -1.187  0.23529
alcohol          0.0006065  0.0044550   0.136  0.89171
age              0.0425412  0.0101749   4.181 2.90e-05 ***
Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 483.17  on 454  degrees of freedom
AIC: 499.17
```

Plan

- Applications
- Modèle de régression logistique
- Interprétation du modèle
- Critères de performances
- Sélection de modèles

Crtitères de performances

matrice de confusion & co.

Matrice de confusion

- Une observation i est affectée à la classe $Y = 1$

  si $\hat{\eta}(x_i) > S$, (S : seuil, par exemple =0.5)

- Performances sur un ensemble de données étiquetées

$g(x) = \hat{y}$	$y = 0$	$y = 1$
$g(x) = 0$	diag. correcte	Faux Négatif
$g(x) = 1$	**Faux Positif**	diag. correcte
	n_0	n_1

- Notions de Performance, Erreur globale.
- Sensibilité : Capacité à diagnostiquer les $\hat{Y} = 1$ parmi les $Y = 1$
- Spécificité : Capacité à diagnostiquer les $\hat{Y} = 0$ parmi les $Y = 0$
- Faux Positifs : diagnostic $\hat{Y} = 1$ à tort.
- Faux Négatifs : diagnostic $\hat{Y} = 0$ à tort

Trouver un compromis acceptable
entre forte sensibilité et forte spécificité

Standard Error for binary classification

Reality / Decision	$y = 0$	$y = 1$
$\hat{y} = 0$	TN	FN
$\hat{y} = 1$	FP	TP

- Accuracy $= \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}}$
- Recall $= \frac{\text{TP}}{\#(\text{real P})} = \frac{\text{TP}}{\text{FN+TP}}$
- Precision $= \frac{\text{TP}}{\#(\text{predicted P})} = \frac{\text{TP}}{\text{FP+TP}}$
- F-score$= 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$

Rem. : Recall= sensitivity.
False-Discovery Rate (FDR)= 1-Precision.

Matrice de confusion

Performances du modèle sur base de test (2 classes)
Problématique de risque de crédit (1 : défaillance crédit).
Base de données : $n = 200$, $n_0 = 120$ $\{0\}$, $n_1 = 80$ $\{1\}$ (pb de crédit)

$g(x) = \hat{y}$	$\{0\}$	$\{1\}$	TOTAL
prédiction $\{0\}$	110	**10**	120
prédiction $\{1\}$	**10**	70	80
TOTAL	120	80	200

- Performance : $\frac{110+70}{200} = \frac{180}{200}$. Taux d' Erreur $= \frac{10+10}{200} = \frac{20}{200} = 10\%$
- Sensibilité $= 70/80$ (capacité à diag. les incidents / les incidents)
- Spécificité $= 110/120$ (capacité à reconnaître les "0" parmi les "0")
- Taux de Faux Positifs $= \frac{10}{120} = 8,33\%$ (risque diag. incident / "0")
- Taux de Faux Négatifs $= \frac{10}{80} = 12,5\%$

Plan

- Applications
- Modèle de régression logistique
- Interprétation du modèle
- Critères de performances
- Sélection de modèles

CRITERES DE PERFORMANCE

courbe ROC

Performance d'un classifieur : courbe ROC

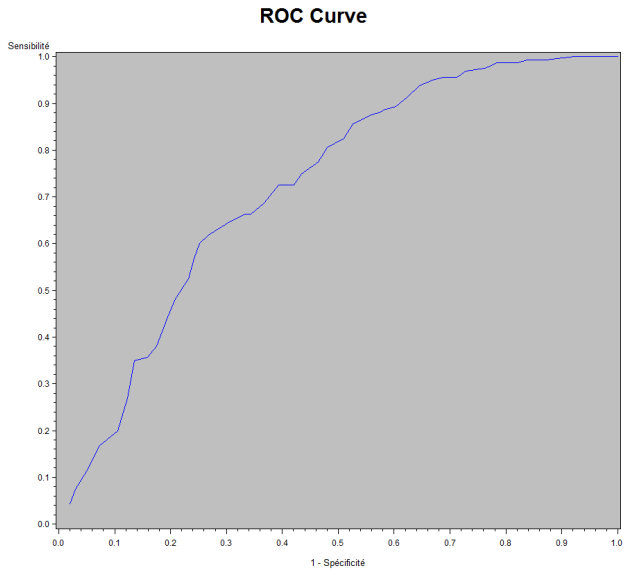
ROC : acronyme de Receiver Operating System

- Adaptée à des modèles de classification à 2 classes
- Usage : comparaison de performance entre plusieurs modèles de classification ou de scoring

- Sensibilité
  - probabilité de bien détecter un événement au seuil s
  - si $score(x) > s$ alors $\hat{Y} = 1$, "événement détecté"
  - $\alpha(s) = P(score(x) > s / x = evenement)$
  - $\alpha(s) = P(\hat{Y} = 1 / Y = 1)$

- Specificité
  - probabilité de bien détecter un non-événement au seuil s
  - $\beta(s) = P(score(x) < s / x = non - evenement)$
  - $\beta(s) = P(\hat{Y} = 0 / Y = 0)$
  - la proportion de faux-événement (faux positif) est $1 - \beta(s) = P(score(s) > s / x = non - evenement)$

Courbe ROC, Performances du modèle

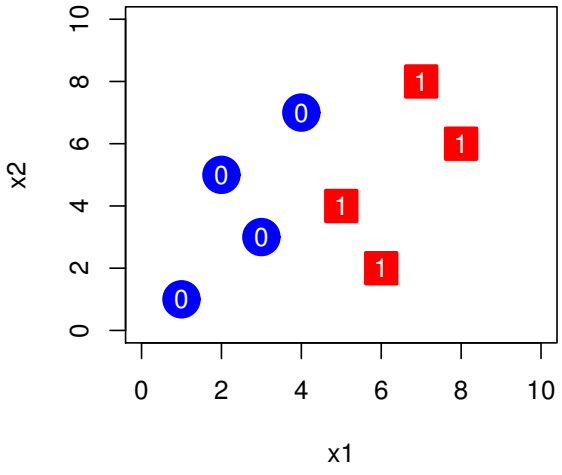


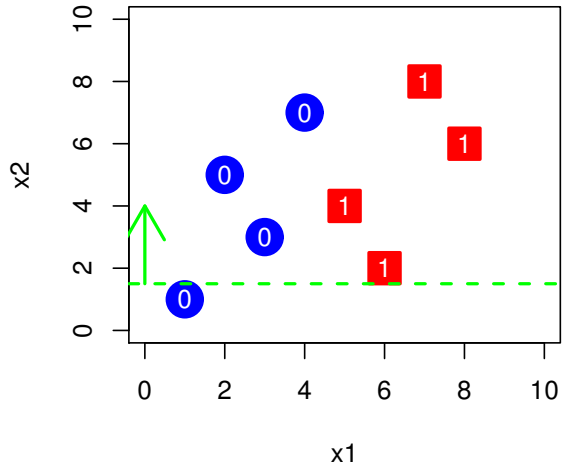
- Sensibilité : capacité à prédire un événement
- Spécificité : capacité à prédire un non-événement
- **Graphique Roc : y : Sensibilité(c) ; x : 1-Spécificité(c)**

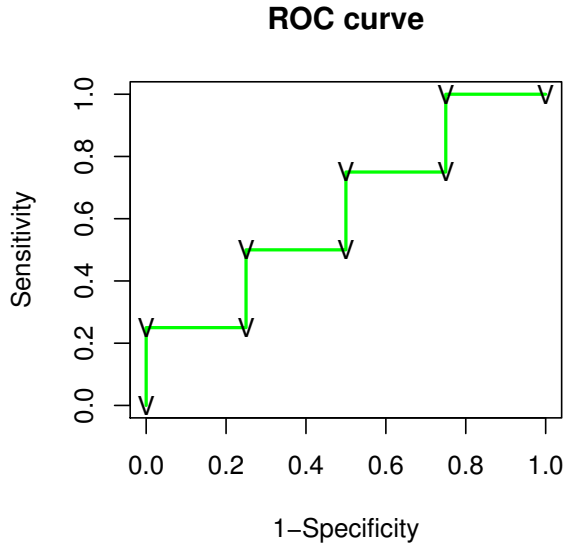
L'aire sous la courbe ROC (AUC) est une mesure du "pouvoir prédictif" du modèle (régression logistique simple de la variable X).

Courbe ROC, illustration

On suppose ici $score(x) = g(x) = x_2$



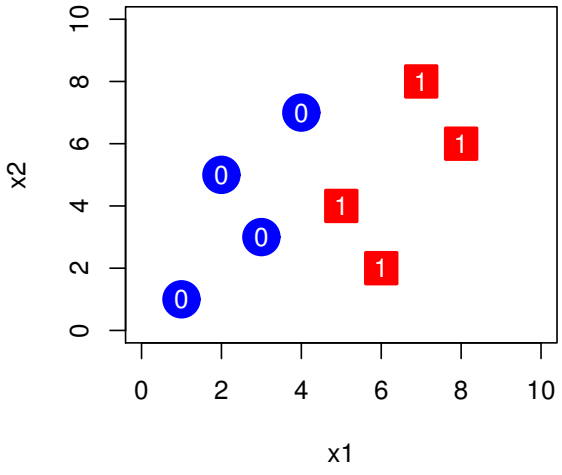


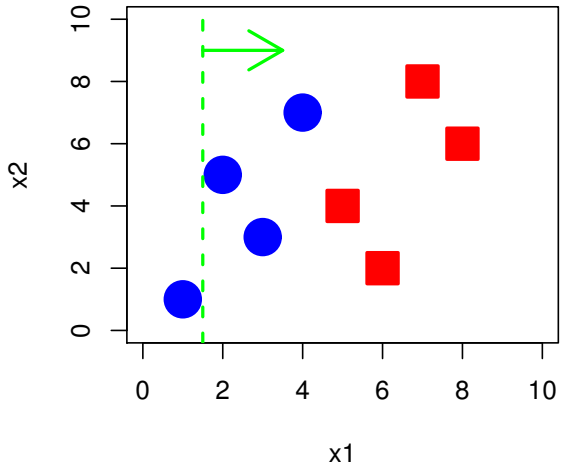


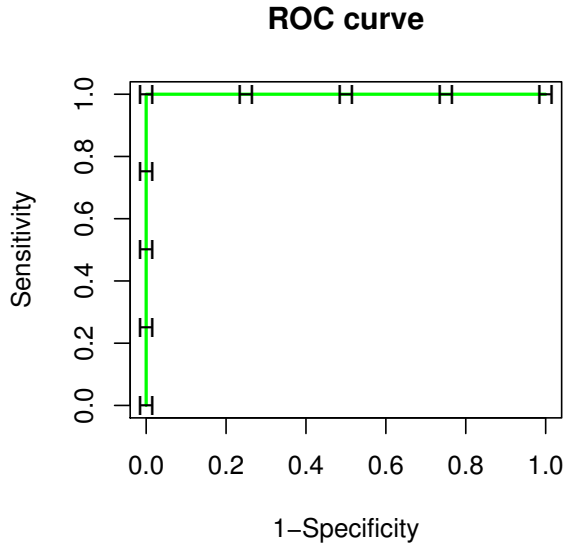
seuilH	α	β	$1-\beta$
0.5	1.00	0.00	1.00
1.5	1.00	0.25	0.75
2.5	0.75	0.25	0.75
3.5	0.75	0.50	0.50
4.5	0.50	0.50	0.50
5.5	0.50	0.75	0.25
6.5	0.25	0.75	0.25
7.5	0.25	1.00	0.00
8.5	0.00	1.00	0.00

Courbe ROC, illustration

On suppose ici $score(x) = g(x) = x_1$

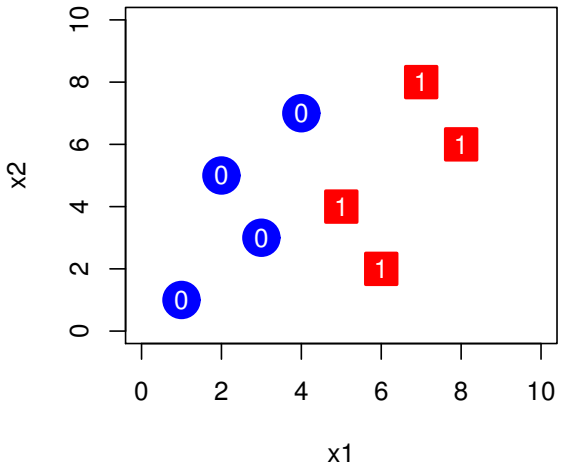


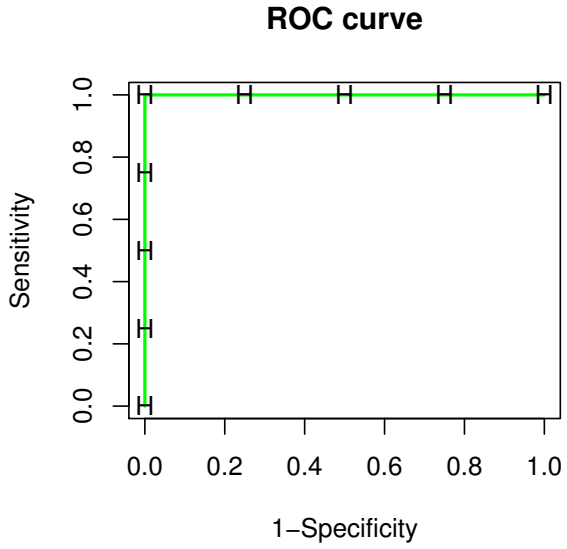




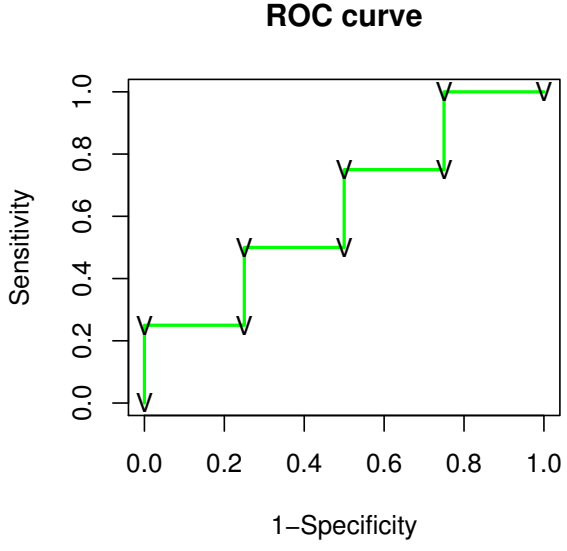
seuilV	α	β	$1-\beta$
0.5	1.00	0.00	1.00
1.5	1.00	0.25	0.75
2.5	1.00	0.50	0.50
3.5	1.00	0.75	0.25
4.5	1.00	1.00	0.00
5.5	0.75	1.00	0.00
6.5	0.50	1.00	0.00
7.5	0.25	1.00	0.00
8.5	0.00	1.00	0.00

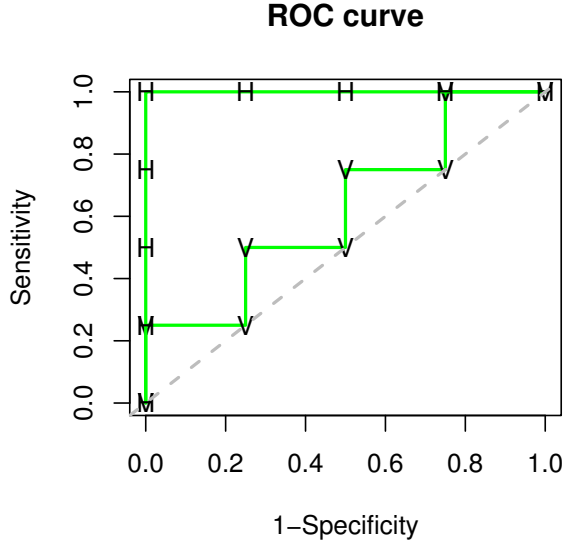
Courbe ROC, illustration





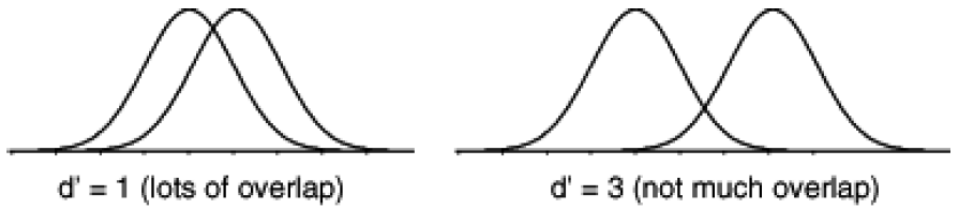
$score(x) = x_1$

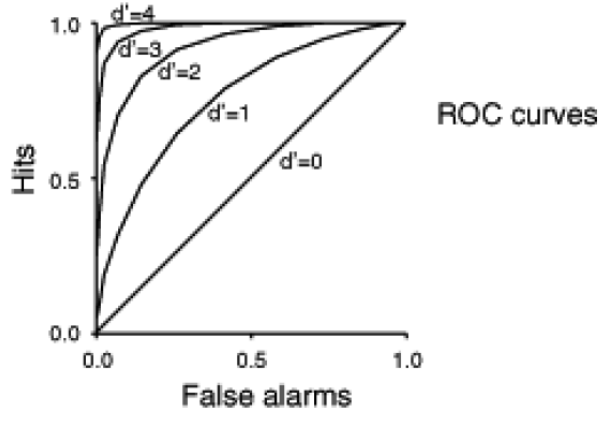




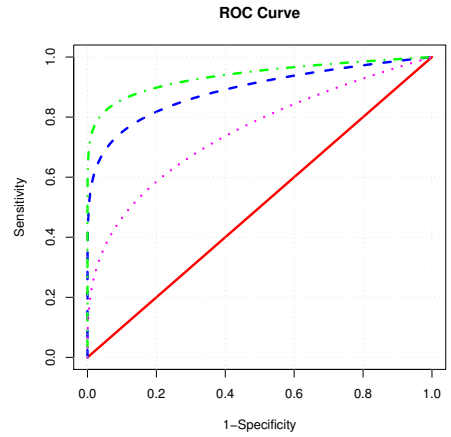
$score(x) = x_1)$

The Gold Standard for Scoring : the ROC curve (k=2)





Courbe ROC



- courbe ROC diagonale : le modèle n'est pas plus performant qu'un modèle aléatoire
- Plus la courbe est proche du coin supérieur gauche du carré, meilleur est le modèle. Il permet de capture le plus possible de vrais événements avec le moins possible de faux événements
- la courbe ROC permet une comparaison des modèles globalement (AUC) et localement.
- Cette courbe ne dépend pas de la probabilité de $Y = 0$ et $Y = 1$.

Plan

- Applications
- Modèle de régression logistique
- Interprétation du modèle
- Sélection de modèles

SELECTION DE MODELE

Significativité des Coefficients β

- Le modèle :
  $\eta(x) = Prob(Y = 1/X = x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$
- **Test de Wald :** (test bilatéral)
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$
- Statistique de test (Wald, autre présentation) : $= \frac{\hat{\beta}_1^2}{Var(\hat{\beta}_1)}$
  suit asymptotiquement une loi du $\chi^2(1)$
- Décision :
  - Rejet de H_0 si $w_{obs} \geq \chi_1^2(1-\alpha)$
  - $Prob(W > w_{obs}) \leq \alpha$
  - Décision, en fonction de la valeur de la p-value

Application

Facteur lié (ou pas) à l'apparition d'une maladie :

Paramètre	DDL	Valeur estimée	écart-type	Khi-2 de Wald	$Pr > Khi - 2$
Intercept	1	-5.6880	1.3068	18.9444	$<$.0001
sbp	1	0.00650	0.00573	1.2882	0.2564
tobacco	1	0.0794	0.0266	8.9028	0.0028
ldl	1	0.1739	0.0597	8.4982	0.0036
adiposity	1	0.0186	0.0293	0.4027	0.5257
famhist Present	1	0.4627	0.1139	16.4879	$<$ .0001
typea	1	0.0396	0.0123	10.3286	0.0013
obesity	1	-0.0629	0.0442	2.0214	0.1551
alcohol	1	0.000122	0.00448	0.0007	0.9784
age	1	0.0452	0.0121	13.9014	0.0002

Régression logistique pas à pas descendante

- On part du modèle complet
- A chaque étape, on enlève la variable
  ayant le Wald le moins significatif
  à condition que son niveau de signification soit supérieur à 10%

• *CodeSAS*

```
proc logistic data=heart;
class chd(desc) famhist(desc);
model chd=sbp ..... alcohol age/selection=backward;
output out=rout p=rpred;
run;
```

Régression logistique pas à pas ascendante

- On part du modèle vide
- A chaque étape, on introduit la variable X_j
  qui aura le niveau de signification du $\chi^2_{score}(X_j)$ le plus faible une fois introduite dans le modèle,
  à condition que l'apport de X soit significatif.

• *CodeSAS*

```
proc logistic data=heart;
class chd(desc) famhist(desc);
model chd=sbp ..... alcohol age/selection=forward;
output out=rout p=rpred;
run;
```

Régression logistique pas à pas stepwise

- On part du modèle vide
- Régression forward avec élimination de variables

• *CodeSAS*

```
proc logistic data=heart;
class chd(desc) famhist(desc);
model chd=sbp ..... alcohol age/selection=stepwise;
output out=rout p=rpred;
run;
```

Modèle de régression logistique et pénalisation

Comme pour la régression linéaire, on pénalise la Log-vraisemblance

Régression Logistique

Références :

- Agresti, A. (1990) Categorical Data Analysis, New-York : John Wiley & Sons, In
- Collet D. (1999) Modeling binary data, Chapman & Hall/CRC, Londres.
- Hastie, R & Tibshirani, R. & Friedman J. (2009) The Elements of Statistical Learning : datamining, inference and prediction. Springer.
- et bien d'autres...