

Régression Logistique

Une approche pour rendre calculable $P(Y/X)$

Ricco RAKOTOMALALA

PLAN

1. Fondements probabilistes, MMV et Estimateurs
2. Évaluation « empirique »
3. Évaluation « statistique »
4. Interprétation des coefficients
5. Sélection automatique de variables
6. Quelques commentaires et curiosités

Les fichiers XLS associés à ce support sont disponibles en ligne

http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistic_support_pour_slides.xls

http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistic_analyse_outlier_et_influential.xls

http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistic_covariate_pattern.xls

Fondements probabilistes

Principe de la maximisation de la vraisemblance
Estimation des paramètres

Théorème de Bayes

Probabilités conditionnelles – On se place dans le cadre binaire $Y \in \{+, -\}$

Estimer la probabilité conditionnelle $P(Y/X)$

$$P(Y = y_k/X) = \frac{P(Y = y_k) \times P(X/Y = y_k)}{P(X)}$$
$$= \frac{P(Y = y_k) \times P(X/Y = y_k)}{\sum_{l=1}^K P(Y = y_l) \times P(X/Y = y_l)}$$

Dans le cas à 2 classes

$$\frac{P(Y = + / X)}{P(Y = - / X)} = \frac{P(Y = +)}{P(Y = -)} \times \frac{P(X / Y = +)}{P(X / Y = -)}$$

La règle d'affectation devient
Si (ce rapport > 1) Alors $Y = +$

Cette quantité est facile à estimer à partir des données


Quelle hypothèse introduire pour rendre l'estimation de ce rapport possible ?

On parle de méthode **semi-paramétrique** parce qu'on ne fait pas d'hypothèses directement sur la distribution mais sur un rapport de distribution \rightarrow l'hypothèse est moins restrictive.

Hypothèse fondamentale de la régression logistique

$$\ln \left[\frac{P(X / Y = +)}{P(X / Y = -)} \right] = b_0 + b_1 X_1 + \dots + b_J X_J$$

Cette hypothèse couvre une très large classe de distributions

- Loi normale (idem Analyse discriminante)
- Loi exponentielle
- Lois discrètes
- Loi gamma, Beta, Poisson
- Mélange de variables explicatives binaires (0/1) et numériques 

Moralité

1. Champ d'application théoriquement plus large que l'Analyse Discriminante
2. Sa capacité à traiter et proposer une interprétation des coefficients pour les variables explicatives binaires est très intéressante

Le modèle LOGIT

Une autre écriture du rapport de probabilité

Écrivons $\pi(X) = P(Y=+/X)$

On définit le **LOGIT** de $P(Y=+/X)$ de la manière suivante

$$\ln \left[\frac{\pi(X)}{1 - \pi(X)} \right] = a_0 + a_1 X_1 + \dots + a_J X_J$$

$1 - \pi(X) = P(Y= - / X)$

Puisqu'on est dans un cadre binaire

$$\pi(X) = \frac{e^{a_0 + a_1 X_1 + \dots + a_J X_J}}{1 + e^{a_0 + a_1 X_1 + \dots + a_J X_J}}$$

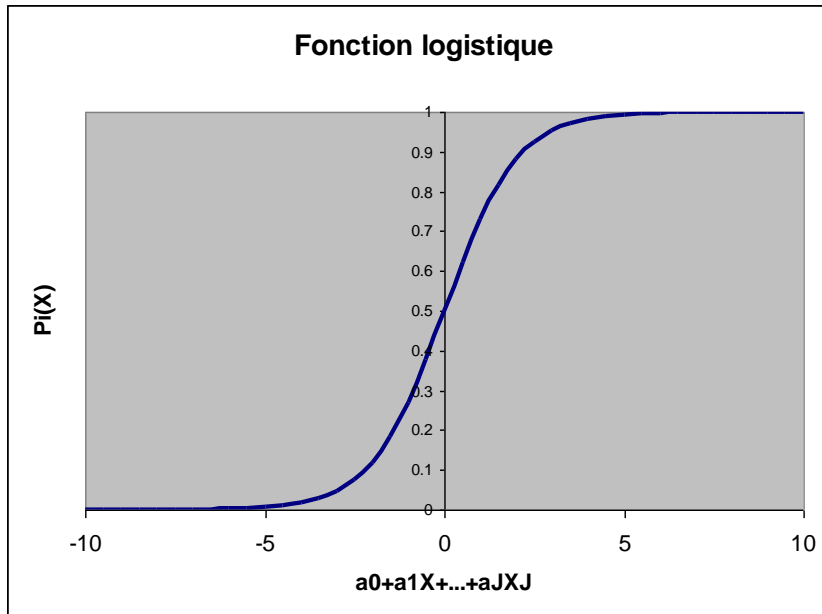
C'est la fonction de répartition de la loi Logistique

$$\frac{\pi(X)}{1 - \pi(X)} = \frac{P(+ / X)}{P(- / X)}$$

Représente un « **odds** » c.à.d. un rapport de chances. Ex. odds = 2 \rightarrow l'individu à 2 fois plus de chances d'être positif que d'être négatif.

La fonction logistique

Quelques éléments de lecture



Fonction logistique

A propos de la fonction de transformation

- $C(X) = a_0 + a_1 \cdot X_1 + \dots + a_j \cdot X_j$ varie de $-\infty$ à $+\infty$
- $0 \leq \pi(X) \leq 1$, c'est une probabilité !!!

A propos de la règle d'affectation

- $\pi(X) / [1 - \pi(X)] > 1 \rightarrow Y=+$
- $\pi(X) > 0.5 \rightarrow Y=+$
- $C(X) > 0 \rightarrow Y=+$

Remarques :

- $C(X)$ et $\pi(X)$ permettent de classer les individus selon leur propension à être +
- Sauf que $\pi(X)$ est une « vraie » probabilité
- D'autres fonctions cumulatives pour transformer $C(X)$. Ex. la loi normale : modèle PROBIT
- Fonction de transformation non-linéaire : on parle de régression non-linéaire dans la littérature

Équivalence entre les approches

$$\begin{aligned}\ln\left[\frac{\pi(X)}{1-\pi(X)}\right] &= a_0 + a_1 X_1 + \dots + a_J X_J \\ &= \ln\left[\frac{P(+)\times P(X/+)}{P(-)\times P(X/-)}\right] \\ &= \ln\left[\frac{P(+)}{P(-)}\right] + \ln\left[\frac{P(X/+)}{P(X/-)}\right] \\ &= \ln\left[\frac{P(+)}{P(-)}\right] + [b_0 + b_1 X_1 + \dots + b_J X_J]\end{aligned}$$

S'appuyer sur l'hypothèse semi-paramétrique

Ou S'appuyer sur la définition du LOGIT

→ Aboutissent à la même formulation (à une constante près)

$$a_0 = \ln\left[\frac{P(+)}{P(-)}\right] + b_0$$

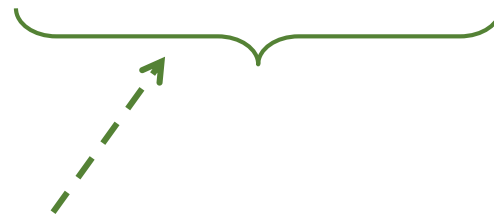
Les cas où la régression logistique est inopérante sont mieux circonscrites : ce sont les cas où les hypothèses de distribution ne sont absolument crédibles au regard des données (ex. distribution multimodales)



Données exemples pour ce support

Détection d'une maladie cardiaque

age	taux_max	angine	coeur
50	126	1	presence
49	126	0	presence
46	144	0	presence
49	139	0	presence
62	154	1	presence
35	156	1	presence
67	160	0	absence
65	140	0	absence
47	143	0	absence
58	165	0	absence
57	163	1	absence
59	145	0	absence
44	175	0	absence
41	153	0	absence
54	152	0	absence
52	169	0	absence
57	168	1	absence
50	158	0	absence
44	170	0	absence
49	171	0	absence



X1 : age du patient (quantitative)
X2 : taux max (quantitative)
X3 : angine de poitrine (binaire)



Y : (+ = présence, - = absence)

Estimation des paramètres

Pourquoi pas les MCO ?

Ω est notre échantillon
 ω est une observation
 $\text{Card}(\Omega)=n$

Dans un cas Y binaire (Positifs vs. Négatifs), nous pouvons coder

$$z(\omega) = \begin{cases} 1, & \text{si } y(\omega) = + \\ 0, & \text{si } y(\omega) = - \end{cases}$$

On constate aisément

$$E[Z(\omega)] = P[Y(\omega) = +]$$

Rapportée dans l'équation de régression

$$E[Z(\omega)] = P[Y(\omega) = +] = \underbrace{c_0}_{\text{endogène}} + \underbrace{c_1 X_1(\omega) + \dots + c_J X_J(\omega)}_{\text{exogènes}}$$

On devrait donc pouvoir mettre en place une régression qui permet d'estimer directement la probabilité d'appartenance $P(Y=+)$???

hélas, non...

- La combinaison linéaire varie entre $-\infty$ et $+\infty$, ce n'est pas une probabilité
- Dans l'échantillon, nous disposons de $Y(\omega)$ mais pas de $P[Y(\omega)=+]$ (Il faudrait que les données soient groupées – ou pondérées – pour que son estimation soit possible)
- Les hypothèses de la MCO, notamment l'homoscédasticité et la normalité des résidus posent problème : statistique inférentielle impossible (évaluation des coefficients, etc.)

Remarques sur la notation

Quelques précisions sur les notations et les expressions

$Y(\omega)$ est la modalité de Y prise par un individu ω , observé

$(X_1(\omega), \dots, X_J(\omega))$ est la description d'un individu ω , dans l'espace des variables explicatives

$P[Y(\omega) = +] = p_+$ est la probabilité a priori d'un individu d'être positif

$P[Y(\omega) = + / X] = \pi(X(\omega))$ est la probabilité qu'un individu ω quelconque soit +, c'est ce qu'on veut modéliser

$\ln \left[\frac{\pi(X(\omega))}{1 - \pi(X(\omega))} \right] = a_0 + a_1 X_1(\omega) + \dots + a_J X_J(\omega)$ est le LOGIT d'un individu ω

ou $\ln \left[\frac{\pi(X(\omega))}{1 - \pi(X(\omega))} \right] = X(\omega) \times a$ avec $a' = (a_0, a_1, \dots, a_J)$
 $X(\omega) = (1, X_1(\omega), \dots, X_J(\omega))$
 $-X_0(\omega) = 1$

---> On veut estimer à partir des n observations

$\hat{a}' = (\hat{a}_0, \hat{a}_1, \dots, \hat{a}_J)$

Estimation des paramètres

Définir la vraisemblance

Le modèle binomial

- (1) Parce que Y est binaire {+, -} ou $Y \in \{1, 0\}$ pour simplifier
- (2) Si Y était d'une autre nature, on utiliserait d'autres modèles (ex. Poisson, Multinomial, ...)

Pour un individu ω , on modélise la probabilité $P(Y/X)$ avec le modèle binomial

$$\pi(\omega)^{Y(\omega)} \times (1 - \pi(\omega))^{[1-Y(\omega)]}$$

$Y(\omega) = 1 \rightarrow P(Y=1/X) = \pi$
 $Y(\omega) = 0 \rightarrow P(Y=0/X) = 1-\pi$

La vraisemblance (LIKELIHOOD) pour un échantillon Ω (les observations sont i.i.d.)

$$L = \prod_{\omega} \pi^Y \times (1 - \pi)^{[1-Y]}$$

Interprétation ?
Valeur max. ?

La log-vraisemblance (LOG-LIKELIHOOD)

$$LL = \sum_{\omega} Y \times \ln(\pi) + [1 - Y] \times \ln(1 - \pi)$$

Estimation des paramètres

Méthode du maximum de vraisemblance

N'oublions pas que

$$\ln \left[\frac{\pi}{1 - \pi} \right] = Xa$$

On veut estimer à partir des n observations

-----> \hat{a}

Principe de la maximisation de la vraisemblance :
produire les paramètres de manière à maximiser la
quantité

$$LL = \sum_{\omega} Y \times \ln(\pi) + (1 - Y) \times \ln(1 - \pi)$$

\hat{a} est un EMV (estimateur du maximum de vraisemblance) avec toutes ses qualités :

- asymptotiquement sans biais
- variance minimale
- asymptotiquement normal (important pour l'inférence)

Remarque : On manipule souvent la quantité $[-2LL]$ que l'on appelle DEVIANCE (cf. analogie avec la SCR de la régression)

Un exemple sous EXCEL

				a0	a1	a2	a3
				14.494	-0.126	-0.064	1.779
age	taux_max	engine	coeur	cœur	C(X)	π	LL
50	126	1	presence	1	1.982	0.879	-0.129
49	126	0	presence	1	0.329	0.582	-0.542
46	144	0	presence	1	-0.438	0.392	-0.936
49	139	0	presence	1	-0.497	0.378	-0.972
62	154	1	presence	1	-1.305	0.213	-1.545
35	156	1	presence	1	1.960	0.877	-0.132
67	160	0	absence	0	-4.093	0.016	-0.017
65	140	0	absence	0	-2.571	0.071	-0.074
47	143	0	absence	0	-0.500	0.378	-0.474
58	165	0	absence	0	-3.280	0.036	-0.037
57	115	1	absence	0	1.802	0.858	-1.955
59	145	0	absence	0	-2.135	0.106	-0.112
44	175	0	absence	0	-2.157	0.104	-0.109
41	153	0	absence	0	-0.382	0.406	-0.520
54	152	0	absence	0	-1.952	0.124	-0.133
52	169	0	absence	0	-2.781	0.058	-0.060
57	168	1	absence	0	-1.566	0.173	-0.190
50	158	0	absence	0	-1.830	0.138	-0.149
44	170	0	absence	0	-1.839	0.137	-0.147
49	171	0	absence	0	-2.531	0.074	-0.077
					-2LL		16.618

\hat{a}

$$LL = \sum_{\omega} Y \times \ln(\pi) + [1 - Y] \times \ln(1 - \pi)$$

Valeur de $-2LL$ obtenue par minimisation avec le SOLVEUR

$$C = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3$$

$$\pi = \frac{e^C}{1 + e^C}$$

Estimation en pratique – Méthode de Newton Raphson

Il n’y a pas de méthode directe pour optimiser **L**

Passer par des méthodes numériques : la méthode de Newton-Raphson est la plus utilisée


Solutions à l'étape i et (i+1)

$$a^{i+1} = a^i - \left(\frac{\partial^2 L}{\partial a \cdot \partial a'} \right)^{-1} \times \frac{\partial L}{\partial a}$$

Vecteur des dérivées partielles premières
Vecteur gradient – dim = J+1

Matrice des dérivées partielles secondes
Matrice hessienne – dim = (J+1)x(J+1)
Son inverse est la matrice de variance covariance des coefficients

$$\frac{\partial L}{\partial a_j} = \sum_{\omega} [y(\omega) - \pi(\omega)] \times x_j(\omega)$$

Dont toutes les composantes sont égales à 0 lorsqu'on a trouvé la solution. 

- Règle d'arrêt : nombre d'itérations max., ou décroissante « trop faible » de L, ou écart faible entre les deux estimations successives \hat{a}
- D'un logiciel à l'autre, les résultats peuvent être différents (précision des calculs, choix de l'algorithme d'optimisation, règle d'arrêt, etc.)
- Beaucoup de calculs complexes, donc risque d'erreur !!!
- Lorsque la discrimination est parfaite, la matrice hessienne n'est plus inversible : le logiciel « plante » !!!

Evaluation "empirique" de la régression

Bilan global de la régression basé sur les prédictions et la déviance

Première évaluation – La matrice de confusion + Mesures d'évaluation

Commune à toutes les techniques supervisées, permet les comparaisons entre méthodes (ex. Reg. Logistique vs. Arbre de décision, etc.)

				a0	a1	a2	a3		
				14.494	-0.126	-0.064	1.779		
age	taux_max	angine	coeur	cœur	C(X)	?	LL	Prédiction	
50	126	1	presence	1	1.98	0.88	-0.13	presence	
49	126	0	presence	1	0.33	0.58	-0.54	presence	
46	144	0	presence	1	-0.44	0.39	-0.94	absence	
49	139	0	presence	1	-0.50	0.38	-0.97	absence	
62	154	1	presence	1	-1.30	0.21	-1.54	absence	
35	156	1	presence	1	1.96	0.88	-0.13	presence	
67	160	0	absence	0	-4.09	0.02	-0.02	absence	
65	140	0	absence	0	-2.57	0.07	-0.07	absence	
47	143	0	absence	0	-0.50	0.38	-0.47	absence	
58	165	0	absence	0	-3.28	0.04	-0.04	absence	
57	115	1	absence	0	1.80	0.86	-1.95	presence	
59	145	0	absence	0	-2.13	0.11	-0.11	absence	
44	175	0	absence	0	-2.16	0.10	-0.11	absence	
41	153	0	absence	0	-0.38	0.41	-0.52	absence	
54	152	0	absence	0	-1.95	0.12	-0.13	absence	
52	169	0	absence	0	-2.78	0.06	-0.06	absence	
57	168	1	absence	0	-1.57	0.17	-0.19	absence	
50	158	0	absence	0	-1.83	0.14	-0.15	absence	
44	170	0	absence	0	-1.84	0.14	-0.15	absence	
49	171	0	absence	0	-2.53	0.07	-0.08	absence	

coeur	Prédiction		Total
	presence	absence	
presence	3	3	6
absence	1	13	14
Total	4	16	20

Taux d'erre	0.20
Sensibilité	0.50
Spécificité	0.93
Précision	0.75

Si $C(X) > 0$ Alors Prédiction = « présence »

Ou, de manière équivalente : Si $\pi(X) > 0.5$ Alors Prédiction = « présence »



Mieux vaut réaliser cette évaluation sur un fichier test, n'ayant pas participé à la construction du modèle : les indicateurs sont non-biaisés

Deuxième évaluation – Les pseudo-R²

Modèle de référence : le modèle initial

Objectif : Produire des indicateurs similaires au R², coefficient de détermination de la régression linéaire.

Comment ? Comparer le modèle avec le modèle initial (trivial) constitué de la seule constante.

Modèle trivial : on n'utilise pas les X pour prédire Y

$$\text{LOGIT}(\pi) = \ln\left[\frac{\pi}{1-\pi}\right] = a_0$$

$$P(Y / X) = P(Y)$$



Estimation

$$\hat{a}_0 = \ln\left[\frac{\hat{p}_+}{1-\hat{p}_+}\right] = \ln\left[\frac{n_+}{n_-}\right]$$

Log-vraisemblance

$$\begin{aligned} LL(0) &= \sum_{\omega} Y(\omega) \times \ln(\hat{p}_+) + [1 - Y(\omega)] \times \ln(1 - \hat{p}_+) \\ &= n \times \ln(1 - \hat{p}_+) + n_+ \times \ln\left(\frac{\hat{p}_+}{1 - \hat{p}_+}\right) \end{aligned}$$

Estimation « classique »

				a0	a1	a2	a3
				-0.847	0.000	0.000	0.000
age	taux_max	angine	cœur	cœur	C(X)	π	LL
50	126	1	presence	1	-0.847	0.300	-1.204
49	126	0	presence	1	-0.847	0.300	-1.204
46	144	0	presence	1	-0.847	0.300	-1.204
49	139	0	presence	1	-0.847	0.300	-1.204
62	154	1	presence	1	-0.847	0.300	-1.204
35	156	1	presence	1	-0.847	0.300	-1.204
67	160	0	absence	0	-0.847	0.300	-0.357
65	140	0	absence	0	-0.847	0.300	-0.357
47	143	0	absence	0	-0.847	0.300	-0.357
58	165	0	absence	0	-0.847	0.300	-0.357
57	115	1	absence	0	-0.847	0.300	-0.357
59	145	0	absence	0	-0.847	0.300	-0.357
44	175	0	absence	0	-0.847	0.300	-0.357
41	153	0	absence	0	-0.847	0.300	-0.357
54	152	0	absence	0	-0.847	0.300	-0.357
52	169	0	absence	0	-0.847	0.300	-0.357
57	168	1	absence	0	-0.847	0.300	-0.357
50	158	0	absence	0	-0.847	0.300	-0.357
44	170	0	absence	0	-0.847	0.300	-0.357
49	171	0	absence	0	-0.847	0.300	-0.357
						-2LL	24.435

Estimation « directe »

$$\hat{a}_0 = \ln\left[\frac{6}{14}\right] \neq -0.847$$

$$-2 \times LL(0) = -2 \times \left[20 \times \ln(1 - 0.3) + 6 \times \ln\left(\frac{0.3}{1 - 0.3}\right) \right] \neq 24.435$$

Deuxième évaluation – Les pseudo-R²

Quelques indicateurs

McFadden's R²

$$R_{MF}^2 = 1 - \frac{LL(a)}{LL(0)}$$

Min = 0 si LL(a) = LL(0)

Max = 1 si L(a) = 1 c.à.d. LL(a) = 0

Cf. l'analogie avec le R² = 1 – SCR/SCT de la régression

COX and Snell's R²

$$R_{CS}^2 = 1 - \left(\frac{L(0)}{L(a)} \right)^{\frac{2}{n}}$$

Min = 0

Max si L(a) = 1 → $\max [R_{CS}^2] = 1 - (L(0))^{\frac{2}{n}}$

Nagelkerke's R²

$$R_N^2 = \frac{R_{CS}^2}{\max [R_{CS}^2]}$$

Min = 0

Max = 1

Prédiction de maladie
cardiaque



LL(0)	-12.21729
L(0)	4.94E-06

LL(a)	-8.308844
L(a)	0.000246

R ² mf	0.319911
R ² cs	0.323514
R ² n	0.458704

Plus on s'écarte de 0, mieux c'est. Mais on ne sait pas trop quoi conclure, c'est « suffisamment » bien ou pas ?

Evaluation "empirique" de la régression

Bilan basé sur la qualité des scores fournis par la régression

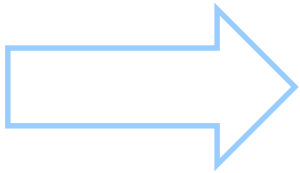
Evaluation des scores

Adéquation entre « scores » fournis par le modèle et « scores » observés dans l'échantillon

La régression fournit pour chaque individu ω le score $\pi(\omega)$ qui est une estimation de la probabilité $P(Y = + / X)$.

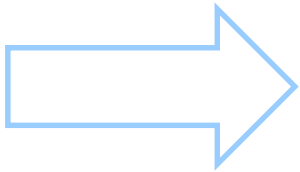
→ Question : est-ce cette estimation est de bonne qualité ?

Objectif du test de conformité



Confronter les scores estimés par le modèle avec le score observé dans l'échantillon

Comment ?



Organiser les données par paquets selon les scores, comparer dans chaque groupe les scores estimés (modèle) et observés (proportion de positifs)

Note : Il est d'usage de travailler directement sur l'échantillon d'apprentissage pour ces calculs, mais on pourrait tout aussi bien fonctionner avec un échantillon test.

Diagramme de fiabilité (Reliability diagram)

Extrait de l'ouvrage « Pratique de la régression logistique » (section 2.2)

N = 100 observations

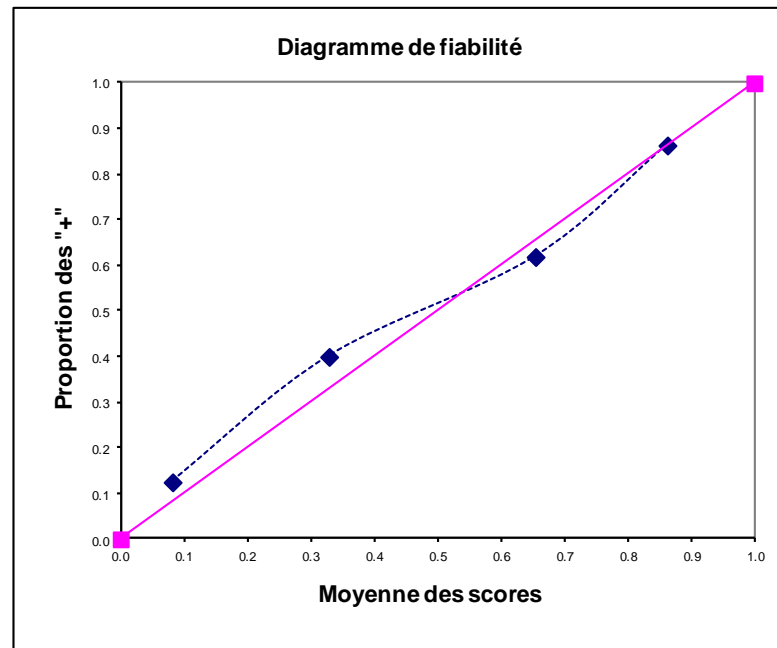
Pour chaque individu ω , on connaît sa classe $Y(\omega)$ (1 ou 0)

Et on a calculé son score $\pi(\omega)$ à partir du modèle

Proportion des "+"	
Moyenne de Y	
PI	Total
0-0.25	0.1250
0.25-0.5	0.4000
0.5-0.75	0.6190
0.75-1	0.8636

Moyenne des scores	
Moyenne de PI	
PI	Total
0-0.25	0.0820
0.25-0.5	0.3289
0.5-0.75	0.6548
0.75-1	0.8629

Observations subdivisées en $G = 4$ blocs via les scores (intervalles de largeur égales : 0-0.25, 0.25-0.5, ...), on compare dans un graphique : la moyenne des scores et la proportion des positifs



Si les points sont alignés sur la diagonale principale, les scores sont de bonne qualité.

Test de Hosmer & Lemeshow

Extrait de l'ouvrage « Pratique de la régression logistique » (section 2.3)

Exemple : Groupe 1

$$n_1 = 10, n_{1(+)} = 2, n_{1(-)} = 8 = 10 - 2$$

$$\text{Scores}_1(+) = \text{Somme}(\text{scores}_1) = 1.1985$$

$$\text{Scores}_1(-) = 10 - \text{Scores}_1(+) = 8.8015$$

$$\text{Statistique (HL)} = (2 - 1.1985)^2/1.1985 + \dots + (9 - 9.3864)^2/9.3864 + (8 - 8.8015)^2/8.8015 + \dots + (1 - 0.6136)^2/0.6136 = 7.8291$$

Groupe	Décile	Effectif	Positifs		Négatifs	
			Observés	Théoriques	Observés	Théoriques
1	0.2871	10	2	1.1985	8	8.8015
2	0.6249	10	4	5.0945	6	4.9055
3	0.7344	10	6	6.7886	4	3.2114
4	0.7874	10	7	7.5886	3	2.4114
5	0.8146	10	7	8.0422	3	1.9578
6	0.8485	10	10	8.3373	0	1.6627
7	0.8775	10	10	8.6720	0	1.3280
8	0.8917	10	8	8.8564	2	1.1436
9	0.9101	10	10	9.0357	0	0.9643
10	1.0000	10	9	9.3864	1	0.6136

Observations subdivisées en $G = 10$ blocs via les scores (intervalles de fréquences égales : seuils = déciles), on compare par calcul la somme des scores (resp. effectifs – somme des scores) et le nombre de positifs (resp. négatifs)



Sous (H_0 : le modèle est compatible avec les données), $HL \cong \chi^2 (G - 2)$

Pour notre exemple à 5%, $\chi^2_{0,95}(8) = 15.51$; le modèle est cohérent avec les données.

Evaluation "statistique" de la régression

Le modèle est-il statistiquement significatif ?

Telle ou telles variables sont elles pertinentes ?

Évaluation « statistique »

S'appuyer sur le modèle probabiliste issu de la maximisation de la vraisemblance

Croiser deux points de vue

Évaluer globalement le modèle c.-à-d.

$H_0 : a_1 = a_2 = \dots = a_j = 0$

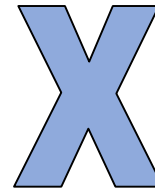
Aucune variable n'est pertinente

Comparer les vraisemblances

Test du Rapport de vraisemblance

Évaluer individuellement les variables c.-à-d.

$H_0 : a_j = 0$



S'appuyer sur la normalité asymptotique
des estimateurs (on sait calculer la matrice
de variance co-variance des coefficients)

Évaluer un groupe de « q » variables c.-à-d.

$H_0 : a_j = \dots = a_{j+q} = 0$

Principe du rapport de vraisemblance

Comparer la vraisemblance des modèles emboîtés

$$LR = -2 \times \ln \left(\frac{L(\text{modèle réduit})}{L(\text{modèle complet})} \right)$$
$$= [-2 \times LL(\text{modèle réduit})] - [-2 \times LL(\text{modèle complet})]$$

LR est forcément positif. Pourquoi ?

Modèle à p variables ($p < J$)
c.-à-d. à $(p+1)$ paramètres estimés
D.D.L $\rightarrow n-p-1$

Modèle à J variables
c.-à-d. à $(J+1)$ paramètres estimés
D.D.L $\rightarrow n-J-1$

$$LR \cong \chi^2(J - p)$$

LR suit asymptotiquement une loi du KHI-2 à $(J-p)$ degrés de liberté.



Dans notre exemple : $-2LL(a) = 16.618$
Ce sera le modèle « complet » de référence.

Test du rapport de vraisemblance

Évaluation globale du modèle

Comparer le modèle complet avec le modèle constitué uniquement de la constante
c.-à-d. tester si tous les coefficients (à part la constante) peuvent être simultanément nuls

				a0	a1	a2	a3
				-0.847	0.000	0.000	0.000
age	taux_max	angine	cœur	cœur	C(x)	π	LL
50	126	1	presence	1	-0.847	0.300	-1.204
49	126	0	presence	1	-0.847	0.300	-1.204
46	144	0	presence	1	-0.847	0.300	-1.204
49	139	0	presence	1	-0.847	0.300	-1.204
62	154	1	presence	1	-0.847	0.300	-1.204
35	156	1	presence	1	-0.847	0.300	-1.204
67	160	0	absence	0	-0.847	0.300	-0.357
65	140	0	absence	0	-0.847	0.300	-0.357
47	143	0	absence	0	-0.847	0.300	-0.357
58	165	0	absence	0	-0.847	0.300	-0.357
57	115	1	absence	0	-0.847	0.300	-0.357
59	145	0	absence	0	-0.847	0.300	-0.357
44	175	0	absence	0	-0.847	0.300	-0.357
41	153	0	absence	0	-0.847	0.300	-0.357
54	152	0	absence	0	-0.847	0.300	-0.357
52	169	0	absence	0	-0.847	0.300	-0.357
57	168	1	absence	0	-0.847	0.300	-0.357
50	158	0	absence	0	-0.847	0.300	-0.357
44	170	0	absence	0	-0.847	0.300	-0.357
49	171	0	absence	0	-0.847	0.300	-0.357
				-2LL			24.435

(A) : -2 LL(0)	24.4346
(B) : -2LL(age,taux,angine)	16.6177
LR: (A) - (B)	7.8169
d.d.l	3
p-value	0.049952

Le modèle est globalement significatif à 5%
c.-à-d. $H_0 : a_1 = a_2 = a_3 = 0$ n'est pas compatible avec les données

Test du rapport de vraisemblance

Évaluer individuellement les variables

Comparer le modèle complet avec le modèle sans la variable à évaluer
c.-à-d. tester si le coefficient associé à la variable est significativement différent de 0
ex. Tester le rôle de la variable « âge »

				a0	a1	a2	a3
				7.450	0.000	-0.059	1.551
age	taux_max	engine	cœur	cœur	$C(X)$	π	LL
50	126	1	presence	1	1.618	0.834	-0.181
49	126	0	presence	1	0.066	0.517	-0.661
46	144	0	presence	1	-0.989	0.271	-1.305
49	139	0	presence	1	-0.696	0.333	-1.100
62	154	1	presence	1	-0.023	0.494	-0.705
35	156	1	presence	1	-0.141	0.465	-0.766
67	160	0	absence	0	-1.926	0.127	-0.136
65	140	0	absence	0	-0.754	0.320	-0.385
47	143	0	absence	0	-0.930	0.283	-0.333
58	165	0	absence	0	-2.219	0.098	-0.103
57	115	1	absence	0	2.262	0.906	-2.361
59	145	0	absence	0	-1.047	0.260	-0.301
44	175	0	absence	0	-2.805	0.057	-0.059
41	153	0	absence	0	-1.516	0.180	-0.198
54	152	0	absence	0	-1.458	0.189	-0.209
52	169	0	absence	0	-2.454	0.079	-0.082
57	168	1	absence	0	-0.844	0.301	-0.358
50	158	0	absence	0	-1.809	0.141	-0.152
44	170	0	absence	0	-2.512	0.075	-0.078
49	171	0	absence	0	-2.571	0.071	-0.074
					-2LL		19.094

(A) : -2 LL(taux,engine)	19.0938
(B) : -2LL(age,taux,engine)	16.6177
LR : (A) - (B)	2.4761
d.d.l	1
p-value	0.115585

La variable « âge » n'est pas significative à 5%
c.-à-d. $H_0 : a_1 = 0$ est compatible avec les données

Test du rapport de vraisemblance

Évaluer un groupe de variables

Comparer le modèle complet avec le modèle sans les variables à évaluer

c.-à-d. tester si les coefficients associés aux variables sont significativement différents de 0

ex. Tester le rôle simultané des variables « âge » et « taux max »

				a0	a1	a2	a3
				-1.386	0.000	0.000	1.792
age	taux_max	angine	cœur	cœur	C(X)	n	LL
50	126	1	presence	1	0.405	0.600	-0.511
49	126	0	presence	1	-1.386	0.200	-1.609
46	144	0	presence	1	-1.386	0.200	-1.609
49	139	0	presence	1	-1.386	0.200	-1.609
62	154	1	presence	1	0.405	0.600	-0.511
35	156	1	presence	1	0.405	0.600	-0.511
67	160	0	absence	0	-1.386	0.200	-0.223
65	140	0	absence	0	-1.386	0.200	-0.223
47	143	0	absence	0	-1.386	0.200	-0.223
58	165	0	absence	0	-1.386	0.200	-0.223
57	115	1	absence	0	0.405	0.600	-0.916
59	145	0	absence	0	-1.386	0.200	-0.223
44	175	0	absence	0	-1.386	0.200	-0.223
41	153	0	absence	0	-1.386	0.200	-0.223
54	152	0	absence	0	-1.386	0.200	-0.223
52	169	0	absence	0	-1.386	0.200	-0.223
57	168	1	absence	0	0.405	0.600	-0.916
50	158	0	absence	0	-1.386	0.200	-0.223
44	170	0	absence	0	-1.386	0.200	-0.223
49	171	0	absence	0	-1.386	0.200	-0.223
					-2LL		21.742

(A) : -2 LL(angine)	21.7422
(B) : -2LL(age,taux,angine)	16.6177
LR: (A) - (B)	5.1245
d.d.l	2
p-value	0.077131

Les variables « âge » et « taux max » ne sont pas significatifs à 5%
c.-à-d. l'hypothèse nulle $H_0 : a_1 = a_2 = 0$ est compatible avec les données

Tests fondés sur la normalité asymptotique des coefficients

Principe - Test de Wald

EMV asymptotiquement normaux.

Le vecteur des coefficients « a » suit une loi normale multidimensionnelle de matrice variance covariance = inverse de la matrice Hessienne (matrice des dérivées secondes de la vraisemblance / aux coefficients)

$$H = X'VX$$

(J+1)x(J+1)

Matrice des variables explicatives, en première colonne la constante.

Matrice diagonale de taille (n x n), formé par les $\pi(1 - \pi)$ estimés par la régression

X'VX			
2.61	130.24	386.29	0.65
130.24	6615.32	19210.78	34.59
386.29	19210.78	57708.88	94.12
0.65	34.59	94.12	0.65

$$\hat{\Sigma} = H^{-1}$$

(J+1) x (J+1)

C'est la matrice de variance co-variance des coefficients estimés. En particulier, sur la diagonale principale, nous observons la variance des coefficients

inv(X'VX)			
63.2765	-0.4882	-0.2627	1.0563
-0.4882	0.0088	0.0004	-0.0413
-0.2627	0.0004	0.0016	0.0030
1.0563	-0.0413	0.0030	2.2635

Ecart type coef.	
const.	7.9547
age	0.0938
taux_max	0.0404
angine	1.5045

$$\sqrt{2.2635} = 1.5045$$

Test de Wald

Évaluer la significativité d'un groupe de « q » variables

$$H_0 : a_j = a_{j+1} = \dots = a_{j+q} = 0$$

La statistique du test (Wald) suit une loi du KHI-2 à q degrés de liberté.

$$W_{(q)} = \hat{a}'_{(q)} \Sigma_{(q)}^{-1} \hat{a}_{(q)} \cong \chi^2(q)$$

$\Sigma_{(q)}$ Sous-matrice des var-covar des q coefficients à évaluer

Sous-vecteur des q coefficients à évaluer

Pour rappel, dans notre exemple Cœur = f(age, taux max, engine)

$$\hat{a} = \begin{pmatrix} 14.494 \\ -0.126 \\ -0.064 \\ 1.779 \end{pmatrix}$$

Ex. Tester $a_1 = a_2 = 0$

$$H_0 : \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

SIGMA				
63.2765	-0.4882	-0.2627	1.0563	
-0.4882	0.0088	0.0004	-0.0413	
-0.2627	0.0004	0.0016	0.0030	
1.0563	-0.0413	0.0030	2.2635	

$$\Sigma_{(2)} = \begin{pmatrix} 0.0088 & 0.0004 \\ 0.0004 & 0.0016 \end{pmatrix} \rightarrow \Sigma_{(2)}^{-1} = \begin{pmatrix} 114.97 & -28.58 \\ -28.58 & 618.40 \end{pmatrix}$$

Calcul de la statistique de Wald

$$W = \begin{pmatrix} -0.126 & -0.064 \end{pmatrix} \begin{pmatrix} 114.97 & -28.58 \\ -28.58 & 618.40 \end{pmatrix} \begin{pmatrix} -0.126 \\ -0.064 \end{pmatrix} = 3.8565$$

wald	3.856537
ddl	2
p-value	0.145400

Les coefficients ne sont pas simultanément significativement différents de 0 à 5% → H0 est compatible avec nos données

Test de Wald

Significativité d'une variable

$$H_0 : a_j = 0$$

$$W_j = \frac{\hat{a}_j^2}{\hat{\sigma}_j^2} \cong \chi^2(1)$$

C'est la valeur lue sur la diagonale principale de la matrice de var-covar des coefficients estimés

Ex. Tester $a_1 = 0$

SIGMA				
63.2765	-0.4882	-0.2627	1.0563	
-0.4882	0.0088	0.0004	-0.0413	
-0.2627	0.0004	0.0016	0.0030	
1.0563	-0.0413	0.0030	2.2635	

$$W_1 = \frac{(-0.126)^2}{0.0088} = 1.7938$$

wald	1.793823
ddl	1
p-value	0.180461

A 5%, la variable « âge » n'est pas significative

Test de Wald

Significativité globale du modèle

$$H_0 : a_1 = a_2 = \dots = a_j = 0$$

Tous les coefficients, mis à part la constante, peuvent ils être simultanément égaux à 0 ?

$$W_{(J)} = \hat{a}'_{(J)} \Sigma_{(J)}^{-1} \hat{a}_{(J)} \cong \chi^2(J)$$

Généralisation du test de q variables

Inversion de la sous-matrice

Ex. Tester $a_1 = a_2 = a_3 = 0$

SIGMA			
63.2765	-0.4882	-0.2627	1.0563
-0.4882	0.0088	0.0004	-0.0413
-0.2627	0.0004	0.0016	0.0030
1.0563	-0.0413	0.0030	2.2635

$$\Sigma_{(J)}^{-1} = \begin{pmatrix} 126.37 & -35.73 & 2.36 \\ -35.73 & 622.89 & -1.48 \\ 2.36 & -1.48 & 0.49 \end{pmatrix}$$

Calcul de la statistique de Wald

$$W = \begin{pmatrix} -0.126 & -0.064 & 1.779 \end{pmatrix} \begin{pmatrix} 126.37 & -35.73 & 2.36 \\ -35.73 & 622.89 & -1.48 \\ 2.36 & -1.48 & 0.49 \end{pmatrix} \begin{pmatrix} -0.126 \\ -0.064 \\ 1.779 \end{pmatrix} = 4.762$$

wald	4.762383
ddl	3
p-value	0.190047

Le modèle n'est pas globalement significatif à 5% → H_0 est compatible avec nos données

Bilan – Évaluation statistique

Test de Rapport de vraisemblance

+ Puissant

-- Plus gourmand en ressources de calcul (reconstruire le modèle à chaque fois)

Test de Wald

-- Moins puissant c.-à-d. plus conservateur, favorise H0

-- Quand la valeur du coef. est élevée, l'estimation de l'écart-type gonfle exagérément, on élimine à tort la variable

-- Pas très bon lorsque les effectifs sont petits (comme c'est le cas ici)

+ Moins gourmand en ressources (inversion de matrice quand même)

Tests à 5%	Rapp. Vraisemblance	Wald
Signif. Globale	Rejet H0 → p-value = 0.049952	Accep. H0 → p-value = 0.190047
Signif. « âge »	Accep. H0 → p-value = 0.1156	Accep. H0 → p-value = 0.1805
Signif. « âge et taux max »	Accep. H0 → p-value = 0.0771	Accep. H0 → p-value = 0.1454

TANAGRA

Adjustement quality

Predicted attribute	coeur
Positive value	presence
Number of examples	20
-2 Log-Likelihood (-2LL)	
Intercept only	24.4346
Model	16.6177
AIC & BIC	
AIC (Model)	24.6177
BIC (Model)	28.6006
Model Chi² test	
Chi-2	7.8169
d.f.	3
P(>Chi-2)	0.0500
R²-like	
McFadden's R²	0.3199
Cox and Snell's R²	0.3235
Nagelkerke's R²	0.4587

Déviante du modèle réduit à la constante seule

Déviante du modèle

KHI-2 du rapport de vraisemblance

Pseudo-R²

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	14.493790	-	-	-
age	-0.125634	0.0938	1.7938	0.1805
taux_max	-0.063560	0.0404	2.4696	0.1161
engine	1.779013	1.5045	1.3982	0.2370

R

```
Call:
glm(formula = coeur ~ age + taux_max + engine, family = binomial,
     data = donnees)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9773  -0.5437  -0.3876   0.5093   1.7577

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  14.49379    7.95464   1.822  0.0684 .
age          -0.12563    0.09380  -1.339  0.1805
taux_max     -0.06356    0.04045  -1.572  0.1161
engine        1.77901    1.50449   1.182  0.2370
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.435  on 19  degrees of freedom
Residual deviance: 16.618  on 16  degrees of freedom
AIC: 24.618

Number of Fisher Scoring iterations: 5
```



Lecture et interprétation des coefficients

Ce qui fait le succès de la régression logistique

Risque relatif, odds, odds-ratio

Quelques définitions

Nombre de coeur	angine		
cœur2	1	0	Total
1	3	3	6
0	2	12	14
Total	5	15	20

Y / X	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	n

Risque relatif

$$RR = \frac{P(+/1)}{P(+/0)} = \frac{a/(a+c)}{b/(b+d)}$$
$$= \frac{3/5}{3/15} = 3$$

Indique le surcroît de « chances » d'être positif du groupe « exposé » par rapport au groupe « témoin »: les personnes qui ont une angine de poitrine lors des efforts ont **3 fois plus de chances** (que les autres) d'avoir une maladie cardiaque.

Odds

$$Odds(+/1) = \frac{P(+/1)}{P(-/1)} = \frac{a/(a+c)}{c/(a+c)}$$
$$= \frac{3/5}{2/5} = 1.5$$

Dans le groupe de personnes ayant une angine de poitrine lors des efforts, on a **1.5 fois plus de chances d'avoir une maladie cardiaque que de ne pas en avoir**. De la même manière, on peut définir $Odds(+/0) = 3/12 = 0.25$

Odds-Ratio

$$OR(1/0) = \frac{Odds(+/1)}{Odds(+/0)} = \frac{a \times d}{b \times c}$$
$$= \frac{3 \times 12}{2 \times 3} = 6$$

Indique à peu près la même chose que le risque relatif : par rapport au groupe exposé, on a **6 fois plus de chances d'être positif** (que d'être négatif) dans le groupe témoin.

Quel indicateur choisir ?

Risque relatif, Odds, Odds-Ratio

Pourquoi choisir l'Odds-ratio ?

Lorsque p_+ (prévalence) est très petit, $OR \sim RR$.

→ Presque toujours, l'un ou l'autre, c'est la même chose.

$$a \ll c \rightarrow a + c \approx c$$

$$b \ll d \rightarrow b + d \approx d$$

$$\Rightarrow RR = \frac{a/(a+c)}{b/(b+d)} \approx \frac{a/c}{b/d} = \frac{a \times d}{b \times c} = OR$$

MAIS l'odds-ratio est invariant selon le mode d'échantillonnage



Tirage aléatoire			
cœur2 x angine	1	0	Total
1	3	3	6
0	2	12	14
Total	5	15	20

RR	3
----	---

Odds(+/1)	1.5
-----------	-----

Odds(+/0)	0.25
-----------	------

OR(+)	6
-------	---

Souvent un vœu pieux : tirage aléatoire à probabilités égales dans la population. Échantillon aléatoire.

Tirage retrospectif (presque) équilibré			
cœur2 x angine	1	0	Total
1	3	3	6
0	1	6	7
Total	5	9	13

RR	1.8
----	-----

Odds(+/1)	3
-----------	---

Odds(+/0)	0.5
-----------	-----

OR(+)	6
-------	---

Souvent pratiqué : on choisit l'effectif des positifs et des négatifs, et on échantillonne au hasard dans chaque groupe
→ l'OR reste de marbre !!!

Odds-ratio

Quel rapport avec la régression logistique ?

Calcul sur un tableau de contingence

Tirage aléatoire			
cœur2 x angine	1	0	Total
1	3	3	6
0	2	12	14
Total	5	15	20

--->

$$OR(1/0) = \frac{3 \times 12}{2 \times 3} = 6$$

Model Chi ² test	
Chi-2	2.6924
d.f.	1
P(>Chi-2)	0.1008

Régression logistique cœur = f(angine)

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.386294	-	-	-
angine	1.791759	1.118	2.5683	0.109

$$e^{1.791759} = 6$$

Le coefficient de la Reg.Log. s'interprète comme le logarithme de l'odds-ratio.
 → On peut mesurer directement le surcroît de risque qu'introduit chaque facteur explicatif (variable 1/0).

A partir de l'intervalle de confiance du coefficient (normalité asymptotique)
 On peut déduire l'intervalle de confiance de l'odds-ratio

Intervalle de confiance de l'odds-ratio (ex. à 5%)

$$bb(a) = 1.791759 - 1.96 \times 1.118 = -0.399$$

$$bh(a) = 1.791759 + 1.96 \times 1.118 = 3.983$$

--->

$$bb(OR) = e^{-0.399} = 0.67$$

$$bh(OR) = e^{3.983} = 53.68$$

$u_{0.975}$

Si l'intervalle contient la valeur « 1 », cela indique que l'influence du facteur sur la variable dépendante n'est pas significative au niveau de risque choisi.

Odds-Ratio

OR partiel et OR des variables continues

Équation comportant toutes les variables

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	14.493790	-	-	-
age	-0.125634	0.0938	1.7938	0.1805
taux_max	-0.063560	0.0404	2.4696	0.1161
angine	1.779013	1.5045	1.3982	0.2370

Pour les variables quantitatives, le coefficient se lit comme le logarithme de l'OR consécutive à l'augmentation d'une unité de la variable explicative.

→ Ici, étrangement, l'augmentation de l'âge d'une année, si l'on contrôle les valeurs de angine et taux max, va réduire le risque de présence de maladie OR = $\text{EXP}(-0.126) = 0.88$

→ Pour tempérer cette conclusion, on remarquera que la variable n'est pas significative, même à 10% (colinéarité ? Effectifs trop faibles ?)

A taux max et âge fixés (ou après avoir enlevé l'effet du taux max et de l'âge), l'OR de la variable angine est $\text{EXP}(1.779) \sim 5.92$

→ La même idée que la corrélation partielle.

Mieux vaut quand même travailler sur des fichiers avec des effectifs plus élevés dans les études réelles !!!

Odds-Ratio

Aller plus loin que les Odds-ratio – Lecture en termes de différentiel de probabilités

Nombre de coeur	angine		
coeur	1	0	Total général
présence	3	3	6
absence	2	12	14
Total général	5	15	20

Proba(Présence)	0.6	0.2
-----------------	-----	-----

Ecart	0.4
-------	-----

$$P(\text{cœur} = \text{présence} / \text{angine} = 0) = 3/15 = 0.2$$

$$P(\text{cœur} = \text{présence} / \text{angine} = 1) = 3/5 = 0.6$$

→ Quand « angine = 1 », la probabilité de la présence de la maladie augmente de $(0.6 - 0.2) = 0.4$

Comment obtenir ce résultat avec la régression logistique ?

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.386	0.6455	4.6123	0.0317
angine	1.792	1.118	2.5683	0.109

$$P(\text{cœur} = + / \text{angine} = 0) = 1/(1+\text{EXP}[-(-1.386)]) = 0.2$$

$$P(\text{cœur} = + / \text{angine} = 1) = 1/(1+\text{EXP}[-(-1.386+1.792)]) = 0.6$$

→ Quand « angine = 1 », la probabilité de la présence de la maladie augmente de $(0.6 - 0.2) = 0.4$

Traitement des variables explicatives nominales

On utilise un autre fichier avec 209 obs. ici.

A plus de 2 modalités

Que se passe-t-il lorsque la variable explicative est nominale à (K>2) modalités ?

→ Dans le tableau de contingence, on prend une modalité de référence, et on calcule les odds-ratio par rapport à cette modalité.

→ On traite (K- 1) tableaux 2 x 2.

Calcul direct dans un tableau croisé					
Nombre de cœur	chest_pain				
cœur	typ_angina	atyp_angina	asympt	_non_anginal	Total
presence	4	6	75	7	92
absence	2	59	27	29	117
Total	6	65	102	36	209

Odds(+/-)	2.000	0.102	2.778	0.241
OR(x/_non_anginal)	8.286	0.421	11.508	

Surcroît de risque de présence de maladie lorsque la douleur à la poitrine n'est pas de type « non anginale ».

Traduire cela dans la régression logistique ?

→ Utiliser un codage disjonctif 0/1 en prenant une modalité de référence.

→ Les coefficients sont des log(Odds-Ratio) par rapport à la modalité de référence.

Résultat de la régression logistique			
OR(Reg.Logistic)	8.286	0.421	11.508

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.421387	-	-	-
typ_angina	2.114534	0.963	4.8216	0.0281
atyp_angina	-0.864391	0.6008	2.07	0.1502
asympt	2.443038	0.4772	26.2106	0

Remarque : On peut tester la significativité « globale » de la variable en évaluant : « les 3 coefficients sont simultanément nuls ».

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
typ_angina	8.2857	1.255	54.7047
atyp_angina	0.4213	0.1298	1.3677
asympt	11.508	4.5166	29.3213

Il faut choisir le bon codage



Traitement des variables explicatives ordinales

Un variable qualitative ordinale à K modalités

3 niveaux possibles de SYSTOLIC
1 : normal ; 2 : élevé ; 3 : très élevé

On doit tenir compte de l'ordre de la modalité cette fois-ci.

→ Dans le tableau de contingence, on calcule l'odds-ratio par rapport à la modalité précédente c.-à-d. on quantifie le surcroît de risque en passant d'un niveau à l'autre

Calcul sur un tableau de contingence				
Nombre de coeur	systolic_level			
coeur		3	2	1
presence		14	31	47
absence		10	36	71
Total		24	67	118
				209

Odds	1.400	0.861	0.662
Odds-Ratio(précédent)	1.626	1.301	

Surcroît de risque de présence de maladie lorsque l'on passe d'un niveau de pression artérielle à l'autre

Régression logistique avec codage emboîté

Odds-Ratio	1.626	1.301
------------	--------------	--------------

Comment traduire cela dans la régression logistique ?

→ Utiliser un codage « emboîté »

→ Les coefficients sont des $\log(\text{Odds-Ratio})$ d'un passage d'une modalité de X à l'autre

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-0.412532	-	-	-
sys2	0.263001	0.3089	0.7251	0.3945
sys3	0.486004	0.4811	1.0205	0.3124

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
sys2	1.3008	0.7101	2.383
sys3	1.6258	0.6332	4.1743

Il faut choisir le bon codage



Ex.
Sys1 = 1 si Systolic >= 1, 0 sinon ; constante, éliminée
Sys2 = 1 si Systolic >= 2, 0 sinon
Sys3 = 1 si Systolic >= 3, 0 sinon

Tenir compte des interactions

Dans le LOGIT, les effets sont initialement additifs, comment le dépasser ?

Prendre en compte les interactions entre les variables binaires (mais aussi pour les variables nominales)

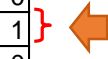
→ On crée autant de nouvelles variables qui prennent les valeurs 1/0 selon l'interaction que l'on veut analyser

→ On parle de modèle « saturé » lorsqu'on tient compte de toutes les interactions possibles

Ex. Effet conjoint d'une tension artérielle élevée et un rythme cardiaque maximum faible sur le diagnostic de la présence d'une maladie cardiaque.

coeur	high_bpress	low_max_rate	bpress_x_lmaxrate
positive	0	0	0
positive	1	1	1
negative	1	0	0
positive	0	0	0
negative	0	1	0
negative	0	0	0
negative	0	0	0
...

Il faut choisir le bon codage



Régression logistique avec les 3 variables

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-0.451985	-	-	-
high_bpress	0.200671	0.2996	0.4487	0.5030
low_max_rate	0.451985	0.6625	0.4654	0.4951
bpress_x_lmaxrate	2.101914	1.2609	2.7791	0.0955

L'effet conjoint pèse dans l'explication

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
high_bpress	1.2222	0.6794	2.1987
low_max_rate	1.5714	0.4289	5.7577
bpress_x_lmaxrate	8.1818	0.6912	96.8481

Une stratégie de modélisation simple serait de comparer des modèles emboîtés (notion de modèles « hiérarchiquement bien formulés ») :

1. M(bpress, max_rate)
2. M(bpress, max_rate, bpress * max_rate)

Cf. Sélection de variables et critères associés

Les possibilités en matière d'analyse et d'interprétation des interactions font le succès de la régression logistique



Comparer le poids relatif des variables

Coefficients non-standardisés vs. Coefficients standardisés

Cas de la régression linéaire

Prédire la consommation à partir du poids et de la puissance d'un véhicule

Modele	Puissance	Poids	Consommation
Daihatsu Cuore	32	650	5.7
Suzuki Swift 1.0 GLS	39	790	5.8
Fiat Panda Mambo L	29	730	6.1
VW Polo 1.4 60	44	955	6.5
Opel Corsa 1.2i Eco	33	895	6.8
Subaru Vivio 4WD	32	740	6.8
Toyota Corolla	55	1010	7.1
Opel Astra 1.6i 16V	74	1080	7.4
Peugeot 306 XS 108	74	1100	9
Renault Safrane 2.2. V	101	1500	11.7
Seat Ibiza 2.0 GTI	85	1075	9.5
VW Golt 2.0 GTI	85	1155	9.5
Citroen ZX Volcane	89	1140	8.8
Fiat Tempra 1.6 Liberty	65	1080	9.3
Fort Escort 1.4i PT	54	1110	8.6
Honda Civic bker 1.4	66	1140	7.7
Volvo 850 2.5	106	1370	10.8
Ford Fiesta 1.2 Zetec	55	940	6.6
Hyundai Sonata 3000	107	1400	11.7
Lancia K 3.0 LS	150	1550	11.9
Mazda Hachtback V	122	1330	10.8
Mitsubishi Galant	66	1300	7.6
Opel Omega 2.5i V6	125	1670	11.3
Peugeot 806 2.0	89	1560	10.8
Nissan Primera 2.0	92	1240	9.2
Seat Alhambra 2.0	85	1635	11.6
Toyota Previa salon	97	1800	12.8
Volvo 960 Kombi aut	125	1570	12.7

Moyenne	77.7143	1196.9643	9.0750
Ecart-type	32.2569	308.9928	2.2329

On sait interpréter ces coefficients (dérivée partielle première) mais, exprimés dans des unités différentes, on ne peut pas comparer leurs rôles (poids) respectifs c.-à-d. quelles sont les variables les plus importantes dans la régression ?

	Poids	Puissance	Constante
coef.	0.0044	0.0256	1.7696
ecart-type	0.0009	0.0083	
t	5.1596	3.0968	
p-value	0.00002	0.00478	

Les p-value nous donnent déjà une meilleure idée...

Solution 1 : Centrer et réduire les données

Coefficients standardisés à partir des données centrées-réduites

	Poids	Puissance	Constante
coef.	0.615016	0.369136	pas de constan

Poids ↗ de 1 écart-type → Conso. ↗ de 0.615 x é.t.
Puissance ↗ de 1 é.t. → Conso. ↗ de 0.369 x é.t.

Solution 2 : Corriger la solution initiale (Sans re-calcul de la régression)

$$\hat{a}_{x_j}^{std} = \hat{a}_{x_j} \times \frac{\hat{\sigma}_{x_j}}{\hat{\sigma}_y}$$

Coeff. Standardisés à partir de la formule de correction (cf. Ménard)

	Poids	Puissance	Constante
coef.	0.615016	0.369136	pas de constan

Comparer le poids relatif des variables

$$\text{Cœur} = f(\text{age}, \text{taux_max})$$

Coefficients standardisés pour la régression logistique (1)

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	16.254441	-	-	-
age	-0.12011	0.0843	2.0305	0.1542
taux_max	-0.074383	0.0387	3.6886	0.0548

é.t. (attrib.)
-
8.1620
16.6559

Cas de la régression logistique

On veut obtenir une interprétation du type : une augmentation de **1** é.t. de la variable X entraîne une variation de **m** é.t. du LOGIT.

On veut également (et surtout) pouvoir comparer les influences respectives des variables dans la régression.

Écart-type du LOGIT estimé
(prédiction du modèle)

-->	e.t. logit	1.4851
-----	------------	--------

Solution 0 : Comprendre le coefficient non standardisé

Test age	
age	40
taux-max	150
Logit	0.2926

age	41.00
taux-max	150
Logit	0.1725

Ecart(Logit)	-0.1201
Ecart ramené à l'écart-type	-0.0809

Test age	
age	40
taux-max	150
Logit	0.2926

age	48.16
taux-max	150
Logit	-0.6877

Ecart(Logit)	-0.9803
Ecart ramené à l'écart-type	-0.6601

Problèmes

- (1) $\pi \rightarrow 0 \Rightarrow \text{LOGIT} \rightarrow -\infty$
- (2) $\pi \rightarrow 1 \Rightarrow \text{LOGIT} \rightarrow +\infty$
- (3) Et de toute manière, on ne peut pas calculer le LOGIT sur les données observées

La correction des coefficients avec l'écart-type du LOGIT n'a pas de sens

Comparer le poids relatif des variables

Coef. standardisés pour la rég. logistique (2)

Solution 1 : Standardisation sur les explicatives seulement

$$\hat{a}_{x_j}^{std} = \hat{a}_{x_j} \times \hat{\sigma}_{x_j}$$

Mais constante non interprétable

Sol. 1 -- Standardisation sur les explicatives seulement

Attribute	Coef.
constant	<i>const. non nulle</i>
age	-0.980339123
taux_max	-1.238914506

Test age	
age	40
taux-max	150
Logit	0.2926

age	48.16
taux-max	150
Logit	-0.6877

Ecart(Logit)	-0.9803
Ecart ramené à l'écart-type	-0.6601

+1 é.t.

Quantifie l'écart absolu. Permet surtout de **comparer le poids relatif des variables** dans la prédiction de Y

Solution 2 : Standardisation sur les explicatives et le LOGIT

$$\hat{a}_{x_j}^{std} = \hat{a}_{x_j} \times \frac{\sigma_{x_j}}{\hat{\sigma}_{LOGIT}}$$

1.4851

Sol. 2 -- Standardisation avec LOGIT

Attribute	Coef.
constant	<i>const. non nulle</i>
age	-0.660102758
taux_max	-0.834212226

Quantifie l'écart en « écart-type ». Permet aussi de **comparer le poids relatif des variables**.

Solution 3 : Standardisation sur les explicatives et l'écart-type théorique que la répartition logistique (Solution SAS)

$$\hat{a}_{x_j}^{std} = \hat{a}_{x_j} \times \frac{\sigma_{x_j}}{\sigma_{théorique}}$$

Sol.3 -- Standardisation et correction param. Loi logistique (SAS)

<i>E.t. théorique</i>	1.813799364
Attribute	Coef.
constant	<i>const. non nulle</i>
age	-0.540489286
taux_max	-0.683049366

Permet **avant tout** de **comparer le poids relatif des variables**.

Écart-type théorique de la loi logistique standard :: Moyenne = 0 et écart-type →

$$\sigma_{théorique} = \frac{\pi}{\sqrt{3}} = 1.8138$$



Sélection de variables

Choisir les variables pertinentes pour la régression

Sélection de variables dans la pratique

Beaucoup de candidats, peu d'élus (souhaitables)

Dans les études réelles, beaucoup de variables disponibles, plus ou moins pertinentes, concurrentes... Trop de variables tue l'interprétation, il y a le danger du sur-apprentissage aussi.

Problème :

- Sélection « experte » manuelle basé sur Wald ou LR fastidieuse voire impossible
- On s'interdit de découvrir des relations auxquelles on n'a pas pensé

Solution :

- Utiliser des techniques numériques pour choisir les « meilleures » variables
 - Principe du Rasoir d'Occam : à performances égales, plus un modèle sera simple, plus il sera robuste ; plus aisée sera son interprétation également.
 - Attention : Ne pas prendre pour argent comptant la solution, plutôt se servir de l'outil pour bâtir des scénarios (qu'on présentera/discutera avec l'expert)
- Travail exploratoire : combinaison de variables, construction de nouvelles variables, etc.

2 approches

1. Sélection de variables = Optimisation d'un critère
2. S'appuyer sur les outils inférentiels = Significativité des variables

Sélection par optimisation

Critère AIC (Akaike) et BIC (Schwartz)

Constat

Plus le nombre de variables augmente, plus la déviance diminue (ou la vraisemblance augmente), même si la variable ajoutée n'est pas pertinente

Cf. par analogie la SCR ou le R^2 dans la régression linéaire, le degré de liberté diminue

Solution

Contrebalancer la réduction de la déviance avec une quantité traduisant la complexité du modèle → Le problème de sélection devient un problème d'optimisation (minimisation)

Critère AKAIKE

$$AIC = -2LL + 2 \times (J + 1)$$

Nombre de paramètres du modèle c.-à-d.
nombre de variables + 1

Critère BIC

$$BIC = -2LL + \ln(n) \times (J + 1)$$

Plus exigeant, pénalise plus la complexité →
sélectionne moins de variables.

Procédure

On va évaluer des successions de modèles emboîtés :

- En les ajoutant au fur et à mesure → FORWARD
- En les retirant au fur et à mesure → BACKWARD
- STEPWISE : En alternant FORWARD / BACKWARD c.-à-d. vérifier que chaque ajout de variable ne provoque pas la sortie d'une autre variable

Règle d'arrêt : l'adjonction ou le retrait d'une variable n'améliore plus le critère

Sélection par optimisation

Détail sélection FORWARD sous R

Start: (AIC=287.09)

```
coeur ~ 1
```

	Df	Deviance	AIC
+ chest_pain_asympt_1	1	207.86	211.86
+ exercice_angina_yes_1	1	210.88	214.88
+ chest_pain_atyp_angina_1	1	233.13	237.13
+ max_hrate	1	256.55	260.55
+ chest_pain_non_anginal_1	1	273.82	277.82
+ age	1	277.68	281.68
+ blood_sugar_f_1	1	280.69	284.69
+ restbpress	1	282.60	286.60
<none>		285.09	287.09
+ restecg_left_vent_hyper_1	1	283.81	287.81
+ restecg_normal_1	1	284.09	288.09

Step: AIC=211.86

```
coeur ~ chest_pain_asympt_1
```

	Df	Deviance	AIC
+ exercice_angina_yes_1	1	177.59	183.59
+ max_hrate	1	202.85	208.85
+ blood_sugar_f_1	1	203.16	209.16
+ chest_pain_atyp_angina_1	1	203.47	209.47
<none>		207.86	211.86
+ age	1	205.98	211.98
+ restbpress	1	206.59	212.59
+ chest_pain_non_anginal_1	1	207.08	213.08
+ restecg_normal_1	1	207.31	213.31
+ restecg_left_vent_hyper_1	1	207.68	213.68

Step: AIC=183.59

```
coeur ~ chest_pain_asympt_1 + exercice_angina_yes_1
```

	Df	Deviance	AIC
+ chest_pain_atyp_angina_1	1	172.93	180.93

```
heart <- read.table(file="heart_for_var_selection.txt",sep="\t",header=TRUE,dec=".")
#description des modèles
str_constant <- "~1"
str_full <- "~age+restbpress+max_hrate+chest_pain_asympt_1+chest_pain_atyp_angina_1+..."
#départ modèle avec la seule constante + sélection forward
modele <- glm(coeur ~1, data = heart, family = binomial)
modele.forward <- stepAIC(modele,scope = list(lower = str_constant, upper = str_full),
trace = TRUE, data = heart, direction = "forward")
summary(modele.forward)
```

AIC de départ, modèle initial : 287.9

Meilleure variable : « chest_pain_asympt_1 »
AIC de M(chest_pain_asympt_1) = 211.86
Point de départ d'une nouvelle recherche

Deuxième meilleure variable, acceptée puisque AIC continue à diminuer : « exercice_angina_yes_1 »
AIC = 183.59

Arrêt lorsque AIC ne diminue plus !!!

Sélection par optimisation

Comparaison des solutions : FORWARD, BACKWARD, BOTH (#STEPWISE)

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.3876     1.1683   1.188  0.23497
chest_pain_asympt_1 -0.2709     0.9811  -0.276  0.78249
exercice_angina_yes_1  2.2536     0.4331   5.204 1.95e-07 ***
chest_pain_atyp_angina_1 -3.1051     1.0511  -2.954  0.00314 **
chest_pain_non_anginal_1 -2.1765     1.0459  -2.081  0.03744 *
blood_sugar_f_1   -1.1871     0.8175  -1.452  0.14646
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family)

Null deviance: 285.09 on 207 degrees of freedom
Residual deviance: 165.69 on 202 degrees of freedom
AIC: 177.69
```

FORWARD

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.1628     0.8261   1.408 0.159255
exercice_angina_yes_1  2.2362     0.4290   5.212 1.86e-07 ***
chest_pain_atyp_angina_1 -2.8548     0.5293  -5.394 6.90e-08 ***
chest_pain_non_anginal_1 -1.9267     0.5218  -3.692 0.000222 ***
blood_sugar_f_1   -1.2097     0.8092  -1.495 0.134923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 285.09 on 207 degrees of freedom
Residual deviance: 165.77 on 203 degrees of freedom
AIC: 175.77
```

STEPWISE

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.1628     0.8261   1.408 0.159255
chest_pain_atyp_angina_1 -2.8548     0.5293  -5.394 6.90e-08 ***
chest_pain_non_anginal_1 -1.9267     0.5218  -3.692 0.000222 ***
blood_sugar_f_1   -1.2097     0.8092  -1.495 0.134923
exercice_angina_yes_1  2.2362     0.4290   5.212 1.86e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 285.09 on 207 degrees of freedom
Residual deviance: 165.77 on 203 degrees of freedom
AIC: 175.77
```

BACKWARD

Bilan

La solution diffère selon le sens de la recherche (normal)

Une variable choisie par AIC n'est pas forcément significative dans la régression ☹️☹️

Gourmandise en temps de calcul : chaque variable à tester (intro ou sortie) → une régression logistique ☹️☹️☹️

Sélection basée sur des critères statistiques

Test du score et Test de Wald

Objectif

- (1) Effectuer une sélection de variables s'appuyant sur des critères statistiques sans avoir à réaliser J^2 régressions : ce serait le cas si on utilisait à tort et à travers le rapport de vraisemblance (LR) → critique surtout sur les grandes bases
- (2) Produire un résultat cohérent avec les tests classiques de significativité
- (3) Avec toujours les mêmes stratégies FORWARD, BACKWARD, STEPWISE

Retrait d'une variable

- (1) La statistique de WALD permet de choisir la variable à éliminer sans avoir à relancer la régression
- (2) Dans une stratégie BACKWARD pure, on n'aurait que J régressions à effectuer

Ajout d'une variable : Test du Score

- (1) Avec le LR : à l'étape p , si on veut ajouter la $(p+1)$ -ème variable, il faudrait effectuer $(J-p)$ régressions. A éviter.
- (2) Principe du Test du SCORE : Utiliser les résultats de la régression à p variables pour calculer les SCORES de chaque $(J-p)$ variable restantes, choisir celle qui a le meilleur score → en FORWARD pur, on aurait au pire J régressions à effectuer

Principe du Test du Score

L'idée de(s) la variable(s) supplémentaire(s)

$$U_j = \frac{\partial L}{\partial a_j} = \sum_{\omega} [y(\omega) - \pi(\omega)] \times x_j(\omega)$$

- Étape courante : effectuer la R.L. avec les p variables déjà sélectionnées
- Calculer, en intégrant la (p+1)-ème variable à évaluer (comme var. supplémentaire)
 - Le vecteur gradient U
 - La matrice Hessienne H
 - La matrice de variance covariance des coefficients $\Sigma = H^{-1}$
 - En déduire la statistique du Score $S = U' \Sigma U$
- S suit une loi du KHI-2 à 1 degré de liberté

$$H(j_1, j_2) = \sum_{\omega} x_{j_1}(\omega) \times x_{j_2}(\omega) \times \pi(\omega) \times [1 - \pi(\omega)]$$

→ On choisit la variable qui présente le score le plus élevé

→ Et on l'intègre dans le modèle si elle est significative au risque α que l'on a choisi

(1) Il s'agit en réalité d'un test d'hypothèses où $H_0 : a_{p+1} = 0$ vs. $H_1 : a_{p+1} \neq 0$

(2) On peut utiliser le même principe pour tester l'ajout simultané de plusieurs variables (DDL = nombre de variables supplémentaires)

(3) Le test du Score est donc une alternative au test du rapport de vraisemblance pour évaluer la significativité d'une ou d'un groupe de variables

Test du Score – Un exemple

Évaluer l'ajout de « âge » dans un modèle où il y a déjà la constante et taux max

EMV : CONST et TAUX_MAX

$$C(X) = -0.0627 \times \text{TAUX_MAX} + 8.4784$$

$$\pi(X) = P(Y = + / X) = \frac{e^{C(X)}}{1 + e^{C(X)}}$$

Coef								
const.	taux_max	age	coeur	cœur	C(X)	π	LL	
1	126	50	presence	1	0.584	0.642	-0.443	
1	126	49	presence	1	0.584	0.642	-0.443	
1	144	46	presence	1	-0.544	0.367	-1.002	
1	139	49	presence	1	-0.231	0.443	-0.815	
1	154	62	presence	1	-1.170	0.237	-1.441	
1	156	35	presence	1	-1.296	0.215	-1.538	
1	160	67	absence	0	-1.546	0.176	-0.193	
1	140	65	absence	0	-0.293	0.427	-0.557	
1	143	47	absence	0	-0.481	0.382	-0.481	
1	165	58	absence	0	-1.860	0.135	-0.145	
1	115	57	absence	0	1.273	0.781	-1.520	
1	145	59	absence	0	-0.607	0.353	-0.435	
1	175	44	absence	0	-2.486	0.077	-0.080	
1	153	41	absence	0	-1.108	0.248	-0.285	
1	152	54	absence	0	-1.045	0.260	-0.301	
1	169	52	absence	0	-2.110	0.108	-0.114	
1	168	57	absence	0	-2.048	0.114	-0.121	
1	158	50	absence	0	-1.421	0.194	-0.216	
1	170	44	absence	0	-2.173	0.102	-0.108	
1	171	49	absence	0	-2.236	0.097	-0.102	
						-2LL	20.682	

A été optimisé uniquement sur CONST et TAUX_MAX

Vecteur U

U(const) = 0 et

U(Taux_Max) = 0 puisque

nous avons trouvé un

optimum de -2LL à partir de ces 2 variables.

0.0000 0.0000 -22.6863

U(const) U(taux_max) U(age)

0.36	45.11	17.90
0.36	45.11	17.54
0.63	91.11	29.11
0.56	77.48	27.31
0.76	117.54	47.32
0.79	122.48	27.48
-0.18	-28.10	-11.77
-0.43	-59.81	-27.77
-0.38	-54.62	-17.95
-0.13	-22.23	-7.81
-0.78	-89.85	-44.53
-0.35	-51.16	-20.82
-0.08	-13.45	-3.38
-0.25	-37.99	-10.18
-0.26	-39.54	-14.05
-0.11	-18.27	-5.62
-0.11	-19.20	-6.51
-0.19	-30.73	-9.72
-0.10	-17.38	-4.50
0.10	16.52	11.73

Ce n'est pas le cas pour

U(age), il ne prenait pas part aux calculs

H = X'VX

3.41	501.80	177.41
501.80	74552.70	26056.59
177.41	26056.59	9440.07


SIGMA = H⁻¹ (-1)

43.46	-0.20	-0.27
-0.20	0.00	0.00
-0.27	0.00	0.00

Statistique du score pour age

Valeur	2.3766
d.d.l	1
p-value	0.1232

V est la matrice diagonale formée par les π x (1 - π)

X est la matrice des var. explicatives qui intègre maintenant la variable « âge » 

CCL : Au risque 5%, la variable « âge » n'est pas significative.

Diagnostic de la régression

Analyse des résidus

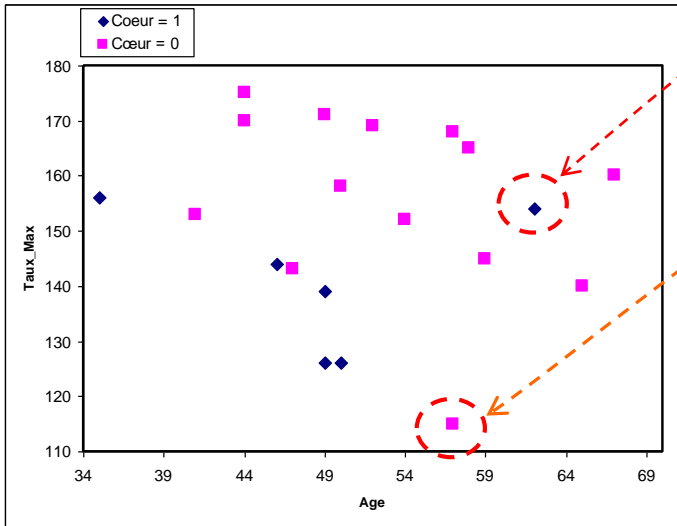
Déterminer s'il y a des observations très mal expliquées

Déterminer si certaines observations s'écartent fortement des autres

Déterminer si certains individus pèsent exagérément sur les résultats (points influents)

Exemple de référence

Cœur = f (age ; taux_max)



	presence	absence	Sum
presence	5	1	6
absence	1	13	14
Sum	6	14	20

Attribute	Coef.	Std-dev	Wald	Signif
constant	16.25444	-	-	-
age	-0.12011	0.0843	2.0303	0.1542
taux_max	-0.074383	0.0387	3.6884	0.0548

Numéro	const	age	taux_max
1	1	50	126
2	1	49	126
3	1	46	144
4	1	49	139
5	1	62	154
6	1	35	156
7	1	67	160
8	1	65	140
9	1	47	143
10	1	58	165
11	1	57	115
12	1	59	145
13	1	44	175
14	1	41	153
15	1	54	152
16	1	52	169
17	1	57	168
18	1	50	158
19	1	44	170
20	1	49	171
Moyenne glob.		52	151
Moyenne cœur = 1		49	141
Moyenne cœur = 0		53	156

QUESTIONS

1. Quels sont les points mal modélisés ? **Résidus**
2. Quels sont les points qui « clochent » ? **Atypiques**
3. Quels sont les points qui pèsent fortement sur le résultat de la modélisation ? **Leviers**
4. Quels sont les points qui, si on les enlevait, nous ferait aboutir à un modèle totalement différent ? **Influents**

Résidus de Pearson

La modélisation de $Y \in \{1 ; 0\}$ peut s'écrire

$$Y(\omega) = \pi(\omega) + \varepsilon(\omega)$$

$$\text{où } \begin{cases} \varepsilon(\omega) = 1 - \pi(\omega), \text{ avec la probabilité } \pi(\omega) \\ \varepsilon(\omega) = -\pi(\omega), \text{ avec la probabilité } 1 - \pi(\omega) \end{cases}$$



$$\begin{cases} E(\varepsilon) = \pi \times [1 - \pi] + (1 - \pi) \times [-\pi] = 0 \\ V(\varepsilon) = \pi \times (1 - \pi) \end{cases}$$

Résidus de Pearson :
(de chaque obs.)

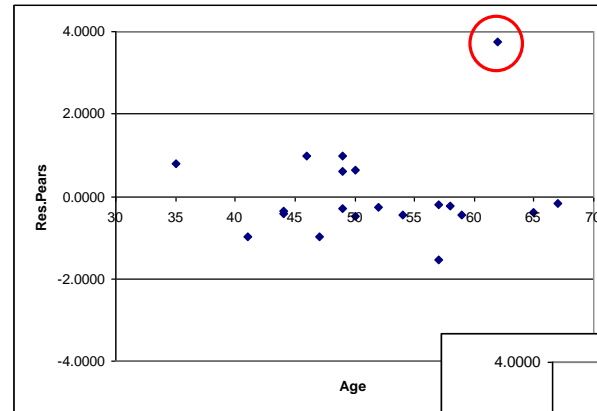
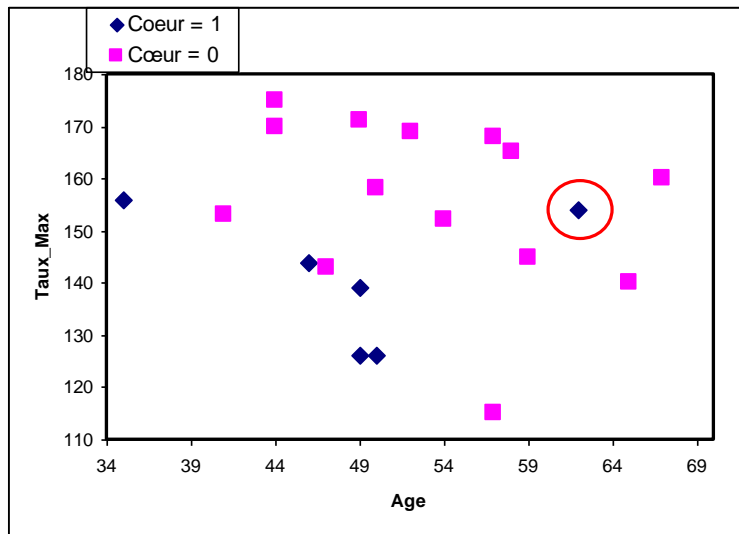
$$r_P = \frac{y - \hat{\pi}}{\sqrt{\hat{\pi}(1 - \hat{\pi})}}$$

De manière à ce que $V(r_P) \neq 1$
Suit approximativement une loi normale $N(0 ; 1)$: valeurs critiques à 5%
+/- 2 (très approximativement)

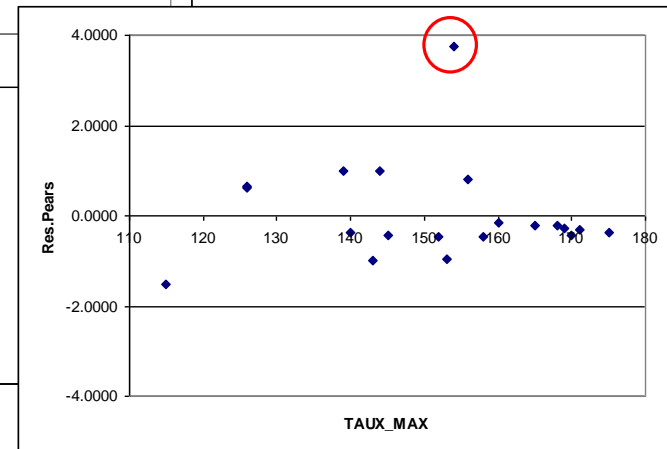
Statistique KHI-2 de Pearson :

$$\chi^2 = \sum_{\omega} r_P^2(\omega)$$

Plus petite elle est, meilleure est l'ajustement
On verra sa distribution plus loin (cf. les covariate pattern)



L'écart est exagéré par la probabilité prédite $\pi(5)=0.06$



Résidus Déviance

La déviance = -2 LL

$$D = -2LL = -2 \times \sum_{\omega} y \times \ln \hat{\pi} + (1 - y) \times \ln(1 - \hat{\pi})$$

Résidus Déviance :
(de chaque obs.)

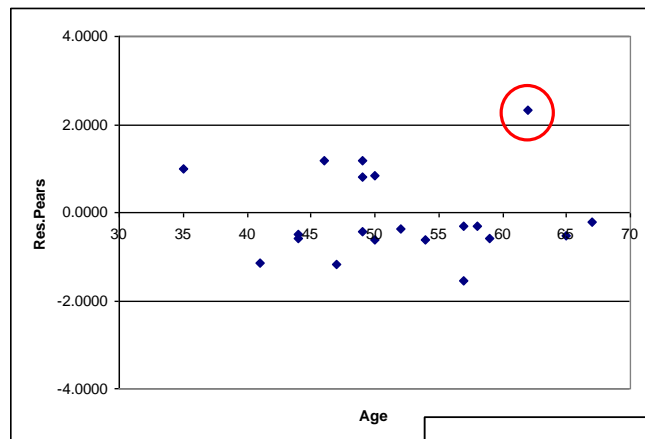
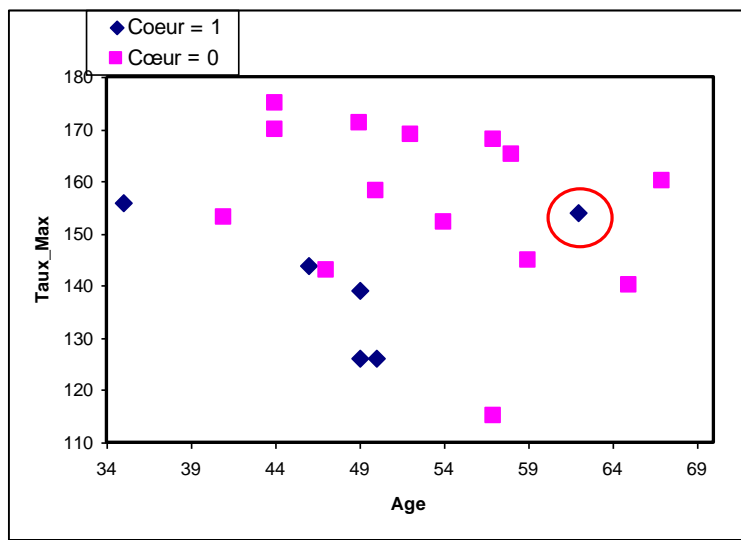
$$r_D = \begin{cases} +\sqrt{2 \times |\ln(\hat{\pi})|}, & \text{si } y = 1 \\ -\sqrt{2 \times |\ln(1 - \hat{\pi})|}, & \text{si } y = 0 \end{cases}$$

Suit approximativement une loi normale N(0 ; 1) : valeurs critiques à 5% # +/- 2 (meilleure approximation que le résidu de Pearson, à privilégier)

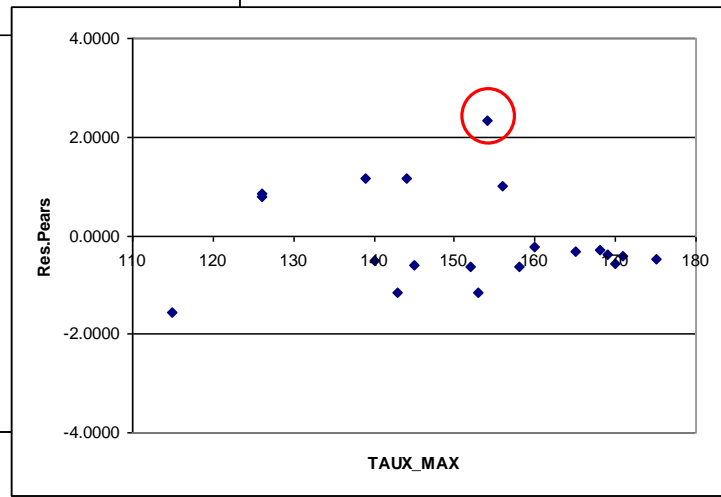
Déviance :

$$D = \sum_{\omega} r_D^2(\omega)$$

Plus petite elle est, meilleure est l'ajustement
On verra sa distribution plus loin (cf. les covariate pattern)



L'écart est moins exagéré, par rapport au résidu de Pearson, mais reste significatif



Levier (1) – Écartement par rapport aux autres points

La HAT-MATRIX s'écrit
(H est symétrique)

$$H = V^{-\frac{1}{2}} X (X' V X)^{-1} X' V^{\frac{1}{2}}$$

V est la matrice diagonale des $\pi(1-\pi)$
X (n,J+1) inclut la constante sur la 1ère colonne

Pour une observation
hatvalues() dans R

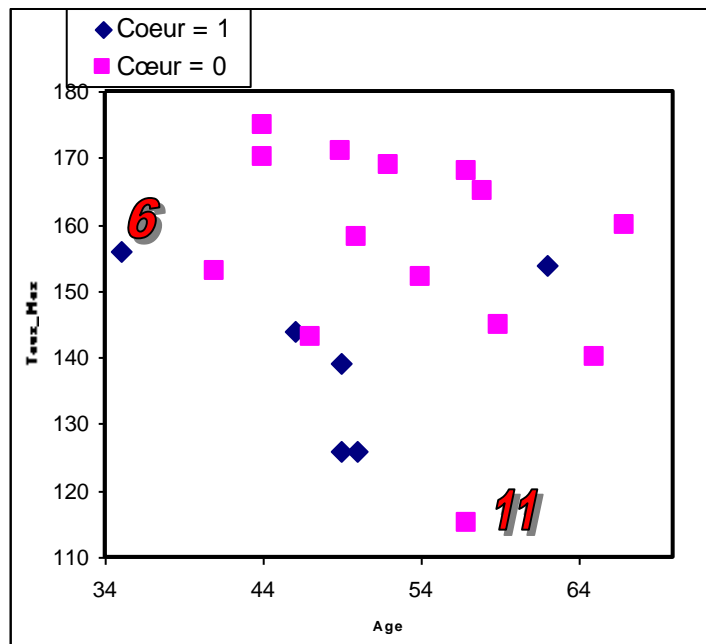
$$h = \hat{\pi}(1-\hat{\pi})x(X' V X)^{-1} x'$$

Lue sur la diagonale principale de H
La distance d'un point par rapport au barycentre
Pondérée par la quantité $\pi(1-\pi)$

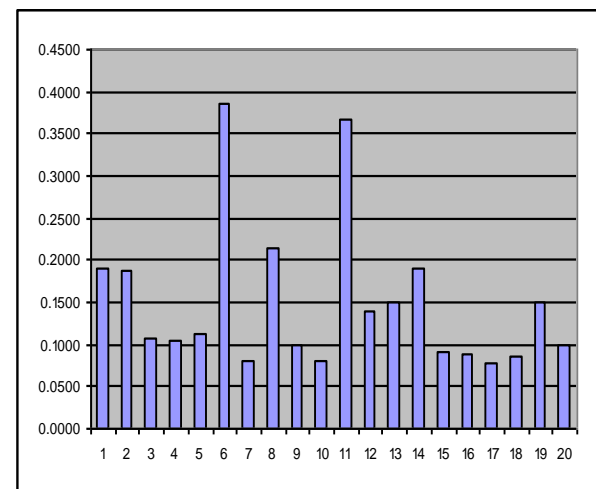
On peut vérifier que

$$\bar{h} = \frac{1}{n} \sum_{\omega} h(\omega) = \frac{J+1}{n}$$

Une règle de détection simple serait $h \geq 2 \times \bar{h}$
Détecter les points qui « décrochent », c'est mieux



					Seuil 0.3	
Numéro	age	taux_max	coeur	PI	h	
1	50	126	1	0.706	0.1914	
2	49	126	1	0.730	0.1885	
3	46	144	1	0.505	0.1073	
4	49	139	1	0.507	0.1048	
5	62	154	1	0.066	0.1121	
6	35	156	1	0.610	0.3848	
7	67	160	0	0.024	0.0804	
8	65	140	0	0.123	0.2149	
9	47	143	0	0.493	0.1000	
10	58	165	0	0.048	0.0809	
11	57	115	0	0.701	0.3663	
12	59	145	0	0.166	0.1401	
13	44	175	0	0.114	0.1489	
14	41	153	0	0.487	0.1908	
15	54	152	0	0.177	0.0899	
16	52	169	0	0.072	0.0871	
17	57	168	0	0.044	0.0768	
18	50	158	0	0.182	0.0855	
19	44	170	0	0.158	0.1498	
20	49	171	0	0.087	0.0997	



Attention :

→ h est surestimé lorsque $\pi \approx 0.5$

→ h est sous-estimé lorsque $\pi \approx 1$ ou $\pi \approx 0$ (ex. $\omega = 7$)

Levier (2) – Mesure de l’influence globale d’une observation

Une autre lecture du levier
(Régression linéaire multiple)

$$\hat{Y}_j = \sum_i h_{i,j} \times Y_i$$

La prédiction du point n°j dépend des valeurs lues dans sa colonne dans la matrice H c.-à-d. $h_{i,j}$ indique l’influence du point i dans la prédiction du point j

Or, on peut montrer que
(s’applique à la Rég.Log.)

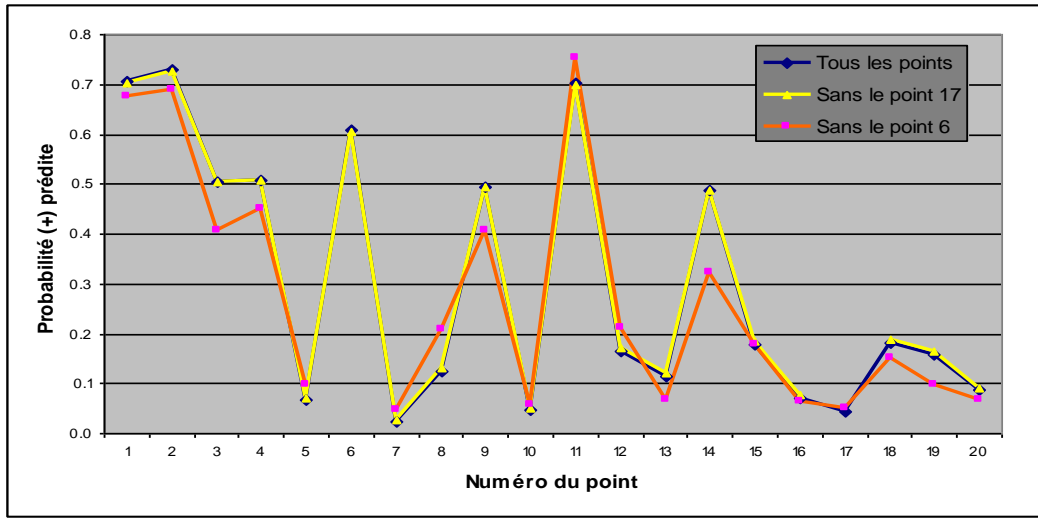
$$h_i = h_{i,i} = \sum_j h_{i,j}^2$$

La valeur sur la diagonale de la matrice H indique l’influence (globale) du point n°i dans la prédiction des valeurs de chaque point n°j

$h_i = J+1$, le point n°i détermine totalement la prédiction de tous les autres points
 $h_i = 0$, le point n°i n’a aucune influence sur la prédiction des autres points

Prédiction de π , avec une rég.log. intégrant tous les points, sans le point n°17 et sans le point n°6

Numéro	PI	h	PI(-17)	PI(-6)
1	0.706	0.1914	0.702	0.675
2	0.730	0.1885	0.726	0.689
3	0.505	0.1073	0.505	0.405
4	0.507	0.1048	0.508	0.450
5	0.066	0.1121	0.071	0.099
6	0.610	0.3848	0.606	
7	0.024	0.0804	0.027	0.048
8	0.123	0.2149	0.130	0.208
9	0.493	0.1000	0.494	0.408
10	0.048	0.0809	0.052	0.058
11	0.701	0.3663	0.698	0.752
12	0.166	0.1401	0.172	0.211
13	0.114	0.1489	0.120	0.068
14	0.487	0.1908	0.487	0.323
15	0.177	0.0899	0.183	0.179
16	0.072	0.0871	0.076	0.063
17	0.044	0.0768		0.050
18	0.182	0.0855	0.188	0.152
19	0.158	0.1498	0.163	0.096
20	0.087	0.0997	0.092	0.066



Le retrait du point n°17 affecte très peu les prédictions
 Le retrait du point n°6 modifie sensiblement les prédictions

Comment quantifier cela ?



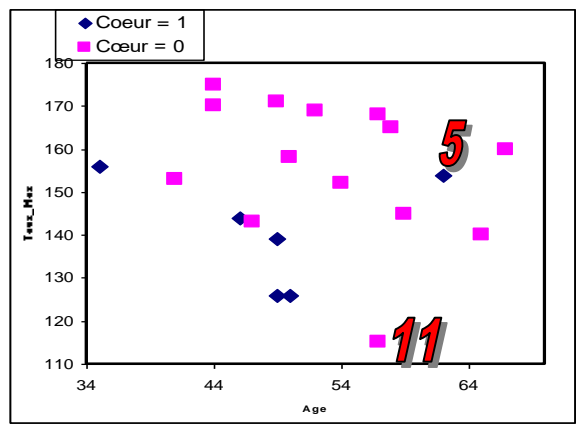
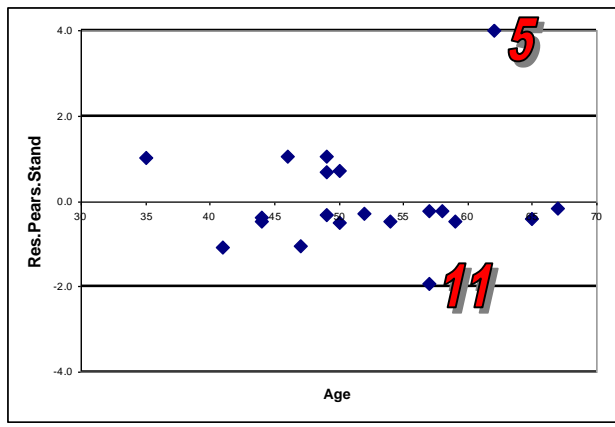
Principe :

Pour chaque observation

- Calculer le résidu, écart Y et Probabilité Prédite
- Sans que le point n'intervienne dans la régression c.-à-d. en donnée supplémentaire
- Sans pour autant avoir à réaliser « n » (taille de la base) régressions
- Les distributions et les seuils (+/- 2 pour un test ~5%) sont les mêmes

Résidu de Pearson Standardisé

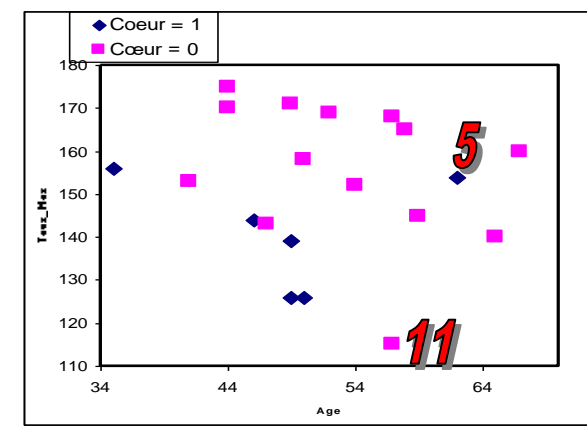
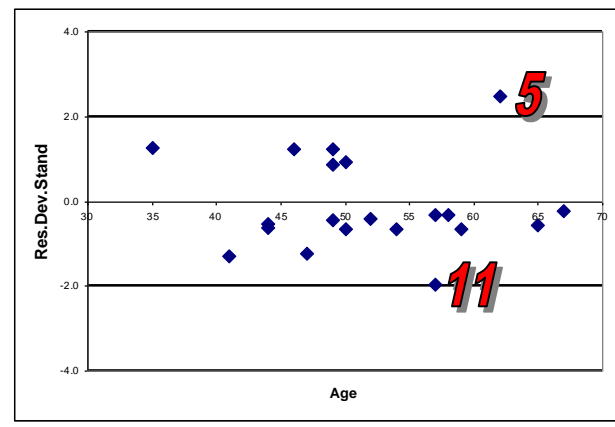
$$r_{P.S} = \frac{r_P}{\sqrt{(1-h)}}$$



Résidu Déviance Standardisé (à privilégier)

$$r_{D.S} = \frac{r_D}{\sqrt{(1-h)}}$$

rstandard() dans R
RSTUDENT dans SPSS



Distance de Cook – Effet global sur tous les coefficients estimés

Principe :
 Comparer les coefficients estimés, avec et sans le point à évaluer
 Test H0 : les coefficients sont tous identiques vs. H1 : 1 au moins est différent
 Sans pour autant avoir à réaliser « n » (taille de la base) régressions

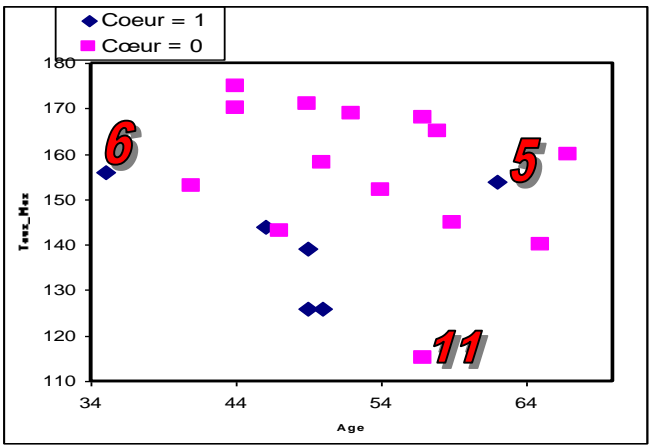
Distance de Cook

Défini sur le résidu déviance standardisé
 cooks.distance() dans R

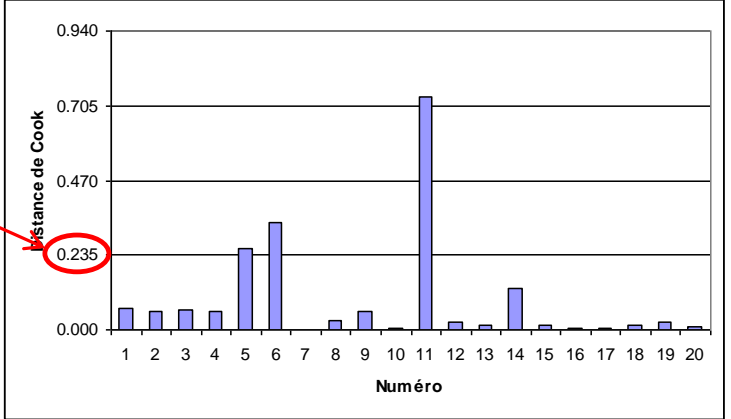
$$D_{D.S} = \frac{r_{D.S}^2}{J+1} \times \frac{h}{(1-h)^2}$$

La règle de détection usuelle est

$$D_{D.S} \geq \frac{4}{n-J-1}$$



$$\frac{4}{20-2-1} \approx 0.235$$



Le point n°6 est correctement modélisé (cf. résidu) mais influe beaucoup sur les résultats (cf. Levier + D.Cook)
 Le point n°11 est mal modélisé (résidu standardisé « limite ») et influe fortement sur les résultats (cf. Levier + D.Cook)
 Le point n°5 est très mal modélisé, mais il pèse moins sur les résultats que les points 6 et 11.

Remarques :

- D peut être défini sur le résidu de Pearson (attention aux petites/grandes valeurs de π)
- Certains ne normalisent pas par le nombre de paramètres (J+1) [ex. SPSS] → la règle de détection est alors $D \geq 1$



Principe : Comparer chaque coefficient estimé, avec et sans le point à évaluer
 Test H0 -- Les coefficients sont identiques dans les deux régressions
Objectif -- Identifier sur quelle variable le point influent pèse le plus
 Cela permet aussi de comprendre de quelle manière le point est atypique

DFBETAS

Formule utilisée dans SAS

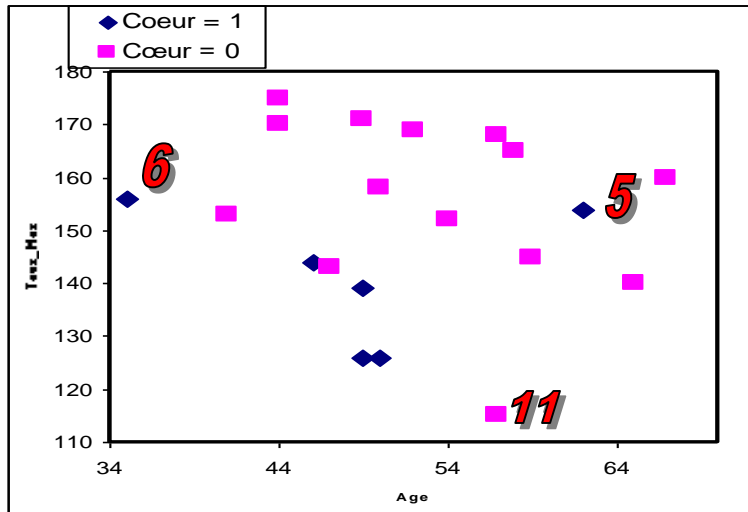
dfbetas() de R semble basée sur le résidu déviance standardisé ?

$$DFBETAS_j = \frac{(X'VX)^{-1} x'_j}{\sqrt{(X'VX)^{-1}_{jj}}} \times \frac{y - \hat{\pi}}{1 - h}$$

La règle de détection usuelle est

$$|DFBETAS_j| \geq \frac{2}{\sqrt{n}}$$

Remarque : Écart type du coefficient $a_j \rightarrow \hat{\sigma}_{\hat{a}_j}$



Numéro	age	taux_max	dfbetas(const)	dfbetas(age)	dfbetas(taux)
1	50	126	0.242	-0.060	-0.276
2	49	126	0.242	-0.083	-0.261
3	46	144	0.172	-0.160	-0.093
4	49	139	0.160	-0.052	-0.155
5	62	154	-1.072	1.241	0.670
6	35	156	0.373	-0.680	0.007
7	67	160	0.040	-0.044	-0.026
8	65	140	0.115	-0.198	-0.027
9	47	143	-0.154	0.118	0.098
10	58	165	0.060	-0.051	-0.050
11	57	115	-0.666	-0.324	1.111
12	59	145	0.102	-0.158	-0.040
13	44	175	0.082	0.012	-0.132
14	41	153	-0.191	0.362	-0.031
15	54	152	0.089	-0.094	-0.068
16	52	169	0.071	-0.041	-0.075
17	57	168	0.056	-0.044	-0.050
18	50	158	0.081	-0.043	-0.095
19	44	170	0.082	0.027	-0.146
20	49	171	0.076	-0.028	-0.092

On comprend mieux :

→ n°6 pèse surtout sur l'âge

→ n°11 pèse surtout sur le taux

→ n°5 est (vraisemblablement) un problème

seuil.bas	-0.447	-0.447	-0.447
seuil.haut	0.447	0.447	0.447

« Covariate pattern » et statistiques associées

Mesures d'évaluation basées sur les résidus

Notion de « covariate pattern »

Quand les données sont constituées de variables nominales/ordinales
Ou quand les données sont issues d'expérimentations

- Plusieurs observations partagent la même description
- Chaque description s'appelle un « covariate pattern »

Ex. hypertension = f (surpoids ; alcool)

- Surpoids = {1 ; 2 ; 3}
- Alcool = {1 ; 2 ; 3}

Il y a $M = 9$ « covariate pattern » possibles
(9 observations « distinctes »)



alcool	surpoids	n	y	pp.obs
1	1	47	16	0.34
1	2	27	11	0.41
1	3	55	39	0.71
2	1	52	20	0.38
2	2	25	15	0.60
2	3	64	41	0.64
3	1	63	38	0.60
3	2	26	15	0.58
3	3	40	33	0.83
Total		399	228	0.57

Effectif dans chaque covariate : n_m

Nombre de positifs dans chaque covariate

y_m

Proportion de positifs dans chaque covariate

Quel intérêt ?

- Développer de nouvelles statistiques d'évaluation basées sur les résidus
- Détecter les groupes (covariates) présentant des caractéristiques (influences) particulières

Estimation et tableau de calcul

Pour l'exemple « Hypertension »

Predicted attribute	hypertension	
Positive value	high	
Number of examples	399	
Model Fit Statistics		
Criterion	Intercept	Model
AIC	546.961	519.871
SC	550.95	531.838
-2LL	544.961	513.871
Model Chi ² test (LR)		
Chi-2	31.0897	
d.f.	2	
P(> Chi-2)	0	
R ² -like		
McFadden's R ²	0.057	
Cox and Snell's R ²	0.075	
Nagelkerke's R ²	0.1006	

Les éléments de diagnostic qui viennent seront l'expression de la confrontation de ces deux quantités

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.673659	-	-	-
surpoids	0.583889	0.1204	23.5307	0.0000
alcool	0.410675	0.1332	9.5009	0.0021

Proportion des positifs observés pour chaque covariate pattern

Proportion des positifs prédits (par le modèle) pour chaque covariate pattern

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
surpoids	1.793	1.4162	2.2701
alcool	1.5078	1.1613	1.9578

Les deux facteurs induisent une augmentation du risque d'hypertension.

alcool	surpoids	n(m)	y(m)	pp.obs(m)	c(m)	pp.pred(m)
1	1	47	16	0.340	-0.7	0.336
1	2	27	11	0.407	-0.1	0.476
1	3	55	39	0.709	0.5	0.620
2	1	52	20	0.385	-0.3	0.433
2	2	25	15	0.600	0.3	0.578
2	3	64	41	0.641	0.9	0.711
3	1	63	38	0.603	0.1	0.536
3	2	26	15	0.577	0.7	0.674
3	3	40	33	0.825	1.3	0.788

Tableau de calcul : confrontation « observation » et « prédiction »

Levier de chaque covariate pattern

Rappel « Levier »

- Indique l'écartement d'un covariate par rapport aux autres
- Indique l'influence d'un covariate dans la prédiction des probas des autres covariates
(Levier = J+1, il détermine toutes les prédictions ; levier = 0, aucune influence)

Pour le covariate « m »
(on a M covariates distincts)

$$h_m = n_m \hat{\pi}_m (1 - \hat{\pi}_m) x_m (X'VX)^{-1} x_m'$$

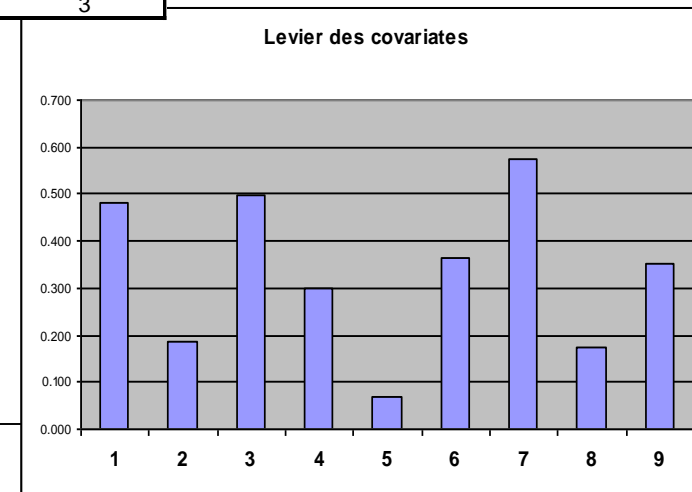
Attention : $(X'VX)^{-1}$ est toujours la matrice de variance covariance des coefficients calculée sur les n observations.

Tableau des leviers, exemple « Hypertension »

N° covariate	alcool	surpoids	n(m)	pp.obs(m)	pp.pred(m) = pi	pi x (1 - pi)	h(m)
1	1	1	47	0.340	0.336	0.223	0.481
2	1	2	27	0.407	0.476	0.249	0.186
3	1	3	55	0.709	0.620	0.236	0.499
4	2	1	52	0.385	0.433	0.246	0.301
5	2	2	25	0.600	0.578	0.244	0.068
6	2	3	64	0.641	0.711	0.206	0.365
7	3	1	63	0.603	0.536	0.249	0.576
8	3	2	26	0.577	0.674	0.220	0.172
9	3	3	40	0.825	0.788	0.167	0.351
Somme							3

Le covariate n°7 détermine pour beaucoup les résultats. Dans une moindre mesure les covariates n°3 et n°1.

→ Cela ne veut pas dire pour autant que ces covariates soient problématiques, l'analyse des résidus doit nous le dire.



Résidus de Pearson

Et covariates pattern

Résidu de Pearson
(pour les covariates)

$$\hat{y}_m = n_m \hat{\pi}_m$$

$$r_m = \frac{y_m - \hat{y}_m}{\sqrt{n_m \hat{\pi}_m (1 - \hat{\pi}_m)}}$$

D'autant plus grand que :

- (1) La prédiction est mauvaise
- (2) Les effectifs sont faibles
- (3) La probabilité est proche de 0 ou 1

Statistique de Pearson
(CHI-SQUARE)

$$\chi^2 = \sum_{m=1}^M r_m^2$$

Si n_m est assez grand, qq soit m , alors la statistique suit une loi du KHI-2 à $(M-J-1)$ degrés de liberté.

Attention : si prédictives continues ($M \# n$) l'approximation par la loi du KHI-2 n'est pas valable !!!

Contribution à la statistique de Pearson (carré du résidu standardisé)

$$\Delta \chi_m^2 = \frac{r_m^2}{1 - h_m}$$

Indique la variation du KHI-2 de Pearson si on supprime la covariate « m » de la régression

N° cov	alcool	surpoids	n(m)	y(m)	pp.obs(m)	p.pred(m)	y^ (m)	pi x (1 - pi)	h(m)	r.pears	chi.pears	delta(chi.pears)
1	1	1	47	16	0.340	0.336	15.8	0.223	0.481	0.057	0.003	0.006
2	1	2	27	11	0.407	0.476	12.9	0.249	0.186	-0.716	0.513	0.630
3	1	3	55	39	0.709	0.620	34.1	0.236	0.499	1.364	1.861	3.716
4	2	1	52	20	0.385	0.433	22.5	0.246	0.301	-0.708	0.502	0.718
5	2	2	25	15	0.600	0.578	14.5	0.244	0.068	0.221	0.049	0.052
6	2	3	64	41	0.641	0.711	45.5	0.206	0.365	-1.239	1.534	2.416
7	3	1	63	38	0.603	0.536	33.7	0.249	0.576	1.077	1.160	2.735
8	3	2	26	15	0.577	0.674	17.5	0.220	0.172	-1.056	1.114	1.346
9	3	3	40	33	0.825	0.788	31.5	0.167	0.351	0.580	0.336	0.518
Somme											7.0711	
d.f											6	
p-value											0.3143	

Mal modélisés, encore que $|r.pears|$ nettement inférieur à 2, si on les enlevait, l'ajustement serait meilleur encore (diminution du KHI-2 de 3.716)

Une valeur de comparaison possible serait $\chi^2(1)$ à 5% # 3.84

Le KHI-2 (7.0711) n'est pas significatif (p-value = 0.3143) c.-à-d. les probas prédites ne sont pas significativement différentes des probas observées → le modèle prédit bien globalement.

Résidus déviance

Et covariate pattern

Résidu Déviance
(pour les covariates)

$$d_m = \text{signe}(y_m - \hat{y}_m) \times \sqrt{2 \left[y_m \ln \frac{y_m}{\hat{y}_m} + (n_m - y_m) \ln \frac{n_m - y_m}{n_m - \hat{y}_m} \right]}$$

Déviance

$$D = \sum_{m=1}^M d_m^2$$

Indique l'écart entre les probas observées et prédites

Si n_m est assez grand, qq soit m , alors la statistique suit une loi du KHI-2 à $(M-J-1)$ degrés de liberté.

Danger : si prédictives continues ($M \# n$) l'approximation par la loi du KHI-2 est illicite !!!

Contribution à la
Déviance

$$\Delta D_m = d_m^2 + r_m^2 \frac{h_m}{1 - h_m}$$

Indique la variation de la Déviance (D) si on supprime la covariate « m » de la régression

Résidus déviance

Calcul déviance

Delta déviance si retrait
du covariate

N° covariate	alcool	surpoids	n(m)	y(m)	y^ (m)	h(m)	r.pears	r.dev	dev.	delta(dev)
1	1	1	47	16	15.8	0.481	0.057	0.057	0.003	0.003
2	1	2	27	11	12.9	0.186	-0.716	-0.719	0.516	0.384
3	1	3	55	39	34.1	0.499	1.364	1.390	1.932	5.587
4	2	1	52	20	22.5	0.301	-0.708	-0.712	0.507	0.473
5	2	2	25	15	14.5	0.068	0.221	0.221	0.049	0.006
6	2	3	64	41	45.5	0.365	-1.239	-1.213	1.471	3.044
7	3	1	63	38	33.7	0.576	1.077	1.082	1.171	2.947
8	3	2	26	15	17.5	0.172	-1.056	-1.033	1.068	1.372
9	3	3	40	33	31.5	0.351	0.580	0.593	0.352	0.306

Somme
d.f. 6
p-value 0.3145

Il faut vraiment se pencher sur ce covariate : si on l'enlevait, l'ajustement serait nettement meilleur (diminution de la déviance de 5.587)

Une valeur de comparaison possible serait $\chi^2(1)$ à 5% # 3.84

Le KHI-2 (7.0690) n'est pas significatif (p -value = 0.3145) c.-à-d. les probas prédites ne sont pas significativement différentes des probas observées → le modèle prédit bien globalement.

C et CBAR

Indiquer l'effet de la suppression d'une variable sur les coefficients

Principe : Indiquer l'effet de la suppression d'un covariate sur les coefficients (écart entre les vecteurs de coefficients)

- Véhicule la même idée que la distance de Cook (c'est une alternative possible)
- Disponible dans SAS (Distance de Cook est disponible dans R)

La statistique CBAR

$$\bar{C}_m = r_m^2 \frac{h_m}{(1-h_m)}$$

La statistique C

$$C_m = r_m^2 \frac{h_m}{(1-h_m)^2}$$

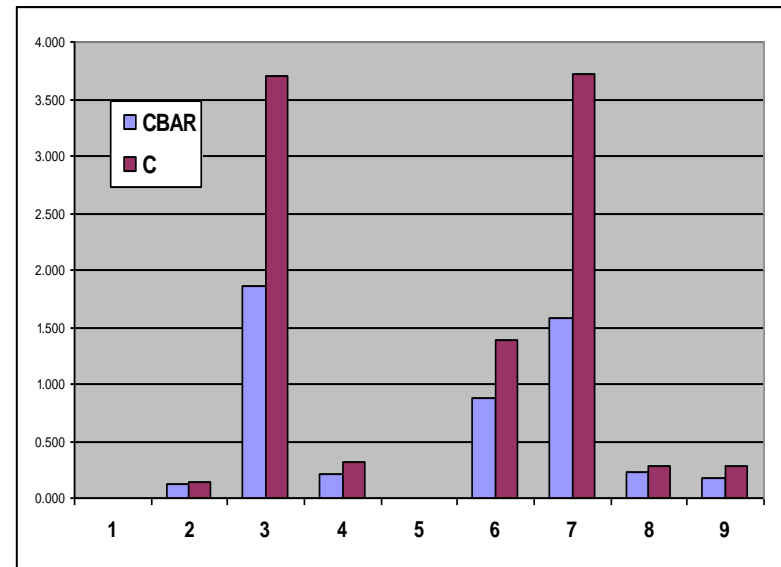
$h_m < 1 \rightarrow$ L'indicateur C rend plus fort encore l'effet d'un levier (h_m) élevé par rapport à CBAR

Tableau de calcul : C et CBAR

N° covariate	alcool	surpoids	h(m)	r.pears	CBAR	C
1	1	1	0.481	0.057	0.003	0.006
2	1	2	0.186	-0.716	0.117	0.144
3	1	3	0.499	1.364	1.855	3.703
4	2	1	0.301	-0.708	0.216	0.309
5	2	2	0.068	0.221	0.004	0.004
6	2	3	0.365	-1.239	0.882	1.389
7	3	1	0.576	1.077	1.575	3.715
8	3	2	0.172	-1.056	0.232	0.280
9	3	3	0.351	0.580	0.182	0.280

Les covariates n°3 et n°7 ont une influence importante sur les coefficients calculés. Déjà repérés avec le levier. Le covariate n°3 se distingue d'autant plus qu'il est (relativement) mal modélisé (cf. résidus).

\rightarrow On peut affiner l'analyse avec les DFBETAS : effet du covariate sur tel ou tel coefficient...



DFBETAS et DFBETA

Et covariate pattern

DFBETA
(pour les covariates)

$$DFBETA_{j,m} = (X'VX)^{-1} x'_m \times \frac{y_m - \hat{y}_m}{1 - h_m}$$

Écart absolu du coefficient estimé avec ou sans le covariate

DFBETAS
(pour les covariates)

$$DFBETAS_{j,m} = \frac{(X'VX)^{-1} x'_m}{\sqrt{(X'VX)^{-1}_{jj}}} \times \frac{y_m - \hat{y}_m}{1 - h_m}$$

Écart normalisé par l'écart type du coefficient estimé

Remarque : Lorsque les données sont (1) sur la même échelle (ex. même unités) ou (2) directement des échelles de valeurs (cf. notre exemple) ou (3) des indicatrices, on a intérêt à utiliser directement le DFBETA → on manipule souvent les covariates pattern dans les deux dernières configurations

Tableau de calcul : DFBETA

N°	alcool	surpoids	y(m)	y^(m)	h(m)	dfbeta.const	dfbeta.alcool	dfbeta.surpoids
1	1	1	16	15.8	0.481	0.029	-0.007	-0.006
2	1	2	11	12.9	0.186	-0.111	0.039	0.004
3	1	3	39	34.1	0.499	0.145	-0.140	0.124
4	2	1	20	22.5	0.301	-0.150	0.008	0.049
5	2	2	15	14.5	0.068	0.005	0.000	0.001
6	2	3	41	45.5	0.365	0.183	-0.025	-0.110
7	3	1	38	33.7	0.576	0.009	0.156	-0.106
8	3	2	15	17.5	0.172	-0.100	-0.056	-0.012
9	3	3	33	31.5	0.351	-0.154	0.049	0.043

Suppression des covariates n°3 et n°7 entraîne une forte modification des coefficients avec, de plus, des effets inversés sur les deux variables.

Comment lire ces valeurs ? Si on supprime le covariate n°7, le coefficient de « surpoids » va être augmenté de 0.106, et celui de « alcool » diminué de 0.156 c.-à-d.

Estimation de a_j sans le covariate « m »

$$\hat{a}_{j,(-m)} = \hat{a}_j - DFBETA_{j,m}$$

DFBETA

L'exemple du covariate n°7

Régression sur tous les covariates

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.673659	-	-	-
surpoids	0.583889	0.1204	23.5307	0.0000
alcool	0.410675	0.1332	9.5009	0.0021

DFBETA pour le covariate n°7

N°	alcool	surpoids	y(m)	y^ (m)	h(m)	dfbeta.const	dfbeta.alcool	dfbeta.surpoids
7	3	1	38	33.7	0.576	0.009	0.156	-0.106

Coefficients sans le covariate n°7

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.689314	-	-	-
surpoids	0.692091	0.1379	25.1921	0.0000
alcool	0.257156	0.1613	2.5415	0.1109

$$0.692 \approx 0.584 - (-0.106)$$

L'écart est imputable aux erreurs de troncature lors de la récupération manuelle des coefs.

Sans le covariate n° 7 (les alcools maigrichons), l'effet de l'alcool sur l'hypertension ne serait plus significative (à 5%). On constate le rôle très important que joue ce covariate !!!

Quelques éléments supplémentaires

Mieux comprendre, anticiper et améliorer les performances de la
Régression Logistique

Non-linéarité sur le LOGIT

$$\text{logit}(Y) = a_0 + a_1 X_1 + \dots + a_J X_J$$

Linéarité sur le LOGIT → une augmentation d'une unité de X1 entraîne une augmentation de « a1 » du LOGIT, quelle que soit la valeur de X1.

Non-linéarité sur le LOGIT → la variation du LOGIT suite à une variation de X1 dépend du niveau (ou de la valeur) de X1.

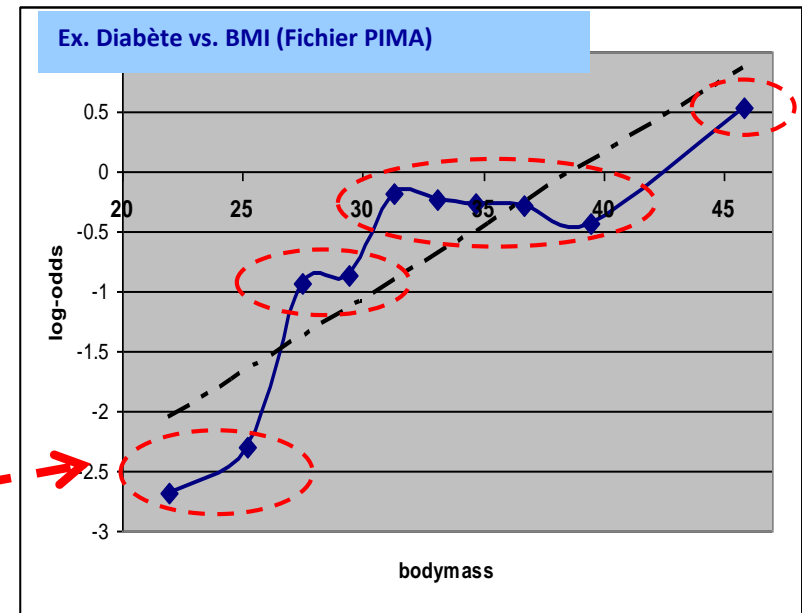
Identification graphique univariée

Le graphique LOGIT vs. Variable indépendante (continue)

Graphique LOGIT

- Découper X en déciles (ou autres)
- Dans chaque intervalle, calculer la proportion de positifs π
- Construire le graphique :
- → en abscisse, centre des intervalles
- → en ordonnée, LOGIT observé c.-à-d. $\text{LN} [\pi / (1 - \pi)]$

On note une évolution par paliers, absolument pas linéaire.
Elle est néanmoins monotone.



Solution : Transformation de variable.

La plus simple : discrétisation de la variable continue (regroupement en classes)

Codage « cumulatif » contraint (pour var. ordinale)
Pour tenir compte de l'information sur la monotonie

$$\begin{aligned} D_1 &= 1, \text{ si BMI} > 26.20; 0 \text{ sinon} \\ D_2 &= 1, \text{ si BMI} > 30.34; 0 \text{ sinon} \\ D_3 &= 1, \text{ si BMI} > 41.62; 0 \text{ sinon} \end{aligned} \quad \rightarrow \quad \begin{aligned} D_2 &= 1 \Rightarrow D_1 = 1 \\ D_3 &= 1 \Rightarrow D_2 = 1 \Rightarrow D_1 = 1 \end{aligned}$$

Rég.Log. Variable initiale

Odds ratios and 95% confidence intervals			
Attribute	Coef.	Low	High
BODYMASS	1.1079	1.0809	1.1357

Rég.Log. Variables discrétisées

Odds ratios and 95% confidence intervals			
Attribute	Coef.	Low	High
D1	4.8802	2.4534	9.7071
D2	1.8513	1.2297	2.787
D3	2.2612	1.3597	3.7602



Passage aux variables discrétisées **diminue le biais** (bien) mais **augmente la variance** (pas bien)

Si évolution non linéaire + non monotone, adopter un codage disjonctif non contraint (0/1)

→ Le mieux est quand même de ne pas discrétiser mais d'utiliser une fonction de transformation

Test de Box-Tidwell - Identification numérique multivariée

Nombre de variables élevé → Détection graphique (individuelle) impossible

S'appuyer sur une procédure statistique pour identifier, et par la suite étudier en détail, les variables susceptibles d'intervenir de manière non-linéaire.

Principe : Pour une variable X à tester, ajouter dans la régression, en plus des autres variables, et en plus de X, le terme d'interaction $Z = X * \ln(X)$. Si le coefficient de Z est significatif, la variable X intervient sous une forme non linéaire (qui reste à identifier).

Remarque : Ce test détecte mal les petits écarts à la linéarité.

Predicted attribute	DIABETE	
Positive value	positive	
Number of examples	757	
Model Fit Statistics		
Criterion	Intercept	Model
AIC	983.528	908.891
SC	988.157	918.15
-2LL	981.528	904.891
Model Chi ² test (LR)		
Chi-2	76.6367	
d.f.	1	
P(> Chi-2)	0	
R ² -like		
McFadden's R ²	0.0781	
Cox and Snell's R ²	0.0963	
Nagelkerke's R ²	0.1325	

Régression BMI seule

Predicted attribute	DIABETE	
Positive value	positive	
Number of examples	757	
Model Fit Statistics		
Criterion	Intercept	Model
AIC	983.528	899.906
SC	988.157	913.794
-2LL	981.528	893.906
Model Chi ² test (LR)		
Chi-2	87.6214	
d.f.	2	
P(> Chi-2)	0	
R ² -like		
McFadden's R ²	0.0893	
Cox and Snell's R ²	0.1093	
Nagelkerke's R ²	0.1504	

Régression BMI ET $Z = \text{BMI} \times \ln(\text{BMI})$

Attribute	Coef.	Std-dev	Wald	Signif
constant	-13.756516	-	-	-
BODYMASS	1.403581	0.3933	12.7355	0.0004
BMI x LN(BMI)	-0.286109	0.0861	11.0508	0.0009

Variable additionnelle significative : BMI intervient de manière non-linéaire dans la relation (C'est cohérent avec l'analyse précédente).

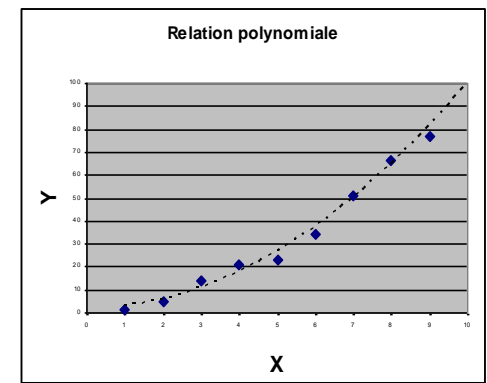
Identification multivariée - Résidus partiels (1)

Identifier la forme appropriée d'une variable dans la régression

Dans la régression linéaire simple : Graphique « nuage de points » pour identifier la forme de la relation entre X et Y

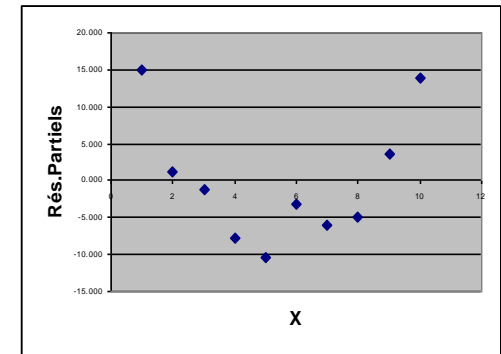


---> Ici la relation semble polynomiale
Rajouter la variable $Z=X^2$ (ou substituer à X) dans la régression



Dans la régression linéaire multiple : le graphique individuel (X_j, Y) n'est plus valable parce qu'il ne tient pas compte des autres variables → On utilise alors les « Résidus Partiels » $\hat{\epsilon}_j = (y - \hat{y}) + \hat{a}_j \times x_j$

$(x_j, \hat{\epsilon}_j)$ Le nuage doit former une droite (si relation linéaire)

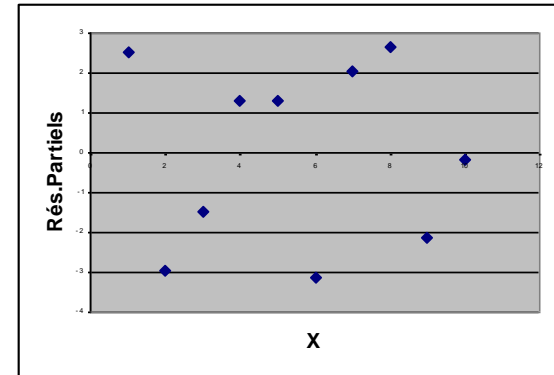


Si le nuage ne forme pas une droite :

Rajouter (ou substituer à) la transformation voulue dans la régression (inspirée de la forme du nuage de points, ex. X^2), recalculer l'équation de régression et construire le graphique des résidus partiels (augmentés)

$$\hat{\epsilon}_j = (y - \hat{y}) + \hat{a}_j \times x_j + \hat{a}_{j+1} \times x_{j+1}^2$$

Tous les coefs. et formes de la variable



Si les bonnes transformations ont été introduites, le nuage doit former une droite.

Identification multivariée - Résidus partiels (2)

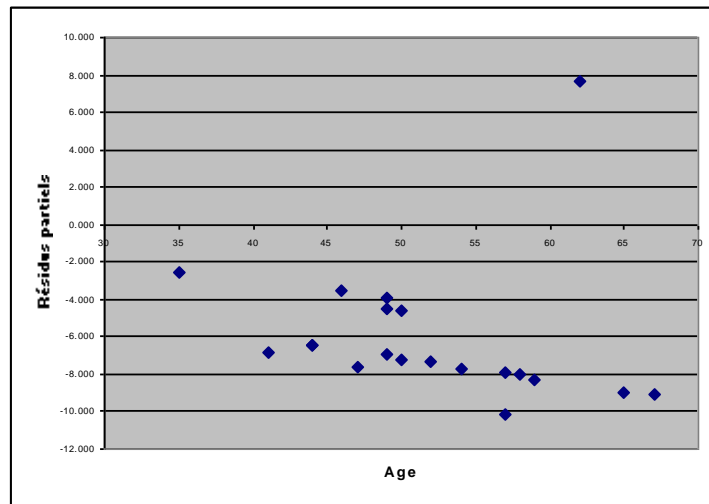
Le cas de la régression logistique

Résidus partiels pour la régression logistique

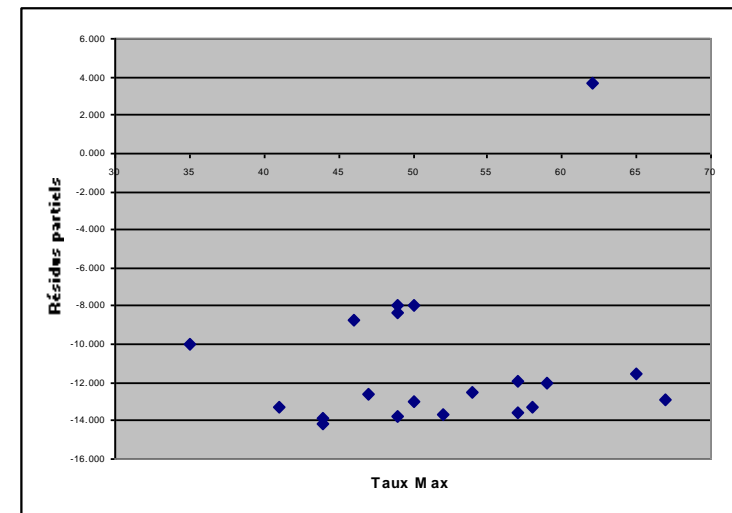
$$r_{x_j} = \frac{y - \hat{\pi}}{\hat{\pi} \times (1 - \hat{\pi})} + \hat{a}_j \times x_j \quad \longrightarrow$$

Élaborer le nuage de points (x_j, r_{x_j}) : si OK, elle forme une droite ; sinon, introduire la variable transformée dans la régression et calculer les résidus partiels augmentés

Heart = f(age, taux_max)



OK pour « âge », mis à part le point atypique



OK pour « taux max », mis à part le point atypique

Non-additivité

Interactions entre variables continues

Améliorer les performances en prédiction

à venir...

Utiliser les MCO pondérés

Pour estimer les paramètres de la régression logistique

Régression pondérée (1)

La méthode des moindres carrés généralisés

Reprenons $Y \in \{0;1\}$ et définissons l'équation de régression avec les hypothèses de la MCO

$$Y = Xa + \varepsilon$$

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \dots + \varepsilon$$

$$E[Y] = \pi$$

$$E[\varepsilon] = 0$$

$$V[\varepsilon] = V[Y]$$

$$= E\{[Y - E(Y)]^2\}$$

$$= E[Y^2] - E(Y)^2$$

$$= \pi - \pi^2 = \pi(1 - \pi)$$

Moyenne de Y \leftrightarrow Probabilité de Y

Par hypothèse

X est non aléatoire (indépendant de ε)

$Y^2 = Y$ puisque $\{0;1\}$

Hétéroscédasticité

Utiliser l'estimateur des moindres carrés généralisés

Z est le LOGIT prédit, corrigé des erreurs de prédiction pour chaque observation (MMV)

$$\hat{z} = [\hat{a}_{0,MMV} + \hat{a}_{0,MMV} x_1 + \dots] + \frac{(y - \pi)}{\pi(1 - \pi)}$$

$$\hat{a}_{MCG} = (X' \hat{V} X)^{-1} (X' \hat{V}) \hat{Z}$$

V (VCV des résidus) est une matrice de taille (n x n) dont la diagonale est formée par les $\pi(1 - \pi)$

En d'autres termes



Pondérer chaque individu par η (différent pour chaque individu)

$$\frac{\hat{z}}{\eta} = a_0 \frac{1}{\eta} + a_1 \frac{x_1}{\eta} + a_2 \frac{x_2}{\eta} + \dots$$

$$\eta(\omega) = \frac{1}{\sqrt{\pi(\omega) \times [1 - \pi(\omega)]}}$$

Conclusion : On sait alors que

$$\hat{a}_{MCG} = \hat{a}_{MMV}$$

Quel intérêt ? (puisque on a besoin des résultats des MMV)

→ Des méthodes itératives plus rapides/stables s'appuient sur ce résultat (IRLS : iteratively re-weighted least square)

→ Utiliser les programmes de régression (moins gourmands en ressources) pour la sélection de variables

Régression pondérée (2)

Calculs

Obtenus par rég. logistique

$$\eta(\omega) = \frac{1}{\sqrt{\pi(\omega) \times [1 - \pi(\omega)]}}$$

π par la MMV	$\pi(1-\pi)$	Poids
0.879	0.106	3.066
0.582	0.243	2.027
0.392	0.238	2.048
0.378	0.235	2.062
0.213	0.168	2.441
0.877	0.108	3.040
0.016	0.016	7.871
0.071	0.066	3.893
0.378	0.235	2.063
0.036	0.035	5.350
0.858	0.122	2.868
0.106	0.095	3.252
0.104	0.093	3.281
0.406	0.241	2.037
0.124	0.109	3.030
0.058	0.055	4.266
0.173	0.143	2.645
0.138	0.119	2.898
0.137	0.118	2.907
0.074	0.068	3.827

η				\hat{z}
const.	age	taux max	angine	cœur
0.33	16.31	41.10	0.33	1.02
0.49	24.17	62.16	0.00	1.01
0.49	22.46	70.31	0.00	1.03
0.48	23.76	67.41	0.00	1.04
0.41	25.40	63.09	0.41	1.39
0.33	11.51	51.32	0.33	1.02
0.13	8.51	20.33	0.00	-0.65
0.26	16.70	35.97	0.00	-0.94
0.48	22.78	69.32	0.00	-1.02
0.19	10.84	30.84	0.00	-0.81
0.35	19.87	40.09	0.35	-1.83
0.31	18.14	44.59	0.00	-1.00
0.30	13.41	53.34	0.00	-1.00
0.49	20.13	75.13	0.00	-1.01
0.33	17.82	50.16	0.00	-1.02
0.23	12.19	39.62	0.00	-0.90
0.38	21.55	63.51	0.38	-1.05
0.35	17.25	54.53	0.00	-1.03
0.34	15.14	58.48	0.00	-1.03
0.26	12.80	44.68	0.00	-0.94

Maximum de Vraisemblance			
angine (a3)	taux_max(a2)	age (a1)	const (a0)
1.779	-0.064	-0.126	14.494
1.504	0.040	0.094	7.955

s
0.138
0.720
1.550
1.644
3.687
0.141
0.017
0.076
0.606
0.038
6.063
0.118
0.116
0.683
0.142
0.062
0.209
0.160
0.159
0.080

Total s **16.407**
s **1.013**

Fonction DROITEREG (EXCEL)

coef.	DROITEREG (MC pondérés)			
	angine (a3)	taux_max(a2)	age (a1)	const (a0)
e.t.	1.779	-0.064	-0.126	14.494
	1.523	0.041	0.095	8.055
	0.16	1.01	# N/A	# N/A
	0.76	16	# N/A	# N/A
	3.12	16.41	# N/A	# N/A

On définit s^2 la variance estimée des résidus
Cf. analogie avec SCR/ddl de la MCO, mais pondérée ici

$$s^2 = \frac{1}{n - J - 1} \sum_{\omega} \eta^2(\omega) \times [y(\omega) - \pi(\omega)]^2$$

Écarts-typés des coefficients mal estimés. Comment corriger ?

e.t. corrigé	angine (a3)	taux_max(a2)	age (a1)	const (a0)
	1.504	0.040	0.094	7.955

$$\hat{\sigma}_{MMV} = \frac{\hat{\sigma}_{MCG}}{s}$$

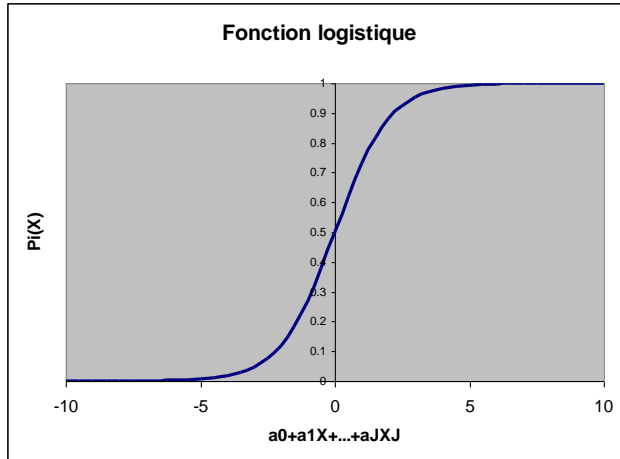
Correction des écarts-typés de la régression par la dispersion des erreurs

Interprétation géométrique

Lien avec l'apprentissage par partitionnement
Régression Logistique = classifieur linéaire

Régression non-linéaire...

Mais séparateur linéaire



Régression Non-Linéaire : Transformer la combinaison linéaire $[C(X) = a_0 + a_1.X_1 + \dots]$ à l'aide d'une fonction de transfert non linéaire (Fonc. Répartition Logistique LOGIT, Fonc. Répart. Normale PROBIT, etc.)

Mais...

Du point de vue du partitionnement, le classifieur induit une séparation linéaire dans l'espace de représentation

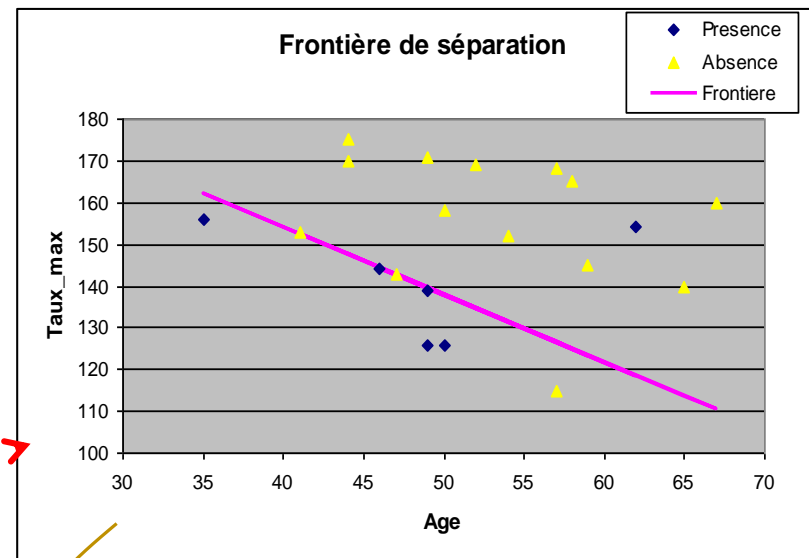
Reprenons le fichier « Maladie cardiaque »
 $Y = f(\text{age}, \text{taux_max})$

MMV $\rightarrow c(X) = 16.254 - 0.120 \times \text{age} - 0.074 \times \text{taux}$

Décision : Si $c(X) > 0$ Alors $Y = +$

\rightarrow c.-à-d. \rightarrow $C(X)=0$ définit une « frontière » entre les $Y=+$ et $Y=-$

D'éq. Explicite $\rightarrow \text{taux} = 218.52 - 1.61 \times \text{age}$



Matrice de confusion			
Nombre de coeur	Pred.		
coeur	absence	presence	Total
absence	13	1	14
presence	1	5	6
Total	14	6	20

Bibliographie

En ligne

Site du cours : http://eric.univ-lyon2.fr/~ricco/cours/cours_regression_logistique.html

Sur le site, l'ouvrage qui vient appuyer ce support

PennState - « STAT 504 - Analysis of discrete data »

<https://onlinecourses.science.psu.edu/stat504/>

Très riche avec des exemples sous SAS et R (redirection sur les sites web)

Ouvrages

D.W. HOSMER, S. LEMESHOW, « Applied Logistic Regression », Wiley, 2003.

S. MENARD, « Applied Logistic Regression Analysis », 2nd Ed., Sage Publications, 1997.

M. BARDOS -- « Analyse discriminante - Application au risque et scoring financier », DUNOD, 2001.

P.L. GONZALES - « Modèles à réponse dichotomique », Chap.6, in *Modèles Statistiques pour Données Qualitatives*, Driesbeke, Lejeune et Saporta, Editeurs, TECHNIP, 2005.