

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 $P(Y/X)$

$$\left[\begin{aligned} P(Y = y_k/X) &= \frac{P(Y = y_k) \times P(X/Y = y_k)}{P(X)} \\ &= \frac{P(Y = y_k) \times P(X/Y = y_k)}{\sum_{l=1}^K P(Y = y_l) \times P(X/Y = y_l)} \end{aligned} \right.$$

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 → 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

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

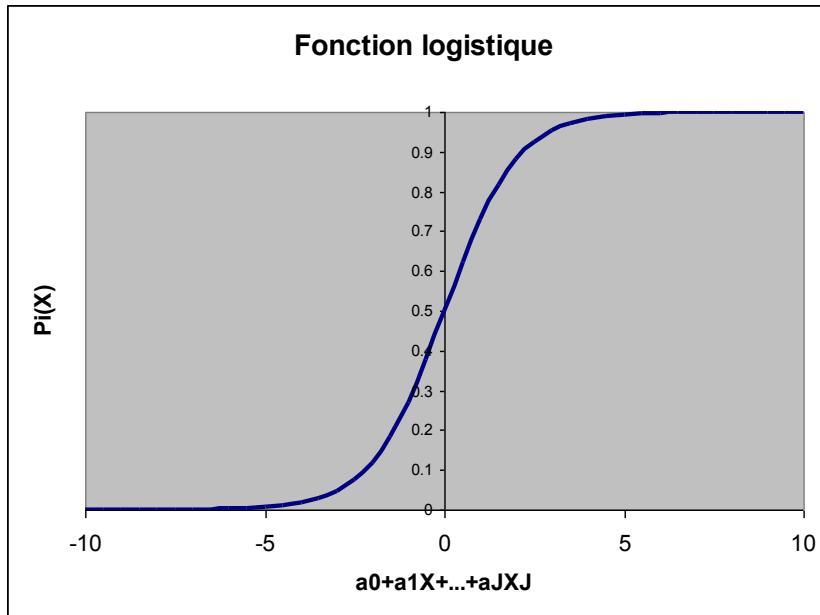
C'est la fonction de répartition de la loi Logistique

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

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

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

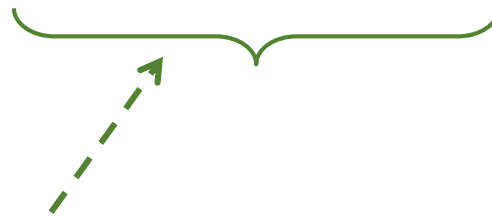
Remarques :

- $C(X)$ 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

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)

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 $a' = (a_0, a_1, \dots, a_J)$
 $X(\omega) = (1, X_1(\omega), \dots, X_J(\omega))$



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	angine	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

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

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

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	

Nombre de	Prédiction		
coeur	presence	absence	Total
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 ?

Lecture et interprétation des coefficients

Ce qui fait le succès de la régression logistique

Risque relatif, odds, odds-ratio

Quelques définitions

Nombre de cœur	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

--- →

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

Régression logistique cœur = f(angine)

Model Chi² test	
Chi-2	2.6924
d.f.	1
P(>Chi-2)	0.1008

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

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

Nombre de cœur	angine		
cœur			
presence	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 $(0.6 - 0.2) = \mathbf{0.4}$

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 $(0.6 - 0.2) = \mathbf{0.4}$

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

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 modele 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

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

FORWARD

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 ☹ ☹ ☹