Apprentissage automatique / Statistique

Régression logistique

PHILIPPE BESSE

INSA de Toulouse
Institut de Mathématiques

Objectif

- Expliquer Z qualitative à 2 modalités $\{0, 1\}$ ou Y nombre de "succès" de Z par $\{X^1, \ldots, X^p\}$ qualitatives et quantitatives
- Prédicteur linéaire $\mathbf{X}\boldsymbol{\beta}$ inadapté
- Cas particulier du MLG : modèle binomial
- Méthode sans doute la plus utilisée (médical, marketing)
- Modèle explicable
- Passe à l'échelle volume

Définition de la cote ou *odd*

- Y une variable qualitative à m modalités
- L'*odd* de la ℓ-ième modalité relativement à la $k$ème est le rapport

$$\Omega_{\ell k} = \frac{\pi_\ell}{\pi_k} \quad \text{avec} \quad \pi_\ell = P[T = \mathcal{T}_\ell] \quad \text{estimé par } \widehat{\Omega}_{\ell k} = \frac{n_\ell}{n_k}$$

Si $m = 2$, $\Omega_{10} = \frac{\pi}{(1-\pi)}$ exprime une cote ou chance de gain

- Si π(succès)=0,8 alors π(échec)=0,2 et *Odd* (succès)=4 : Chance de succès de 4 contre un

Définition du rapport de cotes ou *odds ratio*

- Table de contingence 2×2 croisant T^1 et T^2

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \quad \text{avec} \quad \pi_{ij} = P[\{T^1 = \mathcal{T}_i\} \text{ et } \{T^2 = \mathcal{T}_j\}]$$

$$\Omega_1 = \frac{\pi_{11}}{\pi_{12}} \quad \Omega_2 = \frac{\pi_{21}}{\pi_{22}}$$

- *Odds ratio* ou rapport de cotes : $\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$
- $\Theta = 1$ ssi X^1 et X^2 sont indépendantes
- $\Theta > 1$ si les sujets de la ligne 1 ont plus de chances de prendre la première colonne que les sujets de la ligne 2 et inférieur à 1 sinon

Exemple d'*odds ratio*

- Concours avec 7 garçons reçus sur 10 et
- 4 filles sur 10
- Odd des garçons : 0.7/0.3=2.33
- Odd des filles : 0.4/0.6=0.67
- odds ratio : 2.33/0.67=3.65

Odds ratio dans une table de contingence $J \times K$

$$\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}}$$

estimé par l'*odds ratio* empirique :

$$\widehat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}$$

Notations

- Z variable qualitative à 2 modalités : 1 ou 0...
- $\mathbf{X}\beta$ prend ses valeurs dans $\mathbb{R}$
- Modéliser $\pi = P[Z = 1]$ ou plutôt
- $g(\pi_i) = \mathbf{x}_i'\beta$ avec $g : [0,1] \longmapsto \mathbb{R}$
- g est appelée fonction lien
  - *probit* : g fonction inverse de la fonction de répartition d'une loi normale (pas explicite).
  - *log-log* : $g(\pi) = \ln[-\ln(1-\pi)]$ (dissymétrique)
  - *logit* : $g(\pi) = \textit{logit}(\pi) = \ln \frac{\pi}{1-\pi}$; $g^{-1}(x) = \frac{e^x}{1+e^x}$
- La *régression logistique* est une modélisation linéaire du *log odd*
- Les coefficients expriment des *odds ratio*

Modèle

- $X^1, \ldots, X^q$: explicatives qualitatives ou quantitatives
- I : nombre des combinaisons $x_i^1, \ldots, x_i^q$ des facteurs X^j
- n_i : nombre d'essais avec $x_i^1, \ldots, x_i^q$ fixé $(n = \sum_{i=1}^{I} n_i)$
- y_i nombre de $(Z = 1)$ observés lors des n_i essais,
- Si $\pi_i = P[Z = 1]$ constante pour $x_i^1, \ldots, x_i^q$ fixé Alors
- Y_i sachant n_i suit une loi *binomiale* $\mathcal{B}(n_i, \pi_i)$ d'espérance $E(y_i) = n_i \pi_i$ et de densité : $P(Y = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{(n_i - y_i)}$.
- Hypothèse : $[\mathit{logit}(\pi_i); i = 1, \ldots, n]' \in \text{vect}\{X^1, \ldots, X^q\}$

$$\mathit{logit}(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta} \quad \text{ou} \quad \pi_i = \frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}} \quad i = 1, \ldots, I$$

Estimation

- Estimation $\mathbf{b}$ de β par maximisation de la log-vraisemblance
- Méthodes numériques itératives (Newton Raphson, Scores de Fisher)
- Prévisions des probabilités π_i : $\widehat{\pi}_i = \frac{e^{\mathbf{x}_i'\mathbf{b}}}{1+e^{\mathbf{x}_i'\mathbf{b}}}$
- et des effectifs $\widehat{y}_i = n_i \widehat{p}_i$

Remarques

- $\mathbf{X}$ construite comme pour l'analyse de covariance
- Attention au choix implicite de paramétrisation par le logiciel $(0, 1)$ ou $(-1, 1)$
- Cas précédent : données *groupées*. Si les observations $\mathbf{x}_i$ sont toutes distinctes : $n_i = 1; i = 1, \ldots, I$. Les comportements asymptotiques et tests ne sont plus valides
- En plus des b_j ou *log odds ratio*, estimation possible des *odds-ratio* ou rapports de cote : Y a e^b fois plus de chance d'apparaître quand $X = 1$

Généralisation

- Cas de Y ordinale
- Y qualitative ordinale : niveau de gravité, de satisfaction...
- Problème si plusieurs modèles en concurrence pour chaque fonction logit
- Utilisable si p le nombre de variables explicatives est petit
- Autre choix de Python (`Scikit-learn`) : une classe contre les autres

Régression polytomique

- Une variable explicative X dichotomique de Y à
- k modalités ordonnées.
- $\pi_j(X) = P(Y = j|X)$ avec $\sum_{j=1}^{k} \pi_j(X) = 1$
- Il faut estimer $k-1$ prédicteurs linéaires :

$$g_j(X) = \alpha_j + \beta_j X \quad \text{pour } j = 1, \ldots, k-1$$

- Trois types d'échelle des rapports de cotes :
  - comparaison des catégories adjacentes deux à deux
  - comparaison des catégories adjacentes supérieures cumulées
  - comparaison des catégories adjacentes cumulées

Logits cumulatifs

$$\log \frac{\pi_{j+1} + \cdots + \pi_k}{\pi_1 + \cdots + \pi_j} \quad \text{pour } j = 1, \ldots, k-1$$

- Hypothèse souvent implicite :
- $\beta_j; j = 1, \ldots, k-1$ homogènes
- Même coefficient b : rapport de cotes proportionnels
- ou même fonction logit translatée
- `proc logistic` de SAS propose un test d'homogénéité des β_j
- Interprétation *Pour tout seuil choisi de Y, la cote des risques d'avoir une gravité supérieure à ce seuil est e^b fois plus grande chez les exposés ($X = 1$) que chez les non exposés ($X = 0$)*

Choix de modèle

- Algorithme par élimination ou mixte (stepwise) avec
- Minimisation du critère AIC d'Akaïke (R)
- Versions Lasso (Python) et PLS de la régression logistique
- Extensions : effets aléatoires, mesures répétées

Exemple simple

Influence du débit et du volume d'air inspiré sur la dilatation des vaisseaux sanguins superficiels des membres inférieurs

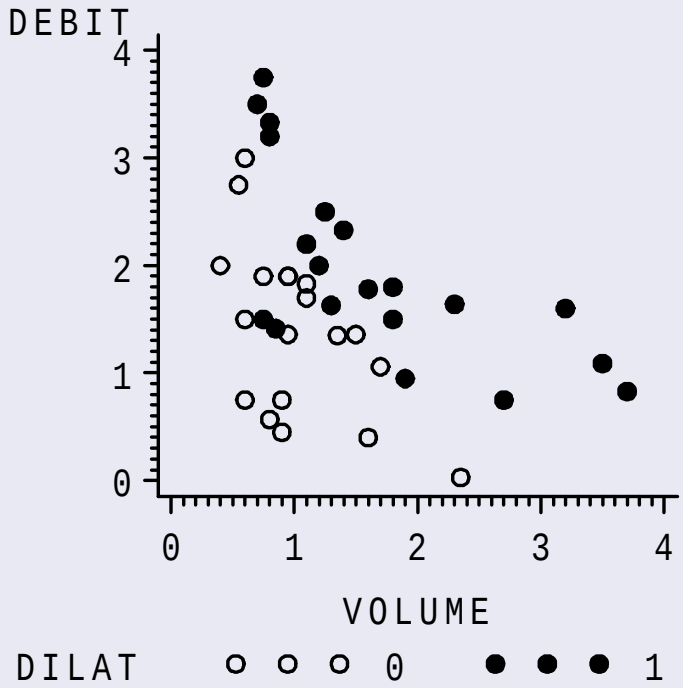


FIGURE – Dilatation : Nuage des modalités de Y

Sorties SAS

```
The LOGISTIC Procedure
                              Intercept
                 Intercept        and
 Criterion         Only       Covariates    Chi-Square for Covariates
 AIC                56.040        35.216          .
 SC                 57.703        40.206          .
 -2 LOG L           54.040        29.216(1)    24.824 with 2 DF (p=0.0001)
 Score                .             .          16.635 with 2 DF (p=0.0002)

              Parameter(2) Standard    Wald(3)     Pr >    Standardized  Odds
Variable DF    Estimate    Error   Chi-Square  Chi-Square  Estimate     Ratio
INTERCPT 1      2.8782     1.3214     4.7443     0.0294           .        .
L_DEBIT  1     -4.5649     1.8384     6.1653     0.0130    -2.085068    0.010
L_VOLUME 1     -5.1796     1.8653     7.7105     0.0055    -1.535372    0.006
```

Régression logistique ordinale

Variables :

1. Etat du conducteur : Normal ou Alcoolisé
2. Sexe du conducteur
3. Port de la ceinture : Oui Non
4. Gravité des blessures : 0 : rien à 3 : fatales

Sorties SAS

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	Gr0	1	1.8699	0.0236	6264.9373	<.0001
Intercept	Gr1	1	2.8080	0.0269	10914.3437	<.0001
Intercept	Gr2	1	5.1222	0.0576	7917.0908	<.0001
sexe	Sfem	1	-0.3118	0.0121	664.3353	<.0001
alcool	A_bu	1	-0.5017	0.0190	697.0173	<.0001
ceinture	Cnon	1	-0.1110	0.0174	40.6681	<.0001

Test de score pour l'hypothèse des cotes proportionnelles

Khi-2	DDL	Pr > Khi-2
33.3161	6	<.0001

Modèle plus simple : GrN vs. GrO

Estimations des rapports de cotes

Effet	Valeur estimée	IC de Wald à 95 %
sexe Sfem vs Shom	1.873	1.786 1.964
alcool A_bu vs Ajeu	2.707	2.512 2.918
ceinture Cnon vs Coui	1.244	1.162 1.332

Diagnostic de cancer

- Wisconsin Breast Cancer Database (`mlbench` de R)
- 9 variables ordinales ou nominales à 10 modalités
- 683 observations

  Clump Thickness
  Uniformity of Cell Size
  Uniformity of Cell Shape
  Marginal Adhesion
  Single Epithelial Cell Size
  Bare Nuclei
  Bland Chromatin
  Normal Nucleoli
  Mitoses
  "benign" et "malignant"

- Avec toutes les variables : ajustement exact (0%) mais erreur de 5,8%
- Modèle réduit : ajustement de 3,5% et erreur de 5,1%

Dépassement de seuil

- Prévision directe des dépassements ($150\mu g/m^3$ au lieu 180)
- Problèmes : ils sont peu nombreux
- Modèle optimal au sens d'Akaïke sans interaction

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			831	744.34	
O3_pr	1	132.89	830	611.46	9.576e-31
vmodule	1	2.42	829	609.04	0.12
s_rmh2o	1	33.71	828	575.33	6.386e-09
station	4	16.59	824	558.74	2.324e-03
TEMPE	1	129.39	823	429.35	5.580e-30

- `vmodule` est-elle utile ?

Comparaison de modèles

- Avec et sans "vmodule", avec et sans interaction
- A partir du quantitatif ou non, MOCAGE

```
Matrices de confusion de l'échantillon test pour différents modèles :
          0   1                   0   1                   0   1                   0   1
  FALSE 163  19           FALSE 162  18         FALSE 163  17         FALSE 160  13
  TRUE    5  21           TRUE    6  22         TRUE    5  23         TRUE    8  27
logistique sans vmodule   avec vmodule       avec interactions       quantitatif
Erreur : 11,5%                11,5%                 10,6%                 10,1%
MOCAGE : 13,6%
```

- Biais systématique
- Besoin de préciser ces estimations d'erreurs

Gestion de la Relation Client

- Données en provenance d'I-BP
- 1425 clients
- 32 variables "comptables"
- Objectif : score d'appétance de la carte visa premier.
  1. Nettoyage des données
  2. Transformations
  3. Comparaison des modélisations

Données bancaires : liste des variables

Identif.	Libellé	Identif.	Libellé
matric	Matricule (identifiant client)	qcred	Moyenne des mouvements créditeurs en Kf
sexec	Sexe (qualitatif)	dmvtp	Age du dernier mouvement (en jours)
ager	Age en années	boppn	Nombre d'opérations à M-1
famil	Situation familiale	facan	Montant facturé dans l'année en francs
	(Fmar : marié, Fcel : célib., Fdiv :divorcé,	lgagt	Engagement long terme
	Fuli :union libre, Fsep : séparés, Fveu :veuf)	vienb	Nombre de produits contrats vie
relat	Ancienneté de relation en mois	viemt	Montant des produits contrats vie en francs
prcsp	Catégorie socio-professionnelle (code num)	uemnb	Nombre de produits épargne monétaire
opgnb	Nombre d'opérations par guichet dans le mois	xlgnb	Nombre de produits d'épargne logement
moyrv	Moyenne des mouvements nets créditeurs	xlgmt	Montant des produits d'épargne logement en francs
	des 3 mois en Kf	ylvnb	Nombre de comptes sur livret
tavep	Total des avoirs épargne monétaire en francs	ylvmt	Montant des comptes sur livret en francs
endet	Taux d'endettement	rocnb	Nombre de paiements par carte bancaire à M-1
gaget	Total des engagements en francs	jntca	Nombre total de cartes
gagec	Total des engagements court terme en francs	nptag	Nombre de cartes point argent
gagem	Total des engagements moyen terme en francs	itavc	Total des avoirs sur tous les comptes
kvunb	Nombre de comptes à vue	havef	Total des avoirs épargne financière en francs
qsmoy	Moyenne des soldes moyens sur 3 mois	dnbjd	Nombre de jours à débit à M
		carvp	**Possession de la carte VISA Premier**

GRC : modélisation

Sélection par méthode descendante de la procédure logistic sur échantillon d'apprentissage

Type 3 Analysis of Effects

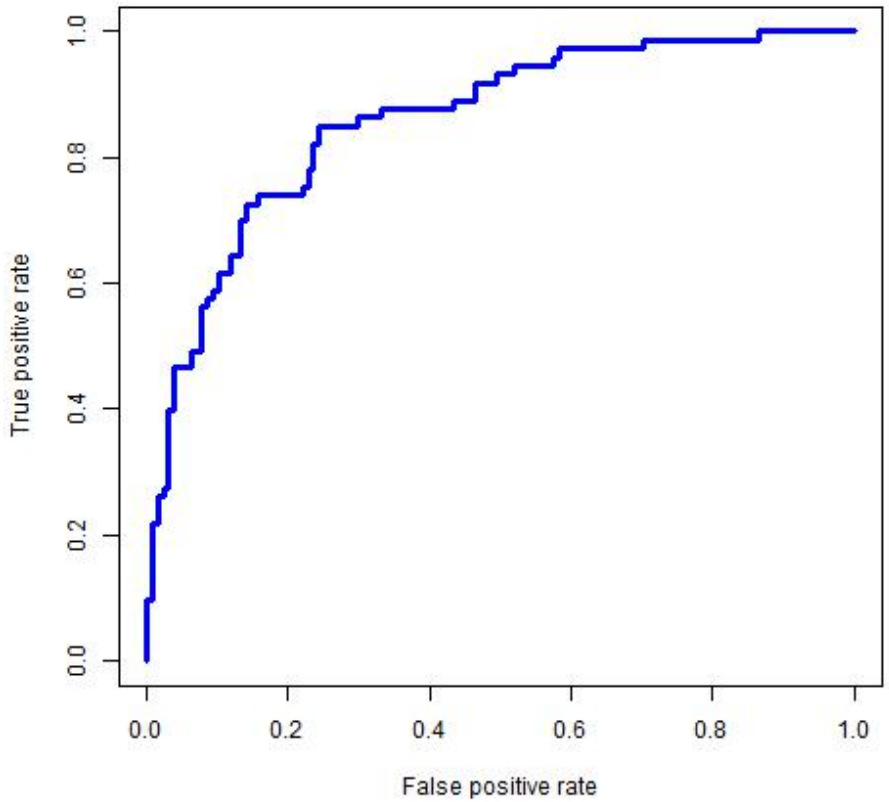
Effect	DF	Wald Chi-Square	Pr > ChiSq
SEXEQ	1	22.7707	<.0001
PCSPQ	4	41.4504	<.0001
kvunbq	1	10.7444	0.0010
uemnbq	2	6.0831	0.0478
nptagq	1	5.0194	0.0251
facanq	1	8.1289	0.0044
relatq	2	18.4219	<.0001
opgnbq	2	15.8660	0.0004
moyrvq	2	65.7911	<.0001
dmvtpq	2	134.7367	<.0001
itavcq	2	9.5263	0.0085

GRC : prévision

Matrices de confusion, estimée sur échantillons d'apprentissage et test

CARVPr Frequency| Percent	predy Frequency| 0	1	Total
0	535	38	573
	61.57	4.37	65.94
1	51	245	296
	5.87	28.19	34.06
Total	586	283	869
	67.43	32.57	100.00

CARVPr Percent |	predy 0	1	Total
0	131	8	139
	65.50	4.00	69.50
1	15	46	61
	7.50	23.00	30.50
Total	146	54	200
	73.00	27.00	100.00



Données bancaires : estimation sur l'échantillon test de la courbe ROC associée à la régression logistique.