MODELISATION DE DONNÉES QUALITATIVES

REGRESSION LOGISTIQUE MULTIPLE

Pierre-Louis Gonzalez

RÉGRESSION LOGISTIQUE MULTIPLE

(cas Y binaire)

La régression logistique multiple généralise la régression logistique simple au cas où il y a plusieurs variables explicatives X_1 ... X_k.

I - LE MODÈLE

On note $x = (x_1 \cdots, x_k)$ une valeur de $X = (X_1, \cdots, X_k)$.

$$\Pi(x) = Prob(Y = 1 \mid X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k}}$$

1 Estimation par le maximum de vraisemblance

On a :

$$\mathbf{Log\ L(\beta) = \sum_{i=1}^{n} \left[y_i\ Log\ \Pi_i + (1 - y_i)\ Log\ (1 - \Pi_i) \right]}$$

où $\Pi_i = \Pi(x_i) = \Pi(x_{i1}, \ldots, x_{ik})$

On obtient $\hat{\beta}$ en annulant les dérivées partielles :

$$\frac{\partial\ Log\ L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - \Pi_i) = 0 \qquad j=0, \ldots, k$$

$$(\text{où } x_{i0} = 1, \text{ pour tout } i)$$

Matrice Variances-Covariances de $\hat{\beta}$:

$$X = \begin{pmatrix} 1 & x_{11} & x_{1k} \\ \vdots & & \\ 1 & x_{n1} & x_{nk} \end{pmatrix} \qquad V = \begin{pmatrix} \hat{\Pi}_1\left(1-\hat{\Pi}_1\right) & & 0 \\ & \ddots & \\ 0 & & \hat{\Pi}_n\left(1-\hat{\Pi}_n\right) \end{pmatrix}$$

On a $\mathbf{Var}\left(\hat{\beta}\right) = \left(X'VX\right)^{-1}$

2 Influence globale des variables

Modèle $$\Pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}$$

Test
$$\begin{cases} H_0 : \beta_1 = \ \ldots \ = \beta_k = 0 \\ H_1 : \text{ au moins un } \beta_j \neq 0 \end{cases}$$

Trois statistiques

- **Statistique de Wald**
- **Rapport des vraisemblances**

$$\Lambda = -2 \ \text{Log} \left[\frac{L\,(\text{constante})}{L\,(\text{toutes les variables})} \right] \longrightarrow \chi^2(\mathbf{k}) \text{ sous } \mathbf{H_0}$$

- **Score**

$$\hat{\beta}_{H_0} = \begin{pmatrix} \mathbf{Log}\ \dfrac{\mathbf{n_1}}{\mathbf{n_0}} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

$$\text{SCORE} = U\left(\hat{\beta}_{H_0}\right)' \hat{I}\left(\hat{\beta}_{H_0}\right)^{-1} U\left(\hat{\beta}_{H_0}\right) \longrightarrow \chi^2(k) \text{ sous } H_0$$

3 Influence d'un groupe de variables

Test $H_0 : \beta_{r+1} = \ \dots \ = \beta_k = 0$

$H_1 :$ au moins un de ces $\beta_j \neq 0$

Statistiques utilisées

Rapport des vraisemblances

$$\Lambda = -2 \, \mathrm{Log} \left[\frac{L\left(X_1, \cdots, X_r\right)}{L\left(X_1, \cdots, X_k\right)} \right] \longrightarrow \chi^2_{k-r} \quad \text{sous } H_0$$

Score

$$U\left(\hat{\beta}_{H_0}\right)' \hat{I}\left(\hat{\beta}_{H_0}\right)^{-1} U\left(\hat{\beta}_{H_0}\right) \longrightarrow \chi^2_{k-r} \quad \text{sous } H_0$$

II SÉLECTION DE VARIABLES : RÉGRESSION LOGISTIQUE PAS À PAS

1 Pas à pas ascendant

À chaque étape, on sélectionne la variable qui aura le «score» le plus élevé une fois introduite dans le modèle.

Soit $\mathbf{X_1}$, **...** , $\mathbf{X_t}$ les variables déjà introduites. On choisit la variable hors modèle $\mathbf{X_j}$ maximisant le score $\mathbf{(X_j)}$ dans le modèle contenant les variables $\mathbf{X_1}$, **...** , $\mathbf{X_t}$, $\mathbf{X_j}$.

L'influence des variables hors modèle $\mathbf{X_{t+1}}$ **...** $\mathbf{X_k}$ en plus des variables $\mathbf{X_1}$ **...** $\mathbf{X_t}$ est testée globalement à l'aide de la **statistique Score (notée «Residual chi-square dans SAS).**

Si les variables hors modèle $\mathbf{X_{t+1}}$, **...** , $\mathbf{X_k}$ sont sans influence marginale sur $\mathbf{Y}$, alors :

$$\text{"Residual chi} - \text{square"} \longrightarrow \chi^2_{(k-t)}$$

2. Pas à pas descendant

À chaque étape, on enlève la variable ayant le plus petit «Wald chi-square» à condition qu'il ne soit pas significatif.

3. Meilleur sous-ensemble.

Cette méthode permet de déterminer les meilleurs modèles à une variable explicative, à deux variables explicatives, à trois variables explicatives…..

Méthode basée sur le score de vraisemblance et utilisant l'algorithme de Furnival et Wilson.

Choix parmi les 2^K -1 modèles possibles.

Cette option n'est pas disponible en présence de variables qualitatives identifiées par l'instruction CLASS.

III RÉSUMÉ DES TESTS DE VALIDITÉ GÉNÉRALE DU MODÈLE

1 Existe-t-il des statistiques permettant de juger de la bonne adéquation du modèle, en jouant un rôle analogue à celui du $\mathbf{R^2}$ classique ?

- → Déviance
- → Rapport de vraisemblance
- → Statistique du score

→ Critère d'Akaike

$$AIC = -2 \text{ Log } L + 2k$$

(**k** = nombre de paramètres à estimer)

→ Critère de Schwartz

$$SC = -2 \text{ Log } L + k \text{ Log } n$$

(**n** = nombre total d'observations)

Les critères de Schwartz et d'Akaike sont utiles pour comparer des modèles différents portant sur les mêmes données (nombre de d.d.l. différents).

On préférera le modèle pour lequel ces statistiques ont la valeur la plus faible.

2 Tables de classement

Objectif **Évaluer la capacité prédictive du modèle**

$$\begin{cases} \hat{Y}_i = 1 \text{ Si } \hat{\Pi}_1 \geq 1/2 \\ \hat{Y}_i = 0 \text{ Si } \Pi_1 < 1/2 \end{cases}$$

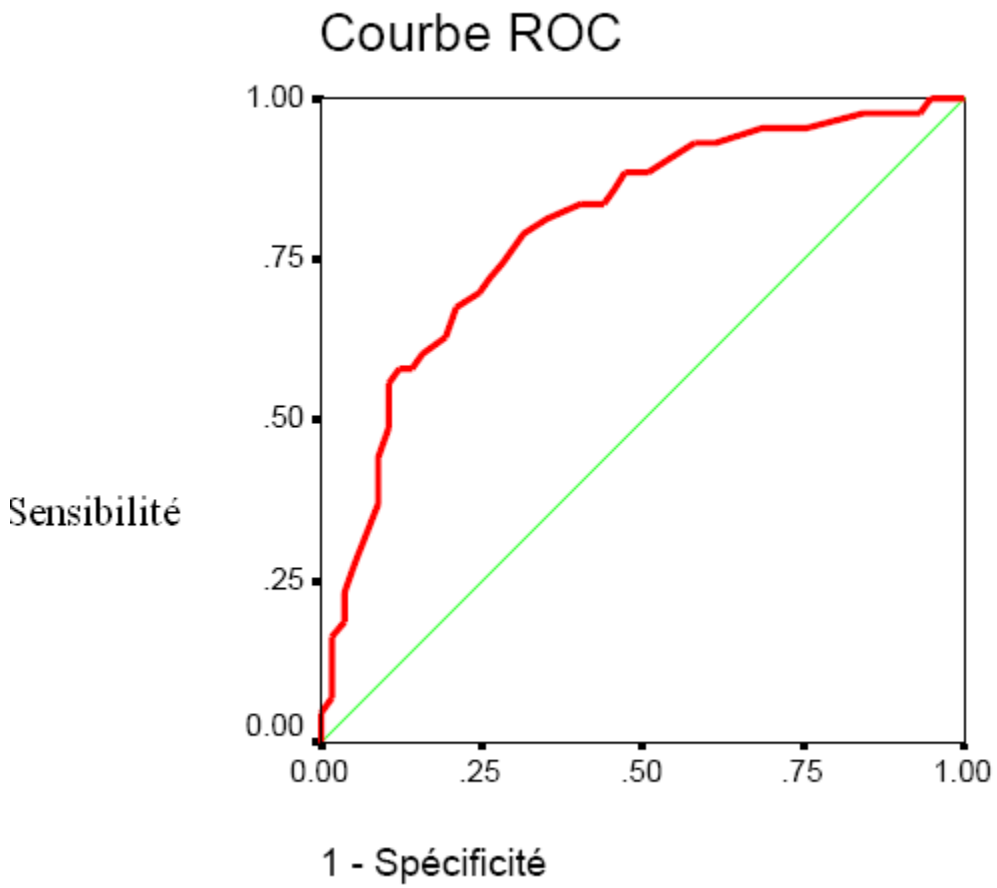
		PRÉDICTION	
		$\hat{Y}_i = 1$	$\hat{Y}_i = 0$
OBSERVATION	$Y_i = 1$	a	b
	$Y_i = 0$	c	d

SENSIBILITE	$\frac{a}{a+b}$
SPÉCIFICITÉ	$\frac{d}{c+d}$
TAUX D'ERREUR PAR EXCÈS	$\frac{c}{a+c}$
TAUX D'ERREUR PAR DÉFAUT	$\frac{b}{b+d}$
BIEN CLASSÉS	$\frac{a+d}{n}$

La capacité prédictive du modèle est d'autant meilleure que :

- **SENSIBILITE**
- **SPECIFICITÉ**
- **BIEN CLASSÉS**

sont élevés.



Site consacré à la courbe ROC

http://www.anaesthetist.com/mnm/stats/roc/

3 Indices de corrélation entre «PRÉDICTION» et «OBSERVATION»

$$t = \text{Card}\left\{\left(Y_j, Y_k\right) \ / \ Y_j \neq Y_k\right\}$$ **nombre de paires différentes**

$$n_c = \text{Card}\left\{\left(Y_j = 1, Y_k = 0\right) , \hat{Y}_j > \hat{Y}_k\right\}$$ **nombre de paires concordantes**

$$n_d = \text{Card}\left\{\left(Y_j = 1 , Y_k = 0\right) \ / \ \hat{Y}_j < \hat{Y}_k\right\}$$ **nombre de paires discordantes**

t = nombre de concordances + nombre de discordances + nombre d'ex aequo

1. $$c = \frac{n_c + 0{,}5\left(t - n_c - n_d\right)}{t} \in [0,1]$$

En l'absence d'ex aequo, c représente l'aire sous la courbe ROC égale à n_c / t souvent noté AUC

2. Somer's $$D = \frac{n_c - n_d}{t} \in [-1,+1]$$

En l'absence d'ex aequo le D de Somer's est l'indice de Gini: 2AUC-1

3. Gamma (Goodman-Kruskal) $$= \frac{n_c - n_d}{n_c + n_d} \in [-1,+1]$$

4. Kendall's tau-a $$= \frac{n_c - n_d}{0{,}5\ n\left(n-1\right)} \in [-1,+1]$$

La capacité prédictive du modèle est d'autant meilleure que les indices de corrélation sont élevés.

IV INTERPRÉTATION DES COEFFICIENTS DE LA RÉGRESSION LOGISTIQUE

1 Cas d'une variable indépendante binaire

Y Variable dépendante	Variable indépendante **X**	
	$\mathbf{x = 1}$	$\mathbf{x = 0}$
$\mathbf{y = 1}$	$\Pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\Pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$\mathbf{y = 0}$	$1-\Pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\Pi(0) = \frac{1}{1+e^{\beta_0}}$

La cote (**odds**) d'obtenir **Y=1** pour les personnes ayant **X=1** est définie par :

$$\frac{\Pi_1}{1-\Pi_1}$$

La cote (**odds**) d'obtenir **Y=1** pour les personnes ayant **X=0** est définie par :

$$\frac{\Pi_0}{1-\Pi_0}$$

Le rapport de cotes (odds ratio) noté ψ est défini par :

$$\psi = \frac{\Pi(1)/(1-\Pi(1))}{\Pi(0)/(1-\Pi(0))}$$

En notant : $$g(1) = \text{Log}\ \frac{\Pi(1)}{1-\Pi(1)}$$

$$g(0) = \text{Log}\ \frac{\Pi(0)}{1-\Pi(0)}$$

le logarithme de l'odds ratio ou log odds est :

$$\text{Log}\,(\psi) = \text{Log}\left[\frac{\Pi(1)/1-\Pi(1)}{\Pi(0)/1-\Pi(0)}\right] = g(1) - g(0)$$

soit la différence de logits.

Le logit $g(x)$ valant $\beta_0 + \beta_1 x$, il vient : $g(1) - g(0) = \beta_1$

d'où $\text{Log}(\psi) = \beta_1 \Rightarrow \psi = e^{\beta 1}$

Interprétation de l'odds ratio

Lorsque $\Pi(1)$ et $\Pi(0)$ sont petits.

$$\psi = \frac{\Pi(1)/1-\Pi(1)}{\Pi(0)/1-\Pi(0)} \cong \frac{\Pi(1)}{\Pi(0)} = \eta = \text{risque relatif}$$

On communique souvent les résultats sous cette forme dans un contexte médical.

Exemple CHD «présence d'une maladie cardiaque»

CHD (y)	$\geq 55\ (1)$	$< 55\ (0)$	TOTAL
Présent (1)	21	22	43
Absent (0)	6	51	57
TOTAL	27	73	100

$$\hat{\psi} = \frac{21/6}{22/51} = 8{,}11$$

Le quotient de malades par rapport aux non malades est huit fois plus important chez les plus de 55 ans que chez les moins de 55 ans.

Résultats de la régression logistique de Y sur l'âge dichotomisé

Variable	Coefficient estimé	Erreur-type	$\frac{\text{coefficient}}{\text{erreur}-\text{type}}$	$\hat{\psi}$
Âge	2,094	0,529	3,96	8,1
Constante	- 0,841	0,255	- 3,30	

$$\hat{\psi} = e^{\beta_1} = e^{2,094} = 8,1$$

De plus, on obtient un **intervalle de confiance de l'odds ratio** (ici à 95 %)

$$\exp\left[\hat{\beta}_1 \pm 1,96\ s\left(\hat{\beta}_1\right)\right] = \exp\left[2,094 \pm 1,96*0,529\right] = \left[2,9\ ;\ 22,9\right]$$

Remarque Les résultats présentés dans cette section dépendent du codage utilisé.

Si on avait codé $\mathbf{X}$ par $\mathbf{1} = (\geq 55)$ et $-\mathbf{1} = (< 55)$

alors : $\text{Log}(\hat{\psi}) = \hat{g}(1) - \hat{g}(-1)$

$$= \hat{\beta}_0 + \hat{\beta}_1 - (\hat{\beta}_0 - \hat{\beta}_1)$$

$$= 2\,\hat{\beta}_1$$

et $\hat{\psi} = \exp(2\,\hat{\beta}_1)$

2 Variable indépendante polytomique

On souhaite étudier la relation entre la maladie cardiaque et le groupe.

CHD	Blanc	Noir	Hispanique	Autre	TOTAL
Présent	5	20	15	10	50
Absent	20	10	10	10	50
Total	25	30	25	20	100
Odds ratio $\hat{\psi}$	1	8	6	4	
I.C. 95 % $\hat{\psi}$		[2,3 ; 27,6]	[1,7 ; 21,3]	[1,1 ; 14,9]	
Log $\hat{\psi}$	0	2,08	1,79	1,39	

Le groupe «blanc» est utilisé comme groupe de référence.

Le résultats du tableau ci-dessus peuvent être retrouvés à l'aide de la **régression logistique de Y = CHD sur les variables indicatrices des groupes «Noir» «Hispanique» «Autre».**

Régression logistique de CHD sur Noir, Hispanique, Autre

Variable	Coefficient estimé	Erreur-type	$\frac{\text{coefficient}}{\text{erreur}-\text{type}}$	$\hat{\psi}$
Groupe (1)	2,079	0,633	3,29	8,0
Groupe (2)	1,792	0,646	2,78	6.0
Groupe (3)	1,386	0,671	2,07	4.0
Constante	- 1,386	0,500	- 2,77	

PROGRAMME SAS

```
options ls=75 ps=50 nodate nocenter nopage;
data CHDMULT;
input groupe $ chd $ effectif;
cards;
ref present 5
ref absent 20
gr1 present 20
gr1 absent 10
gr2 present 15
gr2 absent 10
gr3 present 10
gr3 absent 10
;
proc logistic  order=data ;
freq effectif ;
class groupe / param= glm;
model chd = groupe /  ctable;
run;
```

Estimations par l'analyse du maximum de vraisemblance						
Paramètre		DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		1	-1.3863	0.5000	7.6871	0.0056
groupe	gr1	1	2.0794	0.6325	10.8100	0.0010
groupe	gr2	1	1.7917	0.6455	7.7048	0.0055
groupe	gr3	1	1.3863	0.6708	4.2706	0.0388
groupe	ref	0	0	.	.	.

Estimations des rapports de cotes			
Effet	Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
groupe gr1 vs ref	8.000	2.316	27.633
groupe gr2 vs ref	6.000	1.693	21.261
groupe gr3 vs ref	4.000	1.074	14.895