



2ème année ENSAI, 2010 - 2011

Régression logistique avec R

Laurent Rouvière

ENSAI - Campus de Ker Lann
Rue Blaise Pascal - BP 37203
35172 BRUZ cedex
Tel : 02 99 05 32 63

Mel : laurent.rouviere@ensai.fr

Table des matières

1	Introduction	5
1.1	Rappels sur le modèle linéaire	5
1.2	Le modèle linéaire généralisé : GLM	6
1.2.1	La régression logistique	6
1.2.2	La régression log-linéaire	9
1.2.3	Généralisation : GLM	10
1.3	Exemples de fonctions de liens pour la régression d'une variable binaire	11
2	Analyse discriminante logistique	13
2.1	Le modèle statistique	13
2.1.1	L'échantillonnage	13
2.1.2	Identifiabilité du modèle	14
2.2	L'estimateur du maximum de vraisemblance	15
2.2.1	La vraisemblance	15
2.2.2	Comportement asymptotique de l'estimateur du maximum de vraisemblance	16
2.3	L'algorithme IRLS	18
2.3.1	Rappel sur l'algorithme de Newton-Raphson	18
2.3.2	Calcul des estimateurs	18
2.4	Dimensions explicatives, variables explicatives	19
2.4.1	Variable explicative quantitative	19
2.4.2	Variable explicative qualitative	20
2.4.3	Interactions	21
2.5	Interprétation des coefficients β	22
2.6	Précision des estimateurs et tests	24
2.6.1	Loi asymptotique	24
2.6.2	Intervalles de confiance	24
2.6.3	Tests de nullité de q coefficients libres	26
2.7	Le schéma d'échantillonnage rétrospectif	28
2.8	Un exemple avec R	30
2.8.1	Modèles "simples"	30
2.8.2	Encore d'autres modèles...	32
3	Sélection et validation de modèles	35
3.1	Sélection ou choix de modèle	35
3.1.1	Un outil spécifique : la déviance	35
3.1.2	Test de déviance entre 2 modèles emboîtés	38
3.1.3	Critère de choix de modèles	38
3.1.4	Apprentissage/validation	39

3.1.5	Validation croisée	42
3.1.6	Sélection automatique	43
3.2	Validation du modèle	46
3.2.1	Test d'adéquation de la déviance	46
3.2.2	Test d'Hosmer Lemeshow	47
3.2.3	Analyse des résidus	47
3.2.4	Points leviers et points influents	52
4	Modèle logistique multi-classes	55
4.1	Le modèle saturé ou modèle multinomial	55
4.2	Modèle polytomique ordonné	56
4.2.1	Cas binaire	56
4.2.2	Généralisation	57
4.2.3	L'égalité des pentes	59
4.2.4	Le test d'égalité des pentes	60
4.3	Le modèle polytomique nominal	61
4.3.1	Le modèle	61
4.3.2	Estimation et interprétation des paramètres	62
	Annexes	65
A.1	Rappels sur la méthode du maximum de vraisemblance	65
A.2	Echantillonnage Rétrospectif	67
A.3	Exercices	69
A.4	Correction	73
	Bibliographie	77

Chapitre 1

Introduction

Notations :

- $X = (1, \mathbf{X}_1, \dots, \mathbf{X}_p)'$: vecteur aléatoire de dimension $p + 1$, les marginales \mathbf{X}_j sont les variables explicatives. On note également $x = (1, \mathbf{x}_1, \dots, \mathbf{x}_p)'$ une réalisation de X ;
- Y variable (univariée) à expliquer.
- $(X_1, Y_1), \dots, (X_n, Y_n)$: un n -échantillon aléatoire (i.i.d. et de même loi que le couple (X, Y)), tel que $X_i = (1, X_{i1}, \dots, X_{ip})'$;
- $(x_1, y_1), \dots, (x_n, y_n)$ une réalisation de $(X_1, Y_1), \dots, (X_n, Y_n)$.
- \mathbf{X} : la matrice des observations :

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}.$$

1.1 Rappels sur le modèle linéaire

Le contexte

Nous cherchons à expliquer une variable Y par p variables $X = (1, \mathbf{X}_1, \dots, \mathbf{X}_p)'$. Pour ce faire, on dispose de n réalisations $(x_1, y_1), \dots, (x_n, y_n)$ du couple (X, Y) . Le but est de modéliser la dépendance de la variable réponse Y sur les variables explicatives $\mathbf{X}_1, \dots, \mathbf{X}_p$. Plusieurs raisons peuvent motiver cette modélisation :

- **la description** : on veut un modèle qui permette de décrire la relation entre Y et X ;
- **l'évaluation** des contributions relatives de chaque prédicteur pour expliquer Y ;
- **la prédiction** : prévoir la valeur de Y pour des nouvelles valeurs des variables explicatives.

On rappelle que le modèle linéaire s'écrit :

$$Y = X'\beta + \epsilon = \beta_0 + \beta_1\mathbf{X}_1 + \dots + \beta_p\mathbf{X}_p + \epsilon,$$

avec $\beta = (\beta_0, \beta_1, \dots, \beta_p)' \in \mathbb{R}^{p+1}$ et $\epsilon \sim \mathcal{N}(0, \sigma^2)$. On distingue alors deux cas :

- Les variables X_i sont déterministes (non aléatoires) :

$$Y \sim \mathcal{N}(X'\beta, \sigma^2), \quad \mathbf{E}[Y] = X'\beta;$$

- Les variables X_i sont aléatoires :

$$(Y|X) \sim \mathcal{N}(X'\beta, \sigma^2), \quad \mathbf{E}[Y|X] = X'\beta.$$

Plaçons nous maintenant dans le cas où la variable à expliquer Y est qualitative (sexe, couleur, présence ou absence d'une maladie...) et possède un nombre fini de modalités g_1, \dots, g_m . Le problème consiste alors à expliquer l'appartenance d'un individu à un groupe à partir des p variables explicatives, on parlera de *discrimination* au lieu de régression.

Il est bien entendu impossible de modéliser directement la variable Y par une relation linéaire (imaginons que Y soit le sexe d'une personne ou la couleur de ces cheveux). Afin de pallier à cette difficulté, on va s'intéresser aux probabilités $\mathbf{P}(Y = g_k | X = x)$. Supposons pour simplifier que la variable Y prenne uniquement deux valeurs : 0 ("groupe 0") ou 1 ("groupe 1"). La connaissance de $\mathbf{P}(Y = 1 | X = x)$ implique celle de $\mathbf{P}(Y = 0 | X = x)$: il suffit par conséquent de modéliser la probabilité $p(x) = \mathbf{P}(Y = 1 | X = x)$. On peut par exemple envisager une relation de la forme

$$p_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon = x' \beta + \varepsilon.$$

Cette approche possède plusieurs inconvénients :

- Remarquons tout d'abord que la variance de $Y | X = x$ vaut $p_\beta(x)(1 - p_\beta(x))$. Contrairement au modèle linéaire traditionnel, cette variance n'est pas constante et par conséquent l'hypothèse d'homoscédasticité des résidus ne sera pas vérifiée.
- Le fait qu'aucune restriction ne soit effectuée sur les β implique que $x' \beta$ peut prendre n'importe quelle valeur dans \mathbb{R} . Ce fait est gênant pour l'estimation d'une probabilité (imaginer une estimation du genre $\mathbf{P}_\beta(Y = 1 | X = x) = -1297.56!!!$).

Pour ces raisons, nous devons étendre le modèle linéaire classique aux cas où :

- Y peut être une variable qualitative (présence ou absence d'une maladie, appartenance à une catégorie...);
- les erreurs peuvent ne pas avoir la même variance (s'affranchir de l'hypothèse d'homoscédasticité).

1.2 Le modèle linéaire généralisé : GLM

1.2.1 La régression logistique

Nous nous plaçons tout d'abord dans un contexte de classification binaire, c'est-à-dire que nous supposons qu'il existe seulement deux groupes à discriminer. Nous verrons dans le chapitre 4 comment étendre les techniques à des modèles *multiclassés* (plus de deux groupes).

Exemple 1.1

Nous souhaitons expliquer la variable Y présence (1)/ absence (0) d'une maladie cardio-vasculaire (Chd) par l'âge des patients. Les données sont représentées sur la figure 1.1.

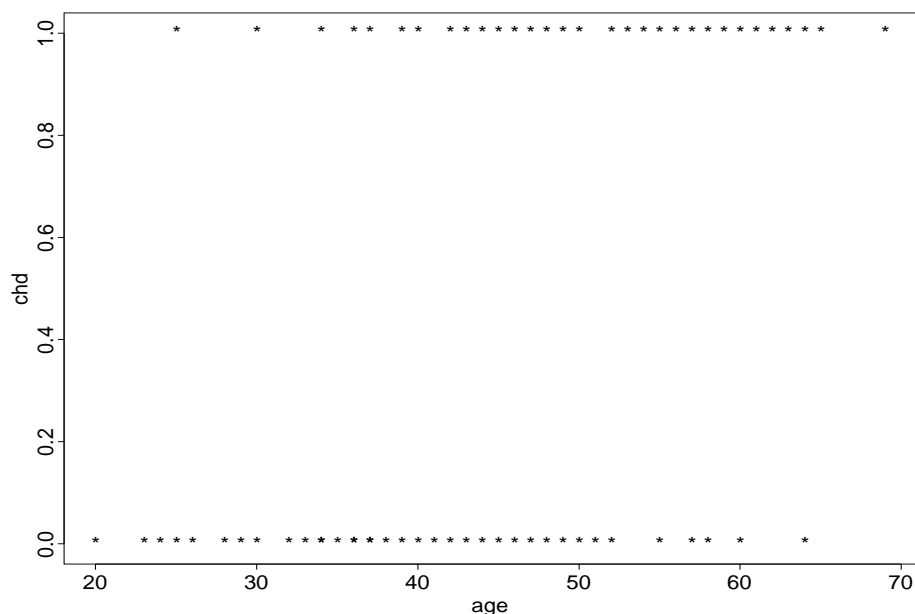


FIGURE 1.1 – Représentation directe de Chd (variable à expliquer Y) en fonction de l'âge (variable explicative X).

Cette figure montre qu'il est difficile de modéliser les données brutes, la variabilité de la variable CHD est élevée pour tout âge. Une méthode permettant de réduire cette variabilité consiste à regrouper les patients par classe d'âge. Nous obtenons le tableau suivant :

Age	Effectif	CHD		Moyenne
		Absent	Présent	
[19 ;29]	10	9	1	0.1
[29 ;34]	15	13	2	0.133333
[34 ;39]	12	9	3	0.25
[39 ;44]	15	10	5	0.333333
[44 ;49]	13	7	6	0.461538
[49 ;54]	8	3	5	0.625
[54 ;59]	17	4	13	0.764706
[59 ;69]	10	2	8	0.8

TABLE 1.1 – Données regroupées en classe d'âge.

La liaison entre l'âge et la présence de la maladie devient beaucoup plus claire. Il apparaît en effet que lorsque l'âge augmente, la proportion d'individus atteint par la maladie augmente. La figure 1.2 permet d'évaluer cette liaison : elle apparaît nettement sous la forme d'une courbe sigmoïde (en forme de "S"). Il semblerait donc "naturel" de modéliser cette proportion de malades par classe d'âge en fonction de l'âge par une courbe sigmoïde.

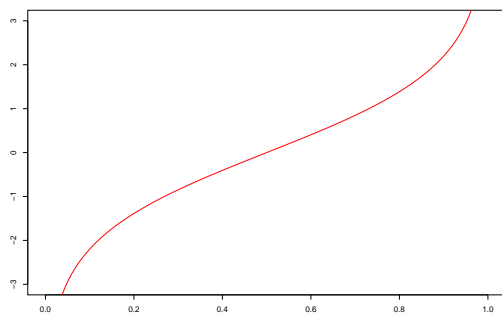


FIGURE 1.3 – Fonction logit .

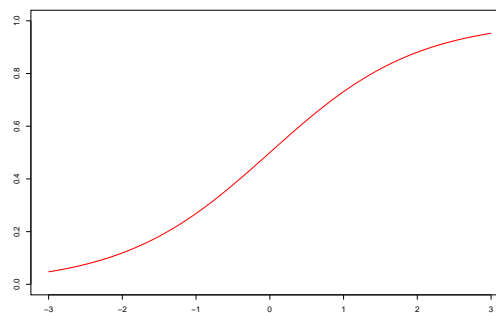


FIGURE 1.4 – Fonction inverse de logit .

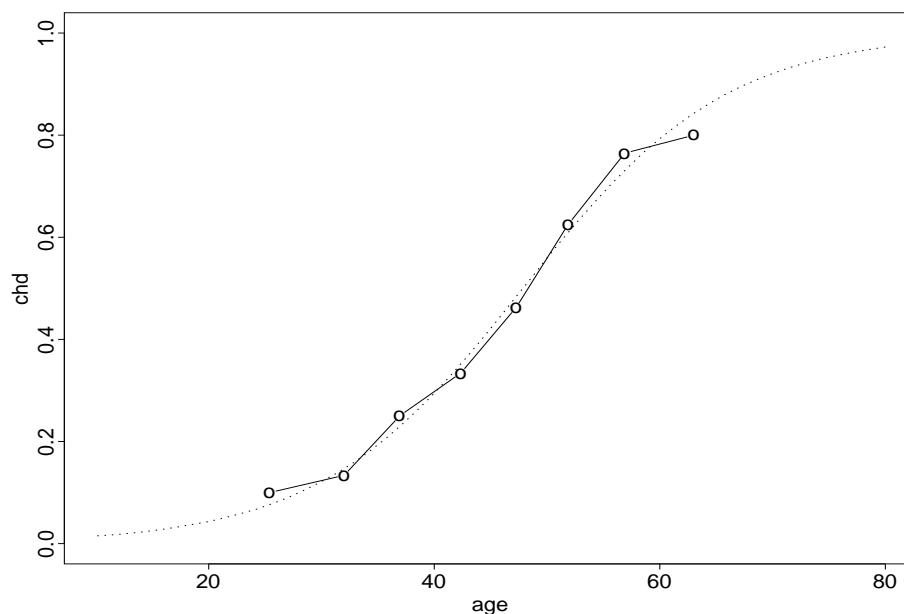


FIGURE 1.2 – Fréquence de Chd par classe d'âge en fonction de l'âge.

La colonne “moyenne” du tableau 1.1 fournit une estimation de $\mathbf{E}[Y|X = x]$ pour chaque classe d'âge. Nous pouvons donc proposer une modélisation de l'espérance conditionnelle de $\mathbf{E}[Y|X = x]$:

$$\mathbf{E}[Y|X = x] = h_{\beta}(x)$$

où l'allure de la courbe représentative de h_{β} est une sigmoïde.

Plusieurs fonctions h_{β} peuvent naturellement être utilisées. Pour le modèle logistique on considère la fonction $h(x) = \exp(x)/(1 + \exp(x))$, ce qui donne le modèle

$$\mathbf{E}[Y|X = x] = p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)},$$

où encore

$$\text{logit } p(x) = \log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x,$$

logit désignant la fonction bijective et dérivable de $]0, 1[$ dans $\mathbb{R} : p \mapsto \log(p/(1 - p))$ (voir figures 1.3 et 1.4). Nous verrons qu'une telle modélisation permettra de retrouver un grand nombre des “bonnes” propriétés du modèle linéaire.

La loi conditionnelle de la variable d'intérêt diffère entre le modèle logistique et le modèle linéaire. Dans le modèle de régression linéaire $Y = \beta_0 + \beta_1 X + \varepsilon$, on fait l'hypothèse que les résidus ε

suivent une loi $\mathcal{N}(0, \sigma^2)$. Il vient $Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$. Pour le modèle logistique, pour une observation x de la variable explicative, on peut exprimer la variable d'intérêt comme suit :

$$Y = p(x) + \varepsilon.$$

La variable aléatoire ε peut prendre simplement deux valeurs : si $y = 1$ alors $\varepsilon = 1 - p(x)$ et si $y = 0$ alors $\varepsilon = -p(x)$. Par conséquent ε prend pour valeur $1 - p(x)$ avec probabilité $p(x)$ et $-p(x)$ avec probabilité $1 - p(x)$: $Y|X = x$ suit une loi de Bernoulli de paramètre $p(x)$.

Définition 1.1 (Régression logistique)

Soit Y une variable à valeurs dans $\{0, 1\}$ à expliquer par p variables explicatives $X = (1, \mathbf{X}_1, \dots, \mathbf{X}_p)'$. Le modèle logistique propose une modélisation de la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $p_\beta(x) = \mathbf{P}_\beta(Y = 1|X = x)$ telle que :

$$\log \frac{p_\beta(x)}{1 - p_\beta(x)} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = x' \beta, \quad (1.1)$$

ou encore

$$\text{logit } p_\beta(x) = x' \beta,$$

logit désignant la fonction bijective et dérivable de $]0, 1[$ dans \mathbb{R} : $p \mapsto \log(p/(1 - p))$.

L'égalité (1.1) peut également s'écrire

$$p_\beta(x) = \mathbf{P}_\beta(Y = 1|X = x) = \frac{\exp(x' \beta)}{1 + \exp(x' \beta)}.$$

Remarque

Dans un modèle logistique, nous effectuons deux choix pour définir le modèle :

1. le choix d'une loi pour $Y|X = x$, ici la loi de Bernoulli ;
2. le choix de la modélisation de $\mathbf{P}(Y = 1|X = x)$ par

$$\text{logit } \mathbf{P}_\beta(Y = 1|X = x) = x' \beta.$$

La fonction logit est bijective et dérivable. Elle est appelée *fonction de lien*.

Remarquons également que

$$\begin{cases} \mathbf{E}_\beta[Y|X = x] = \mathbf{P}_\beta(Y = 1|X = x) \\ \mathbf{V}_\beta(Y|X = x) = \mathbf{P}_\beta(Y = 1|X = x) \left(1 - \mathbf{P}_\beta(Y = 1|X = x)\right) \end{cases}$$

ce qui implique que la variance n'est pas constante et varie selon x .

1.2.2 La régression log-linéaire

Dans le modèle logistique la variable à expliquer est une variable binaire. Le modèle log-linéaire traite le cas d'une *variable de comptage*. Voici quelques exemples :

- nombre de catastrophes aériennes sur une période donnée ;
- nombre de voitures à un feu rouge ;
- nombre d'accidents par jour sur une autoroute...

Définition 1.2 (Régression log-linéaire)

Soit Y une variable de comptage à expliquer par le vecteur $X = (1, \mathbf{X}_1, \dots, \mathbf{X}_p)'$. Le modèle log-linéaire propose une modélisation de la loi de $Y|X = x$ par une loi de poisson de paramètre $\lambda = \lambda(x)$ telle que :

$$\log \mathbf{E}[Y|X = x] = x'\beta.$$

Pour une nouvelle mesure x effectuée, le modèle log-linéaire va donc prédire la valeur $\exp(x'\beta)$.

Remarque

Ici encore, deux choix sont effectués pour définir le modèle :

1. le choix d'une loi pour $Y|X = x$, ici la loi de Poisson ;
2. le choix de la modélisation de $\mathbf{E}(Y|X = x)$ par

$$\log \mathbf{E}[Y|X = x] = x'\beta.$$

La fonction log est bijective et dérivable.

1.2.3 Généralisation : GLM

On peut résumer les remarques précédentes par le tableau :

Choix	logistique	log-linéaire	linéaire
loi de $Y X = x$	Bernoulli	Poisson	Normale
modélisation de $\mathbf{E}[Y X = x]$	$\text{logit } \mathbf{E}[Y X = x] = x'\beta$	$\log \mathbf{E}[Y X = x] = x'\beta$	$\mathbf{E}[Y X = x] = x'\beta$

Une généralisation de ces méthodes est appelée GLM (Generalized Linear Model). L'approche GLM consiste à :

1. choisir une loi pour $Y|X = x$ parmi un ensemble restreint de loi (les lois exponentielles GLM) ;
2. choisir une *fonction de lien* $g(\cdot)$ parmi une ensemble réduit de fonctions bijectives et dérivables.
3. la transformation de l'espérance conditionnelle $\mathbf{E}[Y|X = x]$ par la fonction g est ensuite modélisée par une fonction η qui n'est autre qu'une combinaison linéaire des variables explicatives :

$$g(\mathbf{E}[Y|X = x]) = \eta(x) = x'\beta.$$

On peut résumer un modèle GLM par le schéma suivant :

A expliquer composante aléatoire	Lien	Explicatif Composante systématique
$Y X = x$ suit une loi fixée.	$\mathbf{E}[Y X = x]$ dépend de $\eta(x)$ au travers de la fonction g appelée fonction de lien	On modélise η par une combinaison linéaire des X_j
	$g(\mathbf{E}[Y X]) = \eta(X)$	$\eta(x) = \sum_{j=0}^p x_j \beta_j$
	g est une fonction <i>inversible</i> .	

Remarque

1. Pour choisir un modèle GLM il faut
 - choisir la loi de $Y|X = x$ dans la famille exponentielle des GLM.
 - choisir une fonction de lien inversible g .
2. Pour utiliser un modèle GLM il faudra donc estimer les paramètres $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$. Une fois cette estimation réalisée, on disposera d'une estimation de $\eta(x)$ est fixé ainsi que de $\mathbf{E}[Y|X = x] = g^{-1}(\eta(x))$.

Le tableau 1.2 donne quelques exemples de GLM.

Loi	Nom du lien	Fonction de lien
Bernoulli/Binomiale	lien logit	$g(\mu) = \text{logit}(\mu) = \log(\mu/(1-\mu))$
Poisson	lien log	$g(\mu) = \log(\mu)$
Normale	lien identité	$g(\mu) = \mu$
Gamma	lien réciproque	$g(\mu) = -1/\mu$

TABLE 1.2 – Exemples de GLM.

1.3 Exemples de fonctions de liens pour la régression d'une variable binaire

D'autres fonctions de lien que `logit` peuvent être utilisées dans le cas où la variable à expliquer Y est binaire. On trouve notamment dans la littérature les transformations :

- `probit`, qui n'est autre que l'inverse de la fonction de répartition de la loi normale centrée réduite :

$$\forall p \in [0, 1], \text{probit}(p) = \epsilon \quad \text{avec} \quad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\epsilon} \exp\left(-\frac{1}{2}u^2\right) du = p.$$

- `log-log` définie par :

$$\forall p \in [0, 1], \text{log-log}(p) = \log(-\log(1-p)).$$

Ces transformations sont représentées sur la figure 1.5.

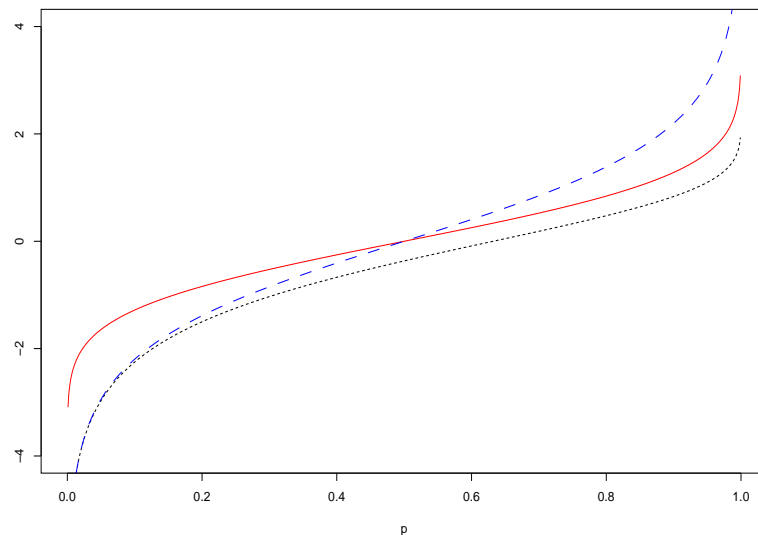


FIGURE 1.5 – Fonctions de liens : probit (trait plein), logit (tirets), log-log (pointillés).

Des trois fonctions de lien présentées, la transformation log-log est bien appropriée aux cas où l'on souhaite modéliser les probabilités de succès de manière asymétrique. Les transformations logit et probit possèdent des propriétés identiques. Dans de nombreux cas, on préfère utiliser la transformation logistique. Plusieurs raisons motivent ce choix :

- d'un point de vue numérique, la transformation logistique est plus simple à manipuler (notamment pour l'écriture des estimateurs du maximum de vraisemblance, voir section 2.2) ;
- on a une interprétation claire des coefficients en terme d'odds ratio pour la transformation logistique (voir section 2.5).
- le modèle logistique est particulièrement bien adapté à un schéma d'échantillonnage rétrospectif (voir section 2.7).

Nous nous focaliserons dans la suite sur le modèle logistique. Les différents résultats obtenus pourront s'étendre aux autres modèles GLM. Il est important de connaître les notations des GLM présentées dans cette partie. C'est en effet sous cette forme qu'elles sont présentées dans la littérature ainsi que dans la plupart des logiciels statistiques (par exemple sur R).

Chapitre 2

Analyse discriminante logistique

Nous rappelons que Y désigne une variable à expliquer binaire (qui prend 2 valeurs 0 ou 1 pour simplifier) ou un label qui dénote l'appartenance à un groupe et $\mathbf{X}_1, \dots, \mathbf{X}_p$ désignent p variables explicatives. On souhaite :

- expliquer la variable Y à l'aide des p variables explicatives $X = (1, \mathbf{X}_1, \dots, \mathbf{X}_p)'$;
- étant donnée une nouvelle mesure x des p variables explicatives, prédire le label y associé à cette nouvelle observation.

Nous avons vu dans le chapitre précédent que le modèle logistique est défini par

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = x' \beta \quad (2.1)$$

où $\beta = (\beta_0, \dots, \beta_p)'$ et $x = (1, \mathbf{x}_1, \dots, \mathbf{x}_p)'$. Nous nous posons maintenant le problème de l'estimation des paramètres β à partir d'un échantillon $(x_1, y_1), \dots, (x_n, y_n)$.

2.1 Le modèle statistique

Les paramètres du modèle logistique sont généralement estimés par la méthode du maximum de vraisemblance (voir annexe A.1). Afin d'écrire "proprement" la vraisemblance, il convient de définir avec précision le modèle statistique dans lequel nous allons nous placer. On rappelle qu'un modèle statistique est un couple $(\mathcal{H}^n, \mathcal{P})$ où \mathcal{H} est l'espace de chaque observation et \mathcal{P} est une famille de lois de probabilité sur \mathcal{H}^n muni de sa tribu borélienne.

2.1.1 L'échantillonnage

On note $(x_1, y_1), \dots, (x_n, y_n)$ une réalisation d'un n -échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. et de même loi que (\mathbf{X}, Y) . Sans perte de généralité, nous supposons que l'échantillonnage a été réalisé en deux temps :

- les n observations x_1, \dots, x_n ont été générés de manière indépendante selon une loi \mathbf{P}_X ;
- pour $i = 1, \dots, n$, y_i est généré selon une loi de Bernoulli de paramètre $p_\beta(x_i)$ tel que $\text{logit } p_\beta(x_i) = x_i' \beta$ (β étant le paramètre à estimer).

Nous distinguerons deux structures de données. On parlera de

- données *individuelles* lorsque tous les x_i sont différents. On appellera alors design l'ensemble $\{x_1, \dots, x_n\}$. Le modèle est alors défini par

$$\mathcal{M}_n = \{ \{0, 1\}^n, \{B(p_\beta(x_1)) \otimes \dots \otimes B(p_\beta(x_n)), \beta \in \mathbb{R}^{p+1} \} \}.$$

- données *répétées* lorsque certaines valeurs de x_i sont répétées plusieurs fois (ce qui peut se produire si par exemple \mathbf{P}_X est discrète à support borné). On note
 - x_1, \dots, x_T les différentes valeurs observées ;
 - $n_t, t = 1, \dots, T$ le nombre de fois où x_t a été observé (on a donc $\sum_{t=1}^T n_t = n$) ;
 - $y_t, t = 1, \dots, T$ le nombre de succès ($Y = 1$) observé au point x_t .
- On appellera design l'ensemble $\{(x_1, n_1), \dots, (x_T, n_T)\}$. Pour cette structure de données, le modèle sera défini par

$$\mathcal{M}_n = \left\{ \{0, \dots, n_1\} \times \dots \times \{0, \dots, n_T\}, \{Bin(n_1, p_\beta(x_1)) \otimes \dots \otimes Bin(n_T, p_\beta(x_T)), \beta \in \mathbb{R}^{p+1}\} \right\}.$$

Exemple 2.1

Dans le cadre d'un essai clinique, on teste 2 traitements (A et B) sur $n = 220$ patients atteints d'une pathologie. On reporte dans le tableau 2.1 le nombre de patients qui ont guéri au bout de 3 mois de traitements.

	Guérison	Non Guérison
A	40	60
B	90	30

TABLE 2.1 – Exemple de données répétées.

Nous sommes ici en présence de données répétées. Un individu (x, y) est un patient (x représente le traitement subi et y vaut 1 si le patient a guéri au bout de 3 mois, 0 sinon). Le design est $\{(A, 100), (B, 120)\}$ et on a $y_1 = 40$ et $y_2 = 90$.

Les propriétés des estimateurs sont très proches pour ces deux types de données. Néanmoins, certains concepts tels que la forme de la vraisemblance où les tests d'adéquation par la déviance peuvent légèrement différer. Dans ce chapitre, nous nous focalisons sur le cas de données individuelles (qui est le cas le plus fréquent). Pour une étude plus approfondie du cas des données répétées, nous renvoyons le lecteur à l'annexe A.3 (pour l'écriture de la vraisemblance) ou aux ouvrages de Hosmer & Lemeshow (2000) et Collet (2003).

2.1.2 Identifiabilité du modèle

On rappelle qu'un modèle $\{\mathcal{H}, \{\mathbf{P}_\theta, \theta \in \Theta\}\}$ est identifiable si $\theta \mapsto \mathbf{P}_\theta$ est injective. Dit brutalement, un modèle est dit identifiable si deux paramètres différents définissent deux lois différentes.

Exemple 2.2

Plaçons nous dans le cadre de la régression d'une variable binaire par une variable explicative à deux modalités A et B . On considère le modèle logistique

$$\mathcal{M}_1 = \left\{ \{0, 1\}, \{B(p_\beta(x)), \beta \in \mathbb{R}^3\} \right\}$$

où

$$\begin{aligned} \text{logit } p_\beta(x) &= \beta_0 + \beta_1 \mathbf{1}_{x=A} + \beta_2 \mathbf{1}_{x=B} \\ &= (\beta_0 + \beta_1) + 0 \mathbf{1}_{x=A} + (\beta_2 - \beta_1) \mathbf{1}_{x=B} \end{aligned}$$

Si on pose $\tilde{\beta} = (\beta_0 - \beta_1, 0, \beta_2 - \beta_1)$ on a $\beta \neq \tilde{\beta}$ et $p_\beta(x) = p_{\tilde{\beta}}(x)$. \mathcal{M}_1 n'est donc pas identifiable. Une solution classique consiste à mettre une contrainte sur les coefficients. Si par exemple, on pose $\beta_2 = 0$ alors le modèle

$$\mathcal{M}'_1 = \left\{ \{0, 1\}, \{B(p_\beta(x)), \beta \in \mathbb{R}^2\} \right\}$$

où $\text{logit } p_\beta(x) = \beta_0 + \beta_1 \mathbf{1}_{x=A}$ est identifiable.

Proposition 2.1

Le modèle \mathcal{M}_n est identifiable si et seulement si $\text{rang}(\mathbf{X}) = p + 1$.

Preuve

Soit $\beta \neq \tilde{\beta}$. Par définition \mathcal{M}_n est identifiable si et seulement si il existe $i \leq n$ tel que $p_\beta(x_i) \neq p_{\tilde{\beta}}(x_i)$. Supposons que : $\forall 1 \leq i \leq n, x'_i \beta = x'_i \tilde{\beta}$. On a alors

$$\alpha_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \alpha_1 \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} + \dots + \alpha_p \begin{pmatrix} x_{1p} \\ \vdots \\ x_{np} \end{pmatrix} = 0$$

avec $\alpha = \beta - \tilde{\beta} \neq 0$. Une colonne de \mathbf{X} est donc combinaison linéaire des autres, d'où $\text{rang}(\mathbf{X}) < p + 1$.

2.2 L'estimateur du maximum de vraisemblance

2.2.1 La vraisemblance

On se place dans le cas de données individuelles. La vraisemblance du modèle (identifiable) \mathcal{M}_n est définie par :

$$L_n : \{0, 1\}^n \times \mathbb{R}^{p+1} \rightarrow \mathbb{R}^+ \\ (y_1, \dots, y_n, \beta) \mapsto B(p_\beta(x_1)) \otimes \dots \otimes B(p_\beta(x_n))(\{y_1, \dots, y_n\}).$$

Remarque

Si on désigne par Y_i une variable aléatoire de loi $B(p_\beta(x_i))$ alors les variables Y_1, \dots, Y_n sont indépendantes mais pas de même loi.

Lorsqu'il n'y aura pas de confusion possible, on commettra l'abus de notation $L_n(y_1, \dots, y_n, \beta) = L_n(\beta)$. Calculons la vraisemblance. On a

$$L_n(\beta) = \prod_{i=1}^n \mathbf{P}_\beta(Y = y_i | X = x_i) = \prod_{i=1}^n p_\beta(x_i)^{y_i} (1 - p_\beta(x_i))^{1-y_i}.$$

En passant au log nous obtenons

$$\begin{aligned} \mathcal{L}_n(\beta) &= \sum_{i=1}^n \{y_i \log(p_\beta(x_i)) + (1 - y_i) \log(1 - p_\beta(x_i))\} \\ &= \sum_{i=1}^n \left\{ y_i \log \left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)} \right) + \log(1 - p_\beta(x_i)) \right\}, \end{aligned}$$

ou encore, d'après (2.1),

$$\mathcal{L}_n(\beta) = \sum_{i=1}^n \{y_i x'_i \beta - \log(1 + \exp(x'_i \beta))\}. \quad (2.2)$$

Le vecteur gradient au point β défini par $\nabla \mathcal{L}_n(\beta) = \left[\frac{\partial \mathcal{L}}{\partial \beta_0}(\beta), \dots, \frac{\partial \mathcal{L}}{\partial \beta_p}(\beta) \right]'$ s'obtient par dérivation

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta_j}(\beta) &= \sum_{i=1}^n \left[y_i x_{ij} - \frac{x_{ij} \exp(x'_i \beta)}{1 + \exp(x'_i \beta)} \right] \\ &= \sum_{i=1}^n [x_{ij}(y_i - p_\beta(x_i))]. \end{aligned}$$

Ce qui donne en écriture matricielle

$$\nabla \mathcal{L}_n(\beta) = \sum_{i=1}^n [x_i(y_i - p_\beta(x_i))] = \mathbf{X}'(\mathbb{Y} - P_\beta)$$

où $\mathbb{Y} = (y_1, \dots, y_n)'$ et $P_\beta = (p_\beta(x_1), \dots, p_\beta(x_n))'$. L'estimateur du maximum de vraisemblance (si il existe) est solution de l'équation (appelée équation du score) :

$$S(\beta) = \nabla \mathcal{L}_n(\beta) = \mathbf{X}'(\mathbb{Y} - P_\beta) = 0. \quad (2.3)$$

On rappelle que si cette équation admet une solution en β notée $g(y_1, \dots, y_n)$ (et que cette solution est un maximum de $\mathcal{L}_n(\beta)$) alors l'estimateur du maximum de vraisemblance est $\hat{\beta} = g(Y_1, \dots, Y_n)$.

Trouver explicitement $\hat{\beta}$ n'est pas possible. En effet, l'équation (2.3) se réécrit :

$$\begin{cases} x_{11}y_1 + \dots + x_{n1}y_n = x_{11} \frac{\exp(\beta_1 x_{11} + \dots + \beta_p x_{1p})}{1 + \exp(\beta_1 x_{11} + \dots + \beta_p x_{1p})} + \dots + x_{n1} \frac{\exp(\beta_1 x_{n1} + \dots + \beta_p x_{np})}{1 + \exp(\beta_1 x_{n1} + \dots + \beta_p x_{np})} \\ \vdots \\ \vdots \\ x_{1p}y_1 + \dots + x_{np}y_n = x_{1p} \frac{\exp(\beta_1 x_{11} + \dots + \beta_p x_{1p})}{1 + \exp(\beta_1 x_{11} + \dots + \beta_p x_{1p})} + \dots + x_{np} \frac{\exp(\beta_1 x_{n1} + \dots + \beta_p x_{np})}{1 + \exp(\beta_1 x_{n1} + \dots + \beta_p x_{np})}. \end{cases}$$

Ce système (qui n'est pas linéaire en β) n'admet pas de solution analytique. On a donc recours à des méthodes numériques qui nécessite de connaître d'éventuelles propriétés sur la régularité de la fonction à optimiser (en terme de convexité par exemple).

2.2.2 Comportement asymptotique de l'estimateur du maximum de vraisemblance

Définition 2.1

Le nuage de points est dit :

- complètement séparable si $\exists \beta \in \mathbb{R}^{p+1} : \forall i$ tel que $Y_i = 1$ on a $x'_i \beta > 0$ et $\forall i$ tel que $Y_i = 0$ on a $x'_i \beta < 0$;
- quasi-complètement séparable si $\exists \beta \in \mathbb{R}^{p+1} : \forall i$ tel que $Y_i = 1$ on a $x'_i \beta \geq 0$, $\forall i$ tel que $Y_i = 0$ on a $x'_i \beta \leq 0$ et $\{i : x'_i \beta = 0\} \neq \emptyset$;
- en recouvrement s'il n'est ni complètement séparable ni quasi-complètement séparable (voir figure 2.1).

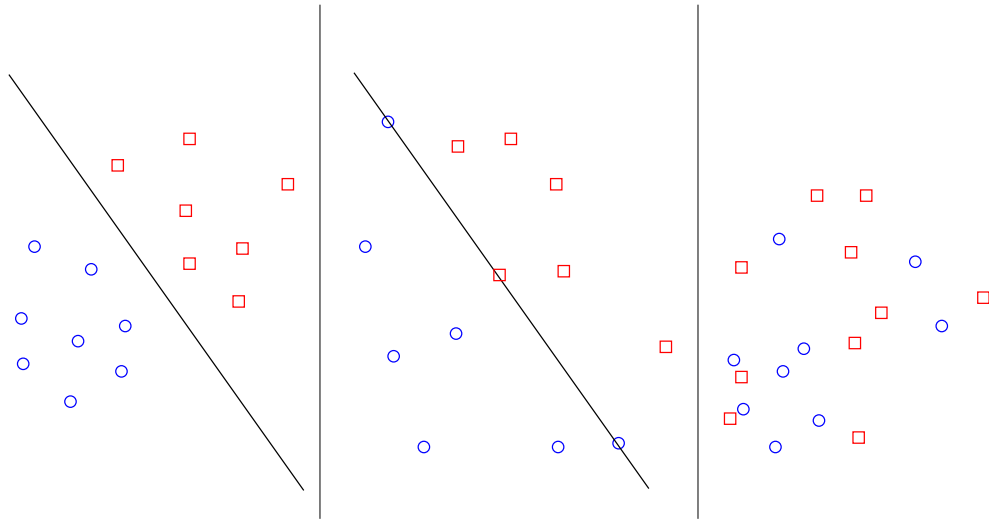


FIGURE 2.1 – Exemple de séparabilité complète (gauche), quasi-complète (milieu) et de recouvrement (droite).

Le théorème suivant nous assure la convergence d'un algorithme itératif vers la vers $\hat{\beta}$ et nous donne le comportement asymptotique de l'estimateur du maximum de vraisemblance. Considérons le jeu d'hypothèses suivant :

- H_1 : $\text{rang}(\mathbf{X}) = p + 1$.
- H_2 : on est en situation de recouvrement.
- H_3 : les x_i sont des réalisations i.i.d. d'une loi à support compact.
- H_4 : la plus petite valeur propre de la matrice $\mathbf{X}'\mathbf{X}$ tend vers l'infini quand $n \rightarrow \infty$.

Théorème 2.1

1. Sous les hypothèses H_1 et H_2 la log-vraisemblance $\beta \rightarrow \mathcal{L}_n(\beta)$ est strictement concave : $\hat{\beta}$ existe et est unique.
2. Sous les hypothèses H_3 et H_4 on a
 - (a) la convergence en probabilité de $\hat{\beta}$ vers β quand $n \rightarrow \infty$.
 - (b) la loi asymptotique de l'estimateur du maximum de vraisemblance

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow \mathcal{N}(0, \mathcal{I}(\beta)^{-1}),$$

où $\mathcal{I}(\beta)$ est la matrice d'information de Fisher (de dimension $(p+1) \times (p+1)$) au point β :

$$\mathcal{I}(\beta)_{ij} = -\mathbf{E} \left[\frac{\partial^2}{\partial \beta_i \partial \beta_j} \mathcal{L}_1(Y, \beta) \right], \quad 0 \leq i, j \leq p,$$

où avec un léger abus de notation $\mathcal{I}(\beta)_{ij}$ désigne le terme de la $(i+1)^{\text{ème}}$ ligne et $(j+1)^{\text{ème}}$ colonne de $\mathcal{I}(\beta)$.

Pour la preuve de la concavité, on pourra se référer au polycopié de Guyon (2005) ou à l'article de Albert & Anderson (1984). La loi asymptotique découle de la théorie du maximum de vraisemblance (voir annexe A.1 et Antoniadis *et al* (1992) pour une preuve détaillée). La concavité a une conséquence numérique importante puisqu'elle justifie qu'un algorithme itératif convergera bien vers la valeur de $\hat{\beta}$. Il n'y a donc pas de risque de converger vers un maximum local non global et la convergence de l'algorithme ne dépend pas du point d'initialisation de l'algorithme.

2.3 L'algorithme IRLS

Deux algorithmes sont généralement implémentés sur les logiciels de statistique pour calculer les estimateurs du maximum de vraisemblance : l'algorithme du score de Fisher et l'algorithme IRLS (Iterative Reweighted Least Square). Nous présentons ici uniquement le second algorithme.

2.3.1 Rappel sur l'algorithme de Newton-Raphson

La méthode de Newton-Raphson permet une résolution numérique des équations du score. Pour simplifier les notations, nous supposons que β est univarié. On part tout d'abord d'une valeur initiale arbitraire de β , notée β^0 et on désigne par $\beta^1 = \beta^0 + h$ une valeur candidate pour être solution de $S(\beta) = 0$, c'est-à-dire $S(\beta^0 + h) = 0$. Par un développement limité à l'ordre un de la fonction S , on obtient l'approximation suivante :

$$S(\beta^0 + h) \simeq S(\beta^0) + hS'(\beta^0).$$

Comme $S(\beta^0 + h) = 0$, on obtient pour h la valeur suivante :

$$h = -[S'(\beta^0)]^{-1} S(\beta^0)$$

et donc

$$\beta^1 = \beta^0 - [S'(\beta^0)]^{-1} S(\beta^0).$$

Dans le cas qui nous concerne $\beta \in \mathbb{R}^{p+1}$ et $S(\beta) = \nabla \mathcal{L}_n(\beta)$. La formule de récurrence se traduit par

$$\beta^1 = \beta^0 - [\nabla^2 \mathcal{L}_n(\beta^0)]^{-1} \nabla \mathcal{L}_n(\beta^0)$$

où $\nabla^2 \mathcal{L}_n(\beta^0)$ désigne la matrice hessienne de la log-vraisemblance au point β^0 :

$$\nabla^2 \mathcal{L}_n(\beta^0)_{ij} = \left[\frac{\partial^2 \mathcal{L}}{\partial \beta_i \partial \beta_j}(\beta^0) \right], \quad 0 \leq i, j \leq p$$

où nous comettons toujours l'abus de désigner par $\nabla^2 \mathcal{L}_n(\beta^0)_{ij}$ le terme de la $(i+1)^{\text{ème}}$ ligne et $(j+1)^{\text{ème}}$ colonne de $\nabla^2 \mathcal{L}_n(\beta^0)$. Le processus est ensuite itéré jusqu'à convergence. Il se résume de la manière suivante :

1. choix d'un point de départ β^0 ;
2. On construit β^{k+1} à partir de β^k

$$\beta^{k+1} = \beta^k + A^k \nabla \mathcal{L}_n(\beta^k),$$

où $\nabla \mathcal{L}_n(\beta^k)$ est le gradient au point β^k et $A^k = -(\nabla^2 \mathcal{L}_n(\beta^k))^{-1}$ est la matrice de "pas" de l'algorithme (l'inverse du hessien de \mathcal{L}_n au point β^k).

2.3.2 Calcul des estimateurs

Calculons la matrice hessienne $\nabla^2 \mathcal{L}_n(\beta) = \left\{ \frac{\partial^2 \mathcal{L}_n(\beta)}{\partial \beta_r \partial \beta_s} \right\}_{0 \leq r, s \leq p}$:

$$\frac{\partial^2 \mathcal{L}_n}{\partial \beta_r \partial \beta_s}(\beta) = - \sum_{i=1}^n x_i^r x_i^s \frac{\exp(x_i' \beta)}{(1 + \exp(x_i' \beta))^2} = - \sum_{i=1}^n x_i^r x_i^s p_\beta(x_i) (1 - p_\beta(x_i)),$$

Algorithme 1 maximisation de la vraisemblance**Require:** β^0 $k \leftarrow 1$ **repeat** $\beta^{k+1} \leftarrow \beta^k + A^k \nabla \mathcal{L}_n(\beta^k)$ $k \leftarrow k + 1$ **until** $\beta^{k+1} \approx \beta^k$ et/ou $\mathcal{L}_n(\beta^{k+1}) \approx \mathcal{L}_n(\beta^k)$

en écriture matricielle nous obtenons

$$\nabla^2 \mathcal{L}_n(\beta) = - \sum_{i=1}^n x_i x_i' p_\beta(x_i) (1 - p_\beta(x_i)) = \mathbf{X}' W_\beta \mathbf{X},$$

où W_β est la matrice diagonale $\text{diag}(p_\beta(x_i)(1 - p_\beta(x_i)), i = 1, \dots, n)$. Nous pouvons maintenant exprimer β^{k+1} en fonction de β^k :

$$\begin{aligned} \beta^{k+1} &= \beta^k + (\mathbf{X}' W_{\beta^k} \mathbf{X})^{-1} \mathbf{X}' (\mathbb{Y} - P_{\beta^k}) \\ &= (\mathbf{X}' W_{\beta^k} \mathbf{X})^{-1} \mathbf{X}' W_{\beta^k} (\mathbf{X} \beta^k + W_{\beta^k}^{-1} (\mathbb{Y} - P_{\beta^k})) \\ &= (\mathbf{X}' W_{\beta^k} \mathbf{X})^{-1} \mathbf{X}' W_{\beta^k} Z^k, \end{aligned}$$

où $Z^k = \mathbf{X} \beta^k + W_{\beta^k}^{-1} (\mathbb{Y} - P_{\beta^k})$. Cette équation est simplement une régression pondérée du vecteur Z^k où les poids W_{β^k} dépendent de \mathbf{X} et β^k ; d'où le nom de IRLS pour cette méthode. Les poids sont réévalués à chaque étape de l'algorithme, une étape étant une simple régression pondérée.

2.4 Dimensions explicatives, variables explicatives

Les remarques formulées dans cette partie s'appliquent pour la plupart des modèles de régression (modèles linéaires et d'analyse de variance par exemple). Pour plus de détails, on pourra se rapporter aux ouvrages de Dreesbeke *et al* (2007) et Cornillon & Matzner-Løber (2007).

On rappelle que la dimension d'un modèle paramétrique (identifiable) $\{\mathcal{H}, \{\mathbf{P}_\theta, \theta \in \Theta\}\}$ est la dimension de l'espace des paramètres Θ . Pour le modèle logistique, les lois sont déterminées par

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

Tout comme pour le modèle linéaire, la dimension d'un modèle logistique s'obtient en sommant les dimensions explicatives associées aux différentes variables explicatives du modèle, lesquelles varient suivant la nature de la variable explicative. Nous étudions dans cette partie les dimensions explicatives pour des variables explicatives quantitatives et qualitatives. Nous étudierons également les cas d'interactions entre variables.

2.4.1 Variable explicative quantitative

Si on dispose d'une seule variable explicative X quantitative (non regroupée en classe) le modèle s'écrit

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 x.$$

Un seul coefficient (β_1) est alloué à X , cette variable est représentée par une seule colonne dans la matrice du design \mathbf{X} . Sa dimension est donc égale à 1.

2.4.2 Variable explicative qualitative

Tout comme pour le modèle d'analyse de variance, une variable qualitative est représentée par les indicatrices associées aux différentes modalités. Considérons un modèle où la seule variable explicative est le `sexe` :

$$\text{logit } p_\alpha(x) = \alpha_0 + \alpha_F \mathbf{1}_F(x) + \alpha_H \mathbf{1}_H(x), \quad (2.4)$$

mais aussi

$$\text{logit } p_\alpha(x) = (\alpha_0 + \alpha_F) + (\alpha_H - \alpha_F) \mathbf{1}_H(x).$$

Il y a une infinité d'écritures possibles... Le modèle (2.4) correspond à une matrice du design \mathbf{X} à trois colonnes où la première colonne est une colonne de 1 et les deux dernières sont obtenues en effectuant un codage disjonctif complet pour chaque individu (le $i^{\text{ème}}$ terme de la 2^{ème} (resp. 3^{ème}) colonne vaut 1 si le $i^{\text{ème}}$ individu de l'échantillon est une femme (resp. un homme)). Par conséquent, la somme des deuxième et troisième colonne vaut 1 ce qui rend l'estimation impossible puisque la matrice \mathbf{X} n'est pas de plein rang ($\mathbf{X}'W_\beta\mathbf{X}$ n'est pas inversible et le modèle n'est pas identifiable). Une solution pour pallier à cette difficulté consiste à mettre une contrainte sur les coefficients α_H et α_F . La solution souvent utilisée par les logiciels est de supprimer une des colonnes de la matrice \mathbf{X} , ce qui revient à considérer que le coefficient de la modalité associée à cette colonne est nul. Cette modalité est prise comme modalité de référence par rapport à laquelle on mesure des déviations. Le choix de cette modalité n'a bien entendu pas d'influence sur les lois \mathbf{P}_β . Il en a cependant une sur la valeur des coefficients estimés ainsi que sur leurs écarts types. Ainsi le nombre de coefficients significativement différents de 0 peut changer suivant le choix de la modalité de référence. Ceci montre clairement que, pour juger l'apport d'une variable qualitative, il n'est pas pertinent d'utiliser les tests de significativité des coefficients. Il sera préférable de réaliser un test entre modèles emboîtés (voir page 38).

Exemple 2.3

Considérons le cas d'une variable explicative à trois niveaux g_1, g_2, g_3 . Les observations sont récoltées dans les tableaux suivants (équivalents)

observation	X	Y
1	g_1	1
2	g_2	1
3	g_3	1
4	g_1	1
5	g_2	0
6	g_1	0

X	$\#\{Y = 1\}$	$\#\{Y = 0\}$
g_1	2	1
g_2	1	1
g_3	1	0

On effectue une régression logistique sur R :

```
> X <- factor(c("g1", "g2", "g3", "g1", "g2", "g1"))
> Y <- factor(c(1, 1, 1, 1, 0, 0))
> model <- glm(Y~X, family=binomial)
> model
```

```
Call: glm(formula = Y ~ X, family = binomial)
```

Coefficients:

```
(Intercept)      Xg2      Xg3
    0.6931     -0.6931     17.8729
```

Degrees of Freedom: 5 Total (i.e. Null); 3 Residual
 Null Deviance: 7.638
 Residual Deviance: 6.592 AIC: 12.59

La modalité g_1 est ici prise comme modalité de référence. Le modèle estimé s'écrit donc :

$$\text{logit } \hat{p}(g_j) = \text{logit } p_{\hat{\beta}}(g_j) = \begin{cases} 0.6931 & \text{si } j = 1 \\ 0 & \text{si } j = 2 \\ 0.6931 + 17.8729 = 18.566 & \text{si } j = 3. \end{cases}$$

ou encore

$$\hat{p}(g_j) = \begin{cases} \frac{\exp(0.6931)}{1+\exp(0.6931)} = 2/3 & \text{si } j = 1 \\ 1/2 & \text{si } j = 2 \\ \frac{\exp(18.566)}{1+\exp(18.566)} = 1.0000 & \text{si } j = 3. \end{cases} \quad (2.5)$$

Bien évidemment, changer la contrainte modifie l'écriture du modèle mais ne modifie la loi de probabilité sous-jacente. Prenons par exemple, la modalité g_3 comme modalité de référence :

```
> model1 <- glm(Y~C(X,base=3),family=binomial)
> model1
```

Call: glm(formula = Y ~ C(X, base = 3), family = binomial)

Coefficients:

(Intercept)	C(X, base = 3)1	C(X, base = 3)2
18.57	-17.87	-18.57

Degrees of Freedom: 5 Total (i.e. Null); 3 Residual
 Null Deviance: 7.638
 Residual Deviance: 6.592 AIC: 12.59

Il est par exemple facile de vérifier que les probabilités $\hat{p}(g_j)$, $j = 1, 2, 3$ sont identiques à celles de (2.5).

2.4.3 Interactions

Tout comme en analyse de la variance, on ne peut se contenter de modèles purement additifs. Reprenons l'exemple développé dans Dreesbeke *et al* (2007) (page 122). Nous considérons le cas où la variable Y représente le fait de faire (codé 1) ou non (codé 0) de la couture. On dispose de deux variables explicatives : l'age et le **sexe**. Le modèle "purement" additif s'écrit :

$$\text{logit } p_{\beta}(x) = \beta_0 + \beta_1 \text{age} + \beta_2 \mathbf{1}_{\text{femme}},$$

la modalité **homme** a été choisie comme modalité de référence. Une telle écriture revient à supposer que les pentes sont identiques pour les hommes et les femmes (voir Figure 2.2).

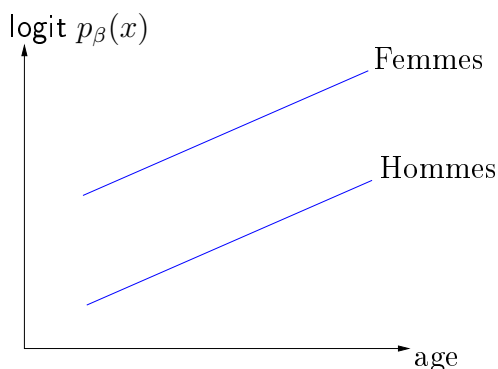


FIGURE 2.2 – Modèle additif.

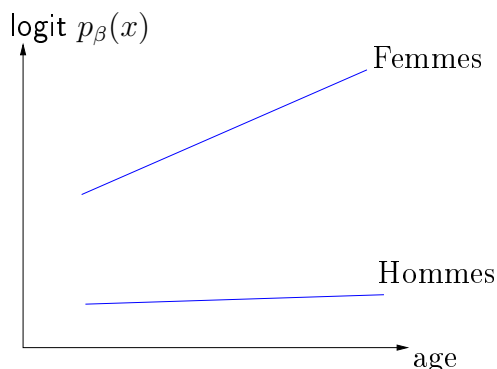


FIGURE 2.3 – Modèle avec interaction.

A priori, il semble que les hommes ne font que très rarement de la couture quel que soit leur âge, il paraît préférable de pouvoir utiliser un modèle du genre (voir Figure 2.3) :

$$\text{logit } p(x) = \beta_0 + \beta_1 \text{age} + \beta_2 \mathbf{1}_{\text{femme}} + \beta_3 \text{age} \mathbf{1}_{\text{femme}}.$$

Ce modèle revient à considérer l'interaction entre les variables **age** et **sexe**. On rappelle que deux variables interagissent si l'effet de l'une sur Y diffère suivant les valeurs de l'autre. Bien entendu, l'ajout d'une interaction augmente la dimension explicative du modèle. Le nombre de composantes supplémentaires s'obtient en faisant le produit du nombre de dimensions des variables qui interagissent (ici les variables **sexe** et **age** sont de dimension 1, on rajoute donc une dimension).

2.5 Interprétation des coefficients β

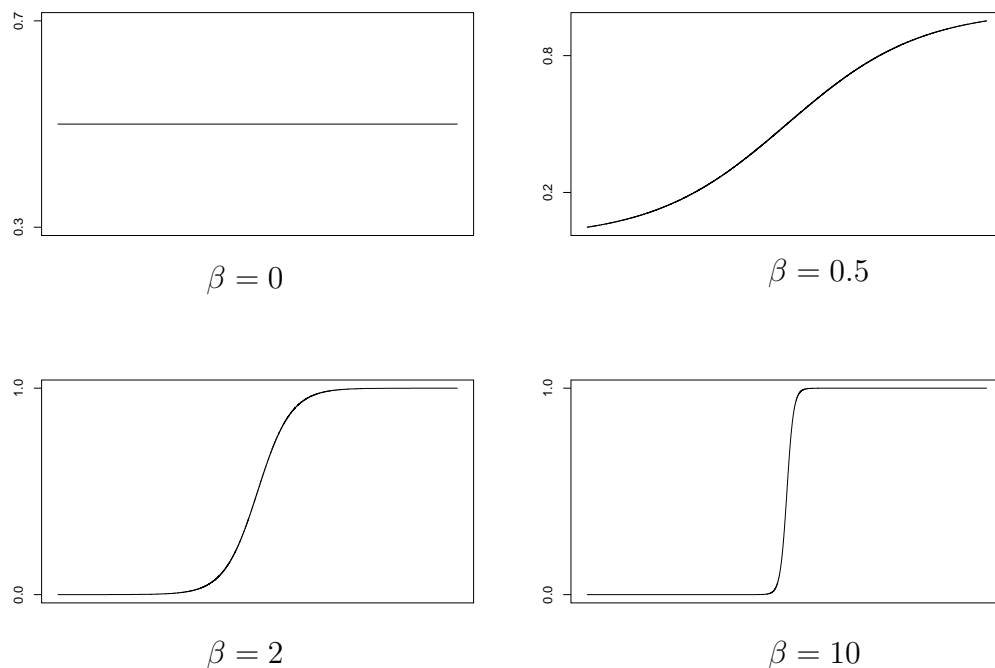


FIGURE 2.4 – $\mathbf{P}_\beta(Y = 1|X = x)$ pour différentes valeurs de β .

Nous avons représenté sur la Figure 2.4 l'allure de la courbe représentative de la fonction $x \mapsto \frac{\exp(x\beta)}{1+\exp(x\beta)}$ pour différentes valeurs du paramètre β . On remarque que :

- pour de faibles valeurs de β on a une large plage de valeurs de x pour lesquelles la fonction se situe aux alentours de 0.5 (la fonction est même constante (0.5) dans le cas extrême $\beta = 0$). Pour ces valeurs $p_\beta(x) = \mathbf{P}_\beta(Y = 1|X = x)$ sera proche de 0.5 et on peut donc penser qu'il sera difficile de discriminer ;
- lorsque β augmente, la zone où la fonction est proche de 0.5 diminue et la fonction est proche de 0 ou 1 pour un grand nombre de valeurs de x . Par conséquent, $\mathbf{P}_\beta(Y = 1|X = x)$ sera souvent proche de 1 ou 0, ce qui risque de minimiser d'éventuelles erreurs de prévisions.

On peut interpréter ainsi : *plus β est grand, mieux on discrimine*. Cependant une telle interprétation dépend des valeurs que x prend (de son échelle). C'est pourquoi en général l'interprétation des coefficients β s'effectue en terme d'*odds ratio*. Les odds ratio sont des outils souvent appréciés dans le domaine de l'épidémiologie (mais pas toujours bien utilisés!).

Les odds ratio servent à mesurer l'effet d'une variable quantitative ou le contraste entre les effets d'une variable qualitative. L'idée générale est de raisonner en terme de probabilités ou de rapport

de cotes (odds). Si on a, par exemple, une probabilité $p = 1/4$ de gagner à un jeu, cela signifie que sur 4 personnes une gagne et les trois autres perdent, soit un rapport de 1 gagnant sur trois perdants, c'est-à-dire $p/(1-p) = 1/3$. Ce rapport $p/(1-p)$ varie entre 0 (0 gagnant) et l'infini (que des gagnants) en passant par 1 (un gagnant pour un perdant).

Définition 2.2

L'odds (chance) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$\text{odds}(x) = \frac{p(x)}{1-p(x)}, \quad \text{où } p(x) = \mathbf{P}(Y = 1|X = x).$$

L'odds ratio (rapport des chances) entre deux individus x et \tilde{x} est

$$\text{OR}(x, \tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})} = \frac{\frac{p(x)}{1-p(x)}}{\frac{p(\tilde{x})}{1-p(\tilde{x})}}.$$

Exemple 2.4

Supposons qu'à un moment donné un cheval x a une probabilité $p(x) = 3/4$ de perdre. Cela signifie que sur 4 courses disputées, il peut espérer en gagner une et en perdre 3. L'odds vaut 3/1 (3 défaites contre 1 victoire, on dit également que ce cheval est à 3 contre 1). Pour la petite histoire, si l'espérance de gain était nulle, cela signifierait que pour 10 euros joués, on peut espérer 30 euros de bénéfice si le cheval gagne). Le rapport $p(x)/(1-p(x))$ varie entre 0 (que des victoires) et l'infini (que des défaites) en passant par 1 (une victoire pour une défaite).

Les odds ratio peuvent être utilisés de plusieurs manières :

1. **Comparaison de probabilités de succès entre deux individus** (voir Tableau 2.2) ;

$$\begin{array}{l} \text{OR}(x, \tilde{x}) > 1 \iff p(x) > p(\tilde{x}) \\ \text{OR}(x, \tilde{x}) = 1 \iff p(x) = p(\tilde{x}) \\ \text{OR}(x, \tilde{x}) < 1 \iff p(x) < p(\tilde{x}) \end{array}$$

TABLE 2.2 – Règles d'interprétation des odds ratio.

2. **Interprétation en terme de risque relatif** : dans le cas où $p(x)$ et $p(\tilde{x})$ sont très petits par rapport à 1, comme dans le cas d'une maladie très rare, on peut faire l'approximation $\text{OR}(x, \tilde{x}) \sim p(x)/p(\tilde{x})$ et interpréter simplement. Par exemple si $\text{OR}(x, \tilde{x}) = 4$ alors la réponse (maladie) est 4 fois plus probable dans le cas où $X = x$ que dans le cas où $X = \tilde{x}$.
3. **Mesure de l'impact d'une variable** : pour le modèle logistique

$$\text{logit } p_{\beta}(x) = \beta_0 + \beta_1 \mathbf{x}_1 + \dots + \beta_p \mathbf{x}_p,$$

il est facile de vérifier que

$$\text{OR}(x, \tilde{x}) = \exp(\beta_1(\mathbf{x}_1 - \tilde{\mathbf{x}}_1)) \dots \exp(\beta_p(\mathbf{x}_p - \tilde{\mathbf{x}}_p)).$$

Pour mesurer l'influence d'une variable sur l'odds ratio, il suffit de considérer deux observations x et \tilde{x} qui diffèrent uniquement par la $j^{\text{ème}}$ variable. On obtient alors

$$\text{OR}(x, \tilde{x}) = \exp(\beta_j(\mathbf{x}_j - \tilde{\mathbf{x}}_j)).$$

Ainsi une variation de la $j^{\text{ème}}$ variable d'une unité (sur l'échelle de cette variable) correspond à un odds ratio $\exp(\beta_j)$ qui est uniquement fonction du coefficient β_j . Le coefficient β_j permet de mesurer l'influence de la $j^{\text{ème}}$ variable sur le rapport $p_{\beta}(x)/(1-p_{\beta}(x))$ lorsque x_j varie d'une unité, et ceux indépendamment de la valeur de x_j . Une telle analyse peut se révéler intéressante pour étudier l'influence d'un changement d'état d'une variable qualitative.

Exemple 2.5

Reprenons l'exemple des cotes pour les courses de chevaux. On cherche à expliquer la performance d'un cheval en fonction du jockey qui le monte. Pour simplifier on suppose que l'on a que deux jockeys A et B . On désigne par Y la variable aléatoire qui prend pour valeurs 0 si le cheval remporte la course, 1 sinon. On considère le modèle logistique

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 \mathbf{1}_{x=B}.$$

On a $\text{OR}(B, A) = \exp(\beta_2)$. Imaginons que pour un échantillon de taille n on obtienne une estimation $\hat{\beta}_2 = \log(2)$. On a alors $\text{OR}(B, A) = 2$, ce qui signifie que la cote du cheval est multipliée par 2 lorsqu'il est monté par B par rapport à A .

2.6 Précision des estimateurs et tests

2.6.1 Loi asymptotique

Nous avons obtenu dans le théorème 2.1 le comportement asymptotique de l'estimateur du maximum de vraisemblance $\hat{\beta}$:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1}),$$

où $\mathcal{I}(\beta)$ est la matrice d'information de Fisher au point β . On déduit

$$(\hat{\beta} - \beta)' n \mathcal{I}(\beta) (\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \chi_{p+1}^2.$$

Un tel résultat n'est pas utilisable tel quel puisque la matrice $\mathcal{I}(\beta)$ est inconnue. On remarque que d'après la loi des grands nombres

$$\begin{aligned} \hat{\mathcal{I}}(\beta)_{ij} &= -\frac{1}{n} \sum_{k=1}^n \frac{\partial^2}{\partial \beta_i \partial \beta_j} \mathcal{L}_1(Y_i, \beta) = -\frac{1}{n} \frac{\partial^2}{\partial \beta_i \partial \beta_j} \sum_{k=1}^n \mathcal{L}_1(Y_i; \beta) = -\frac{1}{n} \frac{\partial^2}{\partial \beta_i \partial \beta_j} \mathcal{L}_n(Y_1, \dots, Y_n, \beta) \\ &= \frac{1}{n} (\mathbf{X}' W_\beta \mathbf{X})_{ij}, \end{aligned}$$

converge presque sûrement vers $\mathcal{I}(\beta)_{ij}$. Comme $\hat{\beta}$ converge faiblement vers β , on obtient grâce au théorème de Slutsky et aux opérations classiques sur la convergence en loi

$$(\hat{\beta} - \beta)' \hat{\Sigma} (\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \chi_{p+1}^2 \quad (2.6)$$

avec $\hat{\Sigma} = (\mathbf{X}' W_{\hat{\beta}} \mathbf{X})$.

2.6.2 Intervalles de confiance

On déduit du paragraphe précédent qu'un estimateur de la variance de $\hat{\beta}_j$ est donné par le $j^{\text{ème}}$ terme de la diagonale de $\hat{\Sigma}^{-1}$. Notons $\hat{\sigma}_j^2$ cet estimateur. Il vient

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \chi_1^2 \quad \text{ou encore} \quad \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \quad (2.7)$$

Un intervalle de confiance (asymptotique) de niveau $1 - \alpha$ pour β_j est donc donné par

$$IC_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j - u_{1-\alpha/2} \hat{\sigma}_j; \hat{\beta}_j + u_{1-\alpha/2} \hat{\sigma}_j \right],$$

où $u_{1-\alpha/2}$ représente le quantile de niveau $(1 - \alpha/2)$ de la loi normale $\mathcal{N}(0, 1)$.

La validité de ces intervalles est toute relative puisqu'il s'agit d'une approximation valable asymptotiquement. Il est toujours possible de compléter cette étude par un bootstrap afin d'obtenir d'autres intervalles de confiance dans le cas où ceux-ci sont particulièrement importants. Cela dit, en pratique, on se contente de l'intervalle de confiance bâti grâce à la matrice d'information de Fisher.

On déduit également de (2.7) les tests de nullité des coefficients du modèle. On note $H_0 : \beta_j = 0$ et $H_1 : \beta_j \neq 0$, alors sous H_0 , $\hat{\beta}_j/\hat{\sigma}_j \xrightarrow{L} \mathcal{N}(0, 1)$. On rejettera H_0 si la valeur observée de $\hat{\beta}_j/\hat{\sigma}_j$ dépasse en valeur absolue le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$.

Exemple 2.6

Reprenons l'exemple du TP1. Un chef d'entreprise souhaite vérifier la qualité de ces machines en fonction de l'âge et de la marque des moteurs. Il dispose

- d'une variable binaire Y (1 si le moteur a déjà connu une panne, 0 sinon) ;
- d'une variable quantitative **age** représentant l'âge du moteur ;
- d'une variable qualitative à 3 modalités **marque** représentant la marque du moteur.

On souhaite expliquer la variable Y à partir des deux autres variables. On construit un modèle logistique permettant d'expliquer Y par l'âge du moteur et sa marque.

```
> model <- glm(panne ~ ., data=donnees, family=binomial)
```

On obtient les variances estimées des estimateurs via

```
> P <- predict(model, type="response")
> W <- diag(P*(1-P))
> X1 <- rep(1, n)
> X2 <- donnees$age
> X3 <- as.numeric(donnees$marque==1)
> X4 <- as.numeric(donnees$marque==3)
> X <- matrix(c(X1, X2, X3, X4), ncol=4)
> #écarts types estimés
> sqrt(diag(solve(Sigma)))
[1] 0.83301450 0.09398045 0.81427979 1.05357830
```

Bien évidemment, on peut retrouver ces écarts types ainsi que les probabilités critiques des tests de nullité des coefficients avec

```
> summary(model)
```

Call:

```
glm(formula = panne ~ ., family = binomial, data = donnees)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4232	-1.2263	0.9082	1.1062	1.5982

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.47808	0.83301	0.574	0.566
age	0.01388	0.09398	0.148	0.883
marque1	-0.41941	0.81428	-0.515	0.607
marque3	-1.45608	1.05358	-1.382	0.167

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 45.717 on 32 degrees of freedom
 Residual deviance: 43.502 on 29 degrees of freedom
 AIC: 51.502

2.6.3 Tests de nullité de q coefficients libres

La théorie du maximum de vraisemblance nous donnant la loi (asymptotique) des estimateurs, il est possible de tester la significativité des variables explicatives. Pour cela, trois tests sont généralement utilisés :

- Le test de Wald ;
- Le test du rapport de vraisemblance ou de la déviance.
- Le test du score ;

Les hypothèses s'écrivent :

$$H_0 : \beta_{j_1} = \beta_{j_2} = \dots = \beta_{j_q} = 0 \quad \text{contre} \quad H_1 : \exists k \in \{1, \dots, q\} : \beta_{j_k} \neq 0.$$

Pour alléger les notations, nous supposons sans perte de généralité que nous testons la nullité des q premiers coefficients du modèle

$$H_0 : \beta_0 = \beta_1 = \dots = \beta_{q-1} = 0 \quad \text{contre} \quad H_1 : \exists k \in \{0, \dots, q-1\} : \beta_k \neq 0.$$

Test de Wald Il est basé sur (2.6). On note $\beta_{0, \dots, q-1}$ le vecteur composé des q premières composantes de β et $\hat{\Sigma}_{0, \dots, q}^{-1}$ la matrice bloc composée des q premières lignes et colonnes de $\hat{\Sigma}^{-1}$. Il est facile de voir que sous H_0

$$\hat{\beta}'_{0, \dots, q-1} \hat{\Sigma}_{0, \dots, q}^{-1} \hat{\beta}_{0, \dots, q-1} \xrightarrow{\mathcal{L}} \chi_q^2.$$

Test du rapport de vraisemblance ou de la déviance La statistique de test est basée sur la différence des rapports de vraisemblance entre le modèle complet et le modèle sous H_0 . On note $\hat{\beta}_{H_0}$ l'estimateur du maximum de vraisemblance contraint par H_0 (il s'obtient en supprimant les q premières variables du modèle). On a alors sous H_0

$$2(\mathcal{L}_n(\hat{\beta}) - \mathcal{L}_n(\hat{\beta}_{H_0})) \xrightarrow{\mathcal{L}} \chi_q^2.$$

Test du score On cherche ici à vérifier si la fonction de score (gradient de la log-vraisemblance) est proche de 0 sous H_0 . Sous H_0 on a

$$S(\hat{\beta}_{H_0})' \hat{\Sigma}_{H_0}^{-1} S(\hat{\beta}_{H_0}) \xrightarrow{\mathcal{L}} \chi_q^2,$$

où $\hat{\Sigma}_{H_0} = \mathbf{X}'W_{\hat{\beta}_{H_0}}\mathbf{X}$.

Pour ces 3 tests, on rejette l'hypothèse nulle si la valeur observée de la statistique de test dépasse le quantile d'ordre $1 - \alpha$ de la loi χ_q^2 . Pour la preuve des convergences des statistiques du maximum de vraisemblance et du score, on pourra se référer à l'annexe D de Antoniadis *et al* (1992). La figure 2.5 permet de visualiser les trois tests. Le test du score revient à tester la nullité de la pente en $\hat{\beta}_{H_0}$ ($\hat{\beta}$ sous H_0), le test de Wald la nullité de la distance entre $\hat{\beta}$ et $\hat{\beta}_{H_0}$ et le test du rapport de vraisemblance la nullité de la différence entre les vraisemblances en ces deux points.

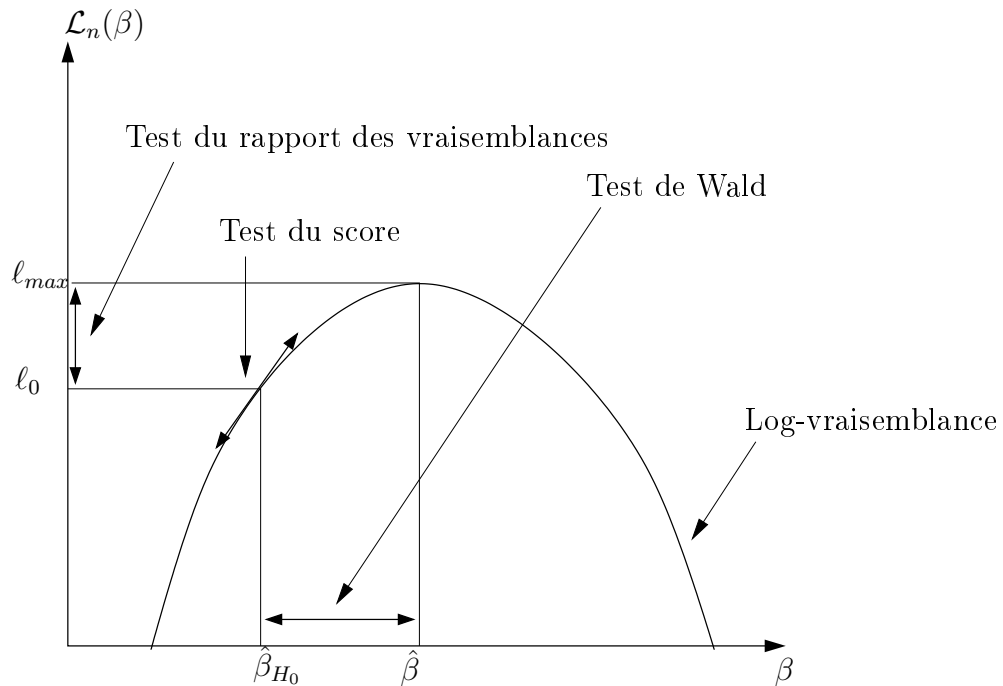


FIGURE 2.5 – Rapport de vraisemblance, score, test de Wald.

Remarque

- La PROC LOGISTIC sous SAS réalise les trois tests pour $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$.
- Pour les tests “variable par variable” ou paramètre par paramètre

$$H_0 : \beta_j = 0 \quad \text{contre} \quad H_1 : \beta_j \neq 0,$$

la PROC LOGISTIC utilise le test de Wald.

Exemple 2.7

Reprenons l'exemple précédent. Le modèle s'écrit

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 \text{age} + \beta_2 \mathbf{1}_{\text{marque}=1} + \beta_3 \mathbf{1}_{\text{marque}=3},$$

la modalité 0 de la variable `marque` est prise comme modalité de référence. On effectue les 3 tests présentés ci-dessus pour

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad H_1 : \exists j \in \{1, 2, 3\} \beta_j \neq 0.$$

Le calcul des statistiques de test et des probabilités critiques s'obtient avec les commandes

```
> #test de Wald
> Sigma <- t(X)%*%W%*%X
> inv_Sigma <- solve(Sigma)
> inv_SigmaH0 <- inv_Sigma[2:4,2:4]
> betaH0 <- coef(model)[2:4]
> statWald <- t(betaH0)%*%solve(inv_SigmaH0)%*%betaH0
> pcWald <- 1-pchisq(statWald,df=3)
> pcWald
      [,1]
[1,] 0.5655233
> #test du rapport de vraisemblance
> modelH0 <- glm(panne~1,data=donnees,family=binomial)
> statRappvrais <- 2*(logLik(model)-logLik(modelH0))
```

```

> pcRappvrais <- 1-pchisq(statRappvrais,df=3)
> pcRappvrais[1]
[1] 0.528955
> #test du score
> prevH0 <- predict(modelH0, type="response")
> scoreH0 <- t(X)%*(as.numeric(donnees$panne)-1-prevH0)
> WHO <- diag(prevH0*(1-prevH0))
> SigmaH0 <- t(X)%*WHO%*X
> statscore <- t(scoreH0)%*solve(SigmaH0)%*scoreH0
> pcscore <- 1-pchisq(statscore,df=3)
> pcscore
      [,1]
[1,] 0.5392691

```

Ces 3 tests acceptent l'hypothèse nulle.

2.7 Le schéma d'échantillonnage rétrospectif

Jusqu'à présent nous avons considéré un échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. de même loi que (X, Y) . Cette phase d'échantillonnage n'est pas nécessairement toujours la mieux adaptée. Considérons l'exemple suivant.

Exemple 2.8

Une clinique cherche à mesurer l'effet du tabac sur le cancer du poumon. Elle prélève parmi ses patients un échantillon composé de $n_1 = 250$ personnes atteintes par le cancer et $n_0 = 250$ personnes ne présentant pas la maladie. Les résultats (simulés!) de l'étude sont présentés dans le tableau suivant :

	Fumeur	Non fumeur
Non malade	48	202
Malade	208	42

TABLE 2.3 – Résultats de l'enquête.

Le statisticien responsable de l'étude réalise un modèle logistique. Les sorties sur R sont :

```

Call: glm(formula = Y ~ X, family = binomial)

Coefficients:
(Intercept)  Xnon_fumeur
      1.466      -2.773

Degrees of Freedom: 499 Total (i.e. Null);  498 Residual
Null Deviance:      692.3
Residual Deviance: 499.9      AIC: 503.9

```

On obtient $p_{\hat{\beta}}(\text{fumeur}) = 0.812$, ce qui peut paraître un peu élevé. La valeur surprenante d'une telle estimation vient du fait l'échantillonnage n'est pas fait selon protocole précédent : il est fait conditionnellement à Y . Il est facile de voir que les répartitions d'individus selon la variable Y ne sont pas les mêmes dans la population et dans l'échantillon. Ceci va entraîner un biais au niveau des estimateurs. On peut modéliser le schéma d'échantillonnage rétrospectif de la manière suivante.

Le schéma d'échantillonnage On s'intéresse toujours à la loi conditionnelle de $Y|X$ qui est une bernoulli de paramètre $p_\beta(x)$ telle que

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

On désigne toujours par P_X la loi de X . La différence ici est que l'échantillon n'est pas i.i.d. de même loi que (X, Y) . On désigne par S une variable aléatoire qui prend pour valeur 0 et 1 et par τ_1 et τ_0 les taux de sondage $\mathbf{P}(S = 1|Y = 1)$ et $\mathbf{P}(S = 1|Y = 0)$. Pour un individu (x, y) généré selon $P_{X,Y}$, on tire une réalisation s de S selon une loi de Bernoulli de paramètre τ_y , si $s = 1$ on garde (x, y) dans l'échantillon, sinon on le jette. On répète le protocole jusqu'à obtenir n individus.

Le modèle logistique Ici un individu est représenté par un triplet (X, Y, S) et l'échantillon constitué s'écrit $(x_1, y_1, 1), \dots, (x_n, y_n, 1)$. Le modèle étudié pour ce schéma d'échantillonnage est donc $\mathcal{M}_n = \{\{0, 1\}^n, B(p_\gamma(x_1)) \otimes \dots \otimes B(p_\gamma(x_n)), \gamma \in \mathbb{R}^{p+1}\}$ avec

$$\text{logit } p_\gamma(x) = \text{logit } \mathbf{P}_\gamma(Y = 1|X = x, S = 1) = \gamma_0 + \gamma_1 x_1 + \dots + \gamma_p x_p.$$

Ainsi, il suffit d'appliquer tout ce qui a été vu dans ce chapitre pour obtenir les estimateurs des γ_j (on déduit également les intervalles de confiance, test...). La question qui se pose est : comment retrouver $p_\beta(x)$ en connaissant $p_\gamma(x)$?

Proposition 2.2

Avec les notations ci dessus, on a

$$\text{logit } p_\gamma(x) = \text{logit } p_\beta(x) + \log \frac{\tau_1}{\tau_0},$$

par conséquent

$$\text{logit } p_\beta(x) = \left(\gamma_0 - \frac{\tau_1}{\tau_0} \right) + \gamma_1 x_1 + \dots + \gamma_p x_p.$$

Preuve

Il suffit de remarquer que grâce au théorème de Bayes on a

$$p_\gamma(x) = \frac{p_\beta(x) \mathbf{P}(S = 1|Y = 1, X = x)}{\mathbf{P}(S = 1)} = \frac{p_\beta(x) \mathbf{P}(S = 1|Y = 1)}{\mathbf{P}(S = 1)}.$$

Ce résultat est intéressant puisqu'il implique que le biais dû au mode d'échantillonnage est exclusivement concentré sur la constante du modèle. Si de plus on connaît les taux de sondage, alors on peut corriger ce biais. Il s'agit là d'une propriété spécifique au modèle logistique.

Ce schéma d'échantillonnage est souvent utilisé lorsque les probabilités $\pi_i = \mathbf{P}(Y = i)$ (probabilité d'appartenance au groupe i) sont très différentes les unes des autres. Dans de tels cas, le schéma classique conduit à travailler avec des effectifs trop petits dans certains groupes, alors que le schéma rétrospectif permet de travailler avec des effectifs comparables. Par exemple, en diagnostic médical, on a souvent des problèmes de discrimination entre deux groupes où l'un des groupes (1 par exemple) est associé à une maladie et l'autre est caractéristique de l'absence de cette maladie. Dans de telles situations, on a bien sûr π_1 beaucoup plus petit que π_0 . L'usage consiste alors à étudier deux échantillons de taille à peu près équivalente ($n_1 \sim n_0$), le premier étant celui des malades, le second celui des individus sains.

Exemple 2.9

Reprenons l'exemple 2.8. Des études préalables ont montré que les probabilités $\mathbf{P}(Y = 1) = \pi_1$ et $\mathbf{P}(Y = 0) = \pi_0$ pouvaient être estimées par 0.005 et 0.995. L'échantillonnage a été effectué de manière à travailler avec des échantillons équilibrés, nous pouvons donc estimer les probabilités $\mathbf{P}(Y = 1|S = 1)$ et $\mathbf{P}(Y = 0|S = 1)$ par 1/2. Ainsi en remarquant que

$$\frac{\tau_1}{\tau_0} = \frac{\mathbf{P}(Y = 1|S = 1) \mathbf{P}(Y = 0)}{\mathbf{P}(Y = 0|S = 1) \mathbf{P}(Y = 1)}$$

on peut estimer $\frac{\tau_1}{\tau_0}$ par $0.995/0.005 \approx -5.293$. On a donc

$$\text{logit } p_{\hat{\beta}}(x) = (1.466 - 5.293) - 2.7731_{\text{non fumeur}}(x).$$

D'où $p_{\hat{\beta}}(\text{fumeur}) = 0.0213$.

2.8 Un exemple avec R

Le traitement du cancer de la prostate change si le cancer a atteint ou non les nœuds lymphatiques entourant la prostate. Pour éviter une investigation lourde (ouverture de la cavité abdominale) un certain nombre de variables sont considérées comme explicatives de la variable Y binaire : $Y = 0$ le cancer n'a pas atteint le réseau lymphatique, $Y = 1$ le cancer a atteint le réseau lymphatique. Le but est d'expliquer Y par les variables suivantes :

- âge du patient au moment du diagnostic : **age** ;
- le niveau d'acide phosphatase sérique : **acide** ;
- Le résultat d'une analyse par rayon X, 0=négatif, 1=positif : **rayonx** ;
- La taille de la tumeur, 0=petite, 1=grande : **taille** ;
- L'état pathologique de la tumeur déterminé par biopsie (0=moyen, 1=grave) : **grade** ;
- Le logarithme népérien du niveau d'acidité : **log.acid**.

	age	acide	rayonx	taille	grade	log.acid.
1	66	0.48	0	0	0	-0.73396918
2	68	0.56	0	0	0	-0.57981850
3	66	0.50	0	0	0	-0.69314718
4	56	0.52	0	0	0	-0.65392647
5	58	0.50	0	0	0	-0.69314718
6	60	0.49	0	0	0	-0.71334989
7	65	0.46	1	0	0	-0.77652879
8	60	0.62	1	0	0	-0.47803580
9	50	0.56	0	0	1	-0.57981850
10	49	0.55	1	0	0	-0.59783700

TABLE 2.4 – Représentation des dix premiers individus.

2.8.1 Modèles “simples”

Nous sommes en présence de 6 variables explicatives $\mathbf{X}_1, \dots, \mathbf{X}_6$ avec :

- $\mathbf{X}_1, \mathbf{X}_2$ et \mathbf{X}_6 quantitatives ;
- $\mathbf{X}_3, \mathbf{X}_4$ et \mathbf{X}_5 qualitatives (2 niveaux pour chacune).

Premier modèle

Considérons tout d'abord les trois variables explicatives qualitatives $X = (\mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5)$:

$$\text{logit } p_{\beta}(x) = \beta_0 + \beta_3 \mathbf{1}_{\{\mathbf{x}_3=1\}} + \beta_4 \mathbf{1}_{\{\mathbf{x}_4=1\}} + \beta_5 \mathbf{1}_{\{\mathbf{x}_5=1\}}.$$

Ce modèle possède 4 paramètres. Les sorties du logiciel R sont :

```
> model_quali<-glm(Y~rayonx+taille+grade,data=donnees,family=binomial)
> model_quali

Call:  glm(formula = Y ~ rayonx + taille + grade, family = binomial,      data = donnees)

Coefficients:
(Intercept)      rayonx1      taille1      grade1
   -2.1455       2.0731       1.4097       0.5499

Degrees of Freedom: 52 Total (i.e. Null);  49 Residual
Null Deviance:      70.25
Residual Deviance: 52.78      AIC: 60.78
```

Si par exemple $(\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5) = (1, 0, 1)$, on aura alors :

$$\text{logit } p_{\hat{\beta}}(x) = \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_5 = -2.1455 + 2.0731 + 0.5499 = 0.4785$$

et

$$p_{\hat{\beta}}(x) = \frac{\exp(0.4785)}{1 + \exp(0.4785)} = 0.6174.$$

Ainsi, dans un contexte de prévision, nous serons tentés d'assigner le label 1 à la nouvelle observation x .

Deuxième modèle

Considérons maintenant le modèle uniquement composé de variables quantitatives,

$$\text{logit } p_{\beta}(x) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_6 \mathbf{x}_6.$$

```
> model_quant<-glm(Y~age+acide+log.acid.,data=donnees,family=binomial)
> model_quant

Call:  glm(formula = Y ~ age + acide + log.acid., family = binomial,      data = donnees)

Coefficients:
(Intercept)      age      acide      log.acid.
  12.34700    -0.02805   -9.96499    10.54332

Degrees of Freedom: 52 Total (i.e. Null);  49 Residual
Null Deviance:      70.25
Residual Deviance: 59.95      AIC: 67.95
```

Troisième modèle

Le modèle "complet" à 6 variables s'écrit

$$\text{logit } p_{\beta}(x) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{1}_{\{\mathbf{x}_3=1\}} + \beta_4 \mathbf{1}_{\{\mathbf{x}_4=1\}} + \beta_5 \mathbf{1}_{\{\mathbf{x}_5=1\}} + \beta_6 \mathbf{x}_6.$$

```
> model_complet<-glm(Y~.,data=donnees,family=binomial)
> model_complet

Call:  glm(formula = Y ~ ., family = binomial, data = donnees)

Coefficients:
(Intercept)      age      acide      rayonx1      taille1      grade1
 10.08672    -0.04289    -8.48006     2.06673     1.38415     0.85376
 log.acid.
  9.60912

Degrees of Freedom: 52 Total (i.e. Null);  46 Residual
Null Deviance:      70.25
Residual Deviance: 44.77      AIC: 58.77
```

2.8.2 Encore d'autres modèles...

Comme dans le cas du modèle "linéaire" on peut également considérer des *interactions* entre les variables explicatives. On dit qu'il y a interaction entre deux variables $F1$ et $F2$ sur une variable Y si l'effet de l'une des variables diffère selon la modalité de l'autre. Remarquons que cette notion n'a rien à voir avec celle de corrélation qui ne concerne que deux variables alors que l'interaction met en jeu une troisième variable Y .

Exemple 2.10 (Construction d'interaction)

On s'intéresse à l'effet de deux traitements \mathbf{X}_1 et \mathbf{X}_2 sur le rhume. Le traitement \mathbf{X}_1 consiste à prendre à intervalle de temps réguliers deux verres de cognac et \mathbf{X}_2 représente un traitement aux antibiotiques (il n'est pas difficile de comprendre l'intérêt d'envisager une interaction). La variable réponse Y correspond à l'état du patient (1 si malade, 0 si bonne santé). N'ayant pas encore trouvé suffisamment de volontaires pour réaliser l'étude, on simule un échantillon suivant le modèle

1. deux facteurs \mathbf{X}_1 et \mathbf{X}_2 à deux niveaux équiprobables ;
2. la loi de Y conditionnellement à \mathbf{X}_1 et \mathbf{X}_2 est donnée dans le tableau 2.5.

$X_2 \backslash X_1$	0	1
0	$B(0.95)$	$B(0.05)$
1	$B(0.05)$	$B(0.95)$

TABLE 2.5 – Lois conditionnelles de Y (B désigne la loi de Bernoulli).

On estime les pourcentages de mal classés sur un échantillon indépendant (voir section 3.1.4) et on reporte dans le tableau suivant les pourcentages de mal classés pour les modèles sans et avec interaction. Nous voyons l'intérêt d'inclure une interaction pour cet exemple.

Sans	0.54
Avec	0.065

TABLE 2.6 – Pourcentages de mal classés.

Pour l'exemple du cancer de la prostate, le modèle avec toutes les interactions d'ordre 2 s'écrit :

```
> model_inter<-glm(Y~.^2,data=donnees,family=binomial)
Warning message:
des probabilités ont été ajustées numériquement à 0 ou 1 in: glm.fit(x = X, y = Y,
```

```

weights = weights, start = start, etastart = etastart,
> model_inter

Call:  glm(formula = Y ~ .^2, family = binomial, data = donnees)

Coefficients:
(Intercept)          age          acide          rayonx1
 2.843e+17   -4.229e+15   -3.117e+17   -5.453e+16
  taille1          grade1          log.acid.          age:acide
 2.516e+16   -5.778e+15    2.026e+17    4.665e+15
 age:rayonx1  age:taille1  age:grade1  age:log.acid.
 2.077e+13   -5.245e+13   -1.670e+14   -2.869e+15
 acide:rayonx1  acide:taille1  acide:grade1  acide:log.acid.
 5.572e+16   -2.420e+16    2.336e+16   -5.687e+15
 rayonx1:taille1  rayonx1:grade1  rayonx1:log.acid.  taille1:grade1
 1.129e+15   -1.176e+15   -4.004e+16   -5.496e+15
 taille1:log.acid.  grade1:log.acid.
 8.625e+15   -1.228e+16

Degrees of Freedom: 52 Total (i.e. Null); 31 Residual
Null Deviance:      70.25
Residual Deviance: 504.6      AIC: 548.6

```

On peut vérifier que ce modèle nécessite l'estimation de 22 paramètres ($1 + 6 + \binom{6}{2}$). Bien entendu, d'autres sous-modèles avec interactions peuvent être utilisés. De plus, nous pouvons nous demander si toutes les variables sont bien explicatives? Dès lors, des méthodes sélection et validation de modèles doivent être envisagées.

Chapitre 3

Sélection et validation de modèles

Ce chapitre se divise en deux parties :

1. **Sélection** : Etant donnés M modèles $\mathcal{M}_1, \dots, \mathcal{M}_M$, comment choisir le “meilleur” à partir de l'échantillon dont on dispose.
2. **Validation** : Est-ce que le modèle sélectionné est “bon” ? En statistique cette question peut être abordée de différentes façons :
 - Est-ce que la qualité d'ajustement globale est satisfaisante : le modèle décrit-il bien les valeurs observées ?
 - Ce type de question fait l'objet des tests d'ajustement ou d'adéquation (goodness of fit).
 - L'ajustement peut être aussi regardé observation par observation (individus aberrants) par des méthodes graphiques (analyse des résidus) ou analytiques.
 - Est-ce que les hypothèses sont vérifiées ? Les méthodes sont essentiellement graphiques (analyse des résidus).
 - L'influence des observations sur l'estimation des paramètres peut être aussi envisagée (distance de Cook, robustesse).

Dans ce chapitre nous allons traiter ces questions à travers l'exemple du modèle logistique. Nous noterons M_β le modèle logistique défini par le vecteur de paramètre β . L'ensemble des méthodes présentées peut s'étendre à d'autres problématiques de sélection-validation de modèles.

3.1 Sélection ou choix de modèle

Si on se restreint à des modèles logistiques, sélectionner un modèle revient à choisir les variables (interactions incluses) qui vont constituer le modèle.

3.1.1 Un outil spécifique : la déviance

Il est difficile de se faire une idée sur l'ajustement en se basant sur la valeur vraisemblance puisqu'elle dépend (entre autres) de la taille de l'échantillon. Pour la régression logistique, un outil spécifique est introduit : la déviance. Elle compare la vraisemblance obtenue à celle d'un modèle de référence : le *modèle complet* (ou *modèle saturé*). Ce modèle ne place pas de contrainte sur la forme du paramètre $p(x)$ de la loi de $Y|X = x$ (cette loi est une Bernoulli de paramètre $p(x)$).

Modèle saturé en présence de données individuelles

En présence de données individuelles $(x_1, y_1), \dots, (x_n, y_n)$, il est facile de voir que, sous le modèle saturé $\mathcal{M}_{sat} = \{\{0, 1\}^n, \{B(p(x_1)) \otimes \dots \otimes B(p_{sat}(x_n)), p_{sat}(x_i) \in \mathbb{R}\}\}$, l'estimateur du maximum de vraisemblance $\hat{p}_{sat}(x_i)$ de $p_{sat}(x_i)$ est donné par $\hat{p}_{sat}(x_i) = y_i$ la log-vraisemblance maximisée \mathcal{L}_{sat} vaut ainsi 0. Ce modèle contient autant de paramètres que de données et reconstitue "parfaitement" l'échantillon.

Modèle saturé en présence de données répétées

En présence de données répétées de design $\{(x_1, n_1), \dots, (x_T, n_T)\}$ il est facile de voir que les estimateurs du maximum de vraisemblance du modèle saturé sont donnés par $\hat{p}_{sat}(x_t) = \bar{y}_t = y_t/n_t, t = 1, \dots, T$. Dans ce cas, la log-vraisemblance maximisée est donnée par :

$$\begin{aligned} \mathcal{L}_{sat} &= \prod_{t=1}^T \binom{n_t}{y_t} \hat{p}_{sat}(x_t)^{y_t} (1 - \hat{p}_{sat}(x_t))^{n_t - y_t} \\ &= \sum_{t=1}^T \log \binom{n_t}{y_t} + \sum_{t=1}^T y_t \log \hat{p}_{sat}(x_t) + (n_t - y_t) \log(1 - \hat{p}_{sat}(x_t)). \end{aligned}$$

Que ce soit en présence de données individuelles ou répétées, le modèle saturé possède autant de paramètres que de points du design. Ce modèle est le modèle le plus complexe (en terme de nombre de paramètres) puisqu'il propose un coefficient différent pour chaque point du design. Tous les autres modèles sont emboîtés dans celui-ci.

Définition 3.1

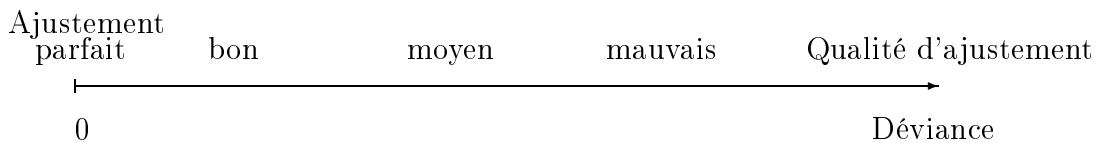
La déviance d'un modèle M_β est définie par

$$D_{M_\beta} = 2(\mathcal{L}_{sat} - \mathcal{L}_n(\hat{\beta})).$$

En présence de données individuelles on a $D_{M_\beta} = -2\mathcal{L}_n(\hat{\beta})$ et en présence de données répétées, la déviance s'écrit

$$\begin{aligned} D_{M_\beta} &= 2 \sum_{t=1}^T n_t \left[\bar{y}_t \log \frac{\bar{y}_t}{p_{\hat{\beta}}(x_t)} + (1 - \bar{y}_t) \log \frac{1 - \bar{y}_t}{1 - p_{\hat{\beta}}(x_t)} \right] \\ &= 2 \sum_{t=1}^T \left[y_t \log \frac{\bar{y}_t}{p_{\hat{\beta}}(x_t)} + (n_t - y_t) \log \frac{n_t - y_t}{n_t - p_{\hat{\beta}}(x_t)} \right], \end{aligned}$$

où $\bar{y}_t = y_t/n_t$. La déviance est égale à 2 fois une différence de log-vraisemblance. Elle constitue un écart en terme de log-vraisemblance entre le modèle saturé d'ajustement maximum et le modèle considéré.



Exemple 3.1 (calcul de déviance)

Considérons l'exemple du cancer de la prostate et calculons d'abord la déviance pour le modèle $Y \sim \text{age} + \text{acide}$. Nous sommes ici en présence de données individuelles, on obtient la déviance via les commandes :

```
> mod1 <- glm(Y~age+acide,data=donnees,family=binomial) #construction du modele
> #calcul de la vraisemblance
> prev <- mod1$fitted.values #on obtient les pi
> vrais <- rep(0,nrow(donnees))
> vrais[donnees$Y==1] <- prev[donnees$Y==1]
> vrais[donnees$Y==0] <- 1-prev[donnees$Y==0]
> vrais <- prod(vrais) #vrais est la vraisemblance du modele
> dev <- -2*log(vrais) #pour ce modele, il n'y a pas de repetitions aux points du
                                design, donc Lsat=0

> dev
[1] 65.72393
```

Bien entendu, le logiciel peut retourner directement la valeur de la déviance

```
> mod1$deviance
[1] 65.72393
```

Si maintenant on considère le modèle $Y \sim \text{age} + \text{taille}$, nous sommes en présence de données répétées. Les données se trouvent dans le fichier "donnees_bin_age_taille.txt" dont voici les premières lignes :

```
"age" "taille" "Y1" "Y0"
49 "0" 0 1
50 "0" 1 0
51 "0" 0 2
52 "0" 0 1
56 "0" 1 3
58 "0" 0 2
```

Les deux premières colonnes représentent les valeurs des variables explicatives. On retrouve ensuite (colonne Y1) le nombre de réponses $Y=1$ et (colonne Y0) le nombre de réponses $Y=0$. Le modèle est construit via la commande :

```
> donnees1 <- read.table("donnees_bin_age_taille.txt",header=T)
> model1 <- glm(cbind(Y1,Y0)~age+taille,data=donnees1,family=binomial)
```

La déviance se calcule comme suit

```
> prev <- model1$fitted #calcul des pi
> ni <- apply(donnees1[,3:4],1,sum)
> ti <- donnees1$Y1
> ybi <- donnees1$Y1/ni
> #calcul des termes combinatoires (facultatif)
> vect_comb <- rep(0,nrow(donnees1))
> for (i in 1:nrow(donnees1)){
+ vect_comb[i] <- log(prod(1:ni[i])/(max(c(prod(1:ti[i]),1))*
                                max(c(prod(1:(ni[i]-ti[i]),1))))))
> vect <- ni*(ybi*log(prev)+(1-ybi)*log(1-prev))
> vrais_model1 <- sum(vect_comb+vect) #log_vraisemblance du modele
> #modele sature
> vect_sat <- ni*(ybi*log(ybi)+(1-ybi)*log(1-ybi))
```

```
> vect_sat[is.na(vect_sat)] <- 0
> vrais_modelsat <- sum(vect_comb+vect_sat)
> #on deduit la deviance
> 2*(vrais_modelsat-vrais_model1)
[1] 37.15260
```

On retrouve cette valeur directement

```
> model1$deviance
[1] 37.15260
```

3.1.2 Test de déviance entre 2 modèles emboîtés

Rappelons que par définition un modèle est emboîté dans un autre plus général (ou plus grand) lorsqu'il est un cas particulier de ce modèle plus général.

Exemple 3.2

Les modèles logistiques définis par

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$$

et

$$\text{logit } p_\gamma(x) = \gamma_0 + \gamma_1 \mathbf{x}_1 + \gamma_2 \mathbf{x}_2 + \gamma_3 \mathbf{x}_3$$

sont emboîtés l'un dans l'autre.

Il est facile de voir que faire un test entre modèles emboîtés est équivalent à tester la nullité de certains coefficients du grand modèle. On peut ainsi utiliser les tests de Wald, du maximum de vraisemblance ou du score pour tester deux modèles emboîtés. Si par exemple, $\mathcal{M}_1 \subset \mathcal{M}_2$, alors on a

$$-2(\mathcal{L}_n(\hat{\mathcal{M}}_1) - \mathcal{L}_n(\hat{\mathcal{M}}_2)) \xrightarrow{L} \chi_{p_2-p_1}^2$$

où p_j désigne la dimension du modèle M_j .

Remarque

La statistique de test peut s'écrire $D_{M_1} - D_{M_2}$, c'est pourquoi ce test est également appelé test de déviance.

3.1.3 Critère de choix de modèles

Le test que nous venons d'étudier permet de sélectionner un modèle parmi deux modèles emboîtés. Or, à partir de p variables explicatives, il est possible de définir un grand nombre de modèles logistiques qui ne sont pas forcément emboîtés. L'utilisation d'un simple test de déviance se révèle alors insuffisante. On a recours à des critères de choix de modèles qui permettent de comparer des modèles qui ne sont pas forcément emboîtés les uns dans les autres.

Les critères AIC et BIC sont les plus utilisés. Ces critères sont basés sur la philosophie suivante : plus la vraisemblance est grande, plus grande est donc la log-vraisemblance et meilleur est le modèle (en terme d'ajustement). Cependant la vraisemblance augmente avec la complexité du modèle, et choisir le modèle qui maximise la vraisemblance revient à choisir le modèle saturé. Ce modèle est clairement sur-paramétré, il "sur-ajuste" les données (overfitting).

Exemple 3.3

On considère un échantillon de taille $n = 100$ simulé suivant le modèle :

$$X_i \sim \mathcal{N}(0, 1), \quad U_i \sim \mathcal{U}[0, 1], \quad \text{et} \quad Y_i = \begin{cases} \mathbf{1}_{U_i \leq 0.25} & \text{si } X_i \leq 0 \\ \mathbf{1}_{U_i \geq 0.25} & \text{si } X_i \geq 0. \end{cases}$$

Les données sont représentées sur la figure 3.1 : environ 3/4 des labels valent 0 pour les valeurs de X_i négatives et 1 pour les valeurs positives. Le modèle saturé ajuste parfaitement les observations. Nous voyons cependant qu'il est difficile, pour ne pas dire impossible à utiliser dans un contexte de prévision. De plus le modèle saturé possède ici $n = 100$ paramètres tandis que le modèle logistique n'en possède que 2. Ceci est nettement plus avantageux pour expliquer Y d'un point de vue descriptif.

Pour choisir des modèles plus parcimonieux, une stratégie consiste à pénaliser la vraisemblance par une fonction du nombre de paramètres.

– Par définition l'AIC (Akaike Informative Criterion) pour un modèle \mathcal{M} de dimension p est défini par

$$AIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\mathcal{M}}) + 2p.$$

– Le critère de choix de modèle le BIC (Bayesian Informative Criterion) pour un modèle \mathcal{M} de dimension p paramètres est défini par

$$BIC(\mathcal{M}) = -2\mathcal{L}_n(\hat{\mathcal{M}}) + p \log(n).$$

On choisira le modèle qui possède le plus petit AIC ou BIC . L'utilisation de ces critères est simple. Pour chaque modèle concurrent le critère de choix de modèles est calculé et le modèle qui présente le plus faible est sélectionné.

Remarque

- Remarquons que certains logiciels utilisent $-AIC$ et $-BIC$ il est donc prudent de bien vérifier dans quel sens doivent être optimisés ces critères (maximisation ou minimisation). Ceci peut être fait aisément en comparant un modèle très mauvais (composé uniquement de la constante par exemple) à un bon modèle et de vérifier dans quel sens varie les critères de choix.
- Les pénalités ne sont pas choisies au hasard... Le critère BIC est issue de la théorie Bayésienne. En deux mots, les modèles candidats sont vus comme des variables aléatoires sur lesquels on met une loi de probabilité a priori. Dans ce contexte, choisir la modèle qui possède le plus petit BIC revient à choisir le modèle qui maximise la probabilité a posteriori.

3.1.4 Apprentissage/validation

Ce critère mesure la performance d'un modèle en terme de prévision. Il convient donc de définir au préalable une règle de prévision pour un modèle logistique. Un modèle M_β fournit une estimation $\hat{p}_\beta(x) = p_\beta(x)$, il est naturel de définir de définir une règle de prévision \hat{g}_β à partir de cette estimation :

$$\hat{g}_\beta(x) = \begin{cases} 1 & \text{si } \hat{p}_\beta(x) \geq s \\ 0 & \text{sinon,} \end{cases}$$

où $s \in [0, 1]$ est un seuil fixé par l'utilisateur. Il existe plusieurs façons de choisir ce seuil, les logiciels statistiques prennent généralement par défaut la valeur 0.5.

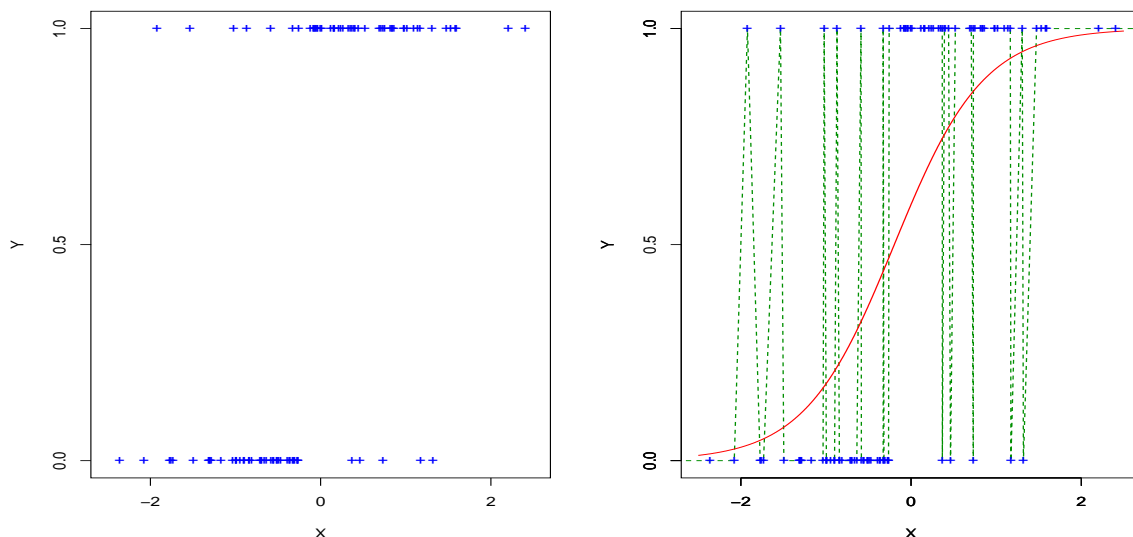


FIGURE 3.1 – Gauche : Représentation des observations (gauche). Droite : Représentation des modèles saturés (pointillés) et logistique (trait plein).

Définition 3.2

Etant donné $\hat{g} : \mathbb{R}^{p+1} \rightarrow \{0, 1\}$ une règle de prévision construite à partir d'un échantillon $\mathcal{D}_n = (X_1, Y_1), \dots, (X_n, Y_n)$, on définit la probabilité d'erreur de \hat{g} par

$$L(\hat{g}) = \mathbf{P}(\hat{g}(X) \neq Y | \mathcal{D}_n).$$

Etant donné K règles $\hat{g}_k, k = 1, \dots, K$, l'approche consiste à

1. estimer les probabilités d'erreur de toutes les règles candidates à l'aide de l'échantillon ;
2. choisir la règle qui possède la plus petite estimation.

Toute la difficulté est de trouver un "bon" estimateur de $L(\hat{g})$. Une première idée serait d'utiliser

$$\hat{L}(\hat{g}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{\hat{g}(X_i) \neq Y_i\}}.$$

Comme le modèle saturé ajuste de manière "parfaite" les données, on a $\hat{L}(\hat{g}_{sat}) = 0$. Cette procédure conduirait à sélectionner de manière quasi-systématique le modèle saturé. La faiblesse de cette approche tient du fait que le même échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ est utilisé pour :

- construire le modèle (ou la règle) ;
- estimer la probabilité d'erreur.

Ceci introduit un biais dans l'estimation de la probabilité d'erreur. La procédure apprentissage-validation s'affranchit de ce problème en séparant de manière aléatoire les données $(X_1, Y_1), \dots, (X_n, Y_n)$ en deux parties distinctes :

- $\mathcal{D}_\ell = \{(X_i, Y_i), i \in \mathcal{I}_\ell\}$ un échantillon d'apprentissage de taille ℓ ;
- $\mathcal{D}_m = \{(X_i, Y_i), i \in \mathcal{I}_m\}$ un échantillon de validation de taille m tel que $\ell + m = n$,

où $\mathcal{I}_\ell \cup \mathcal{I}_m = \{1, \dots, n\}$ et $\mathcal{I}_\ell \cap \mathcal{I}_m = \emptyset$. L'échantillon d'apprentissage est utilisé pour construire les modèles concurrents (pour estimer les paramètres des différents modèles logistiques envisagés) à partir desquels on définit les règles de prévision candidates. L'échantillon de validation est ensuite utilisé pour estimer les probabilités d'erreur de chaque règle (voir figure 3.2).

Plus précisément, une fois les paramètres des modèles candidats estimés sur l'échantillon d'apprentissage, on construit les règles de prévision $\hat{g}_k, k = 1, \dots, K$ associées à ces modèles. La probabilité

d'erreur d'une règle \hat{g}_k

$$L(\hat{g}_k) = \mathbf{P}(\hat{g}_k(X) \neq Y | \mathcal{D}_\ell)$$

est ensuite estimée à l'aide de l'échantillon de validation :

$$\hat{L}(\hat{g}_k) = \frac{1}{m} \sum_{i \in \mathcal{I}_m} \mathbf{1}_{\{\hat{g}_k(X_i) \neq Y_i\}}.$$

On choisira la règle qui minimise $\hat{L}(\hat{g}_k)$. Il est facile de voir que $\hat{L}(\hat{g}_k)$ est un estimateur sans biais de $L(\hat{g}_k)$ puisque, conditionnellement à \mathcal{D}_ℓ , $m\hat{L}(\hat{g}_k)$ suit une loi Binomiale $Bin(m, L(\hat{g}_k))$.

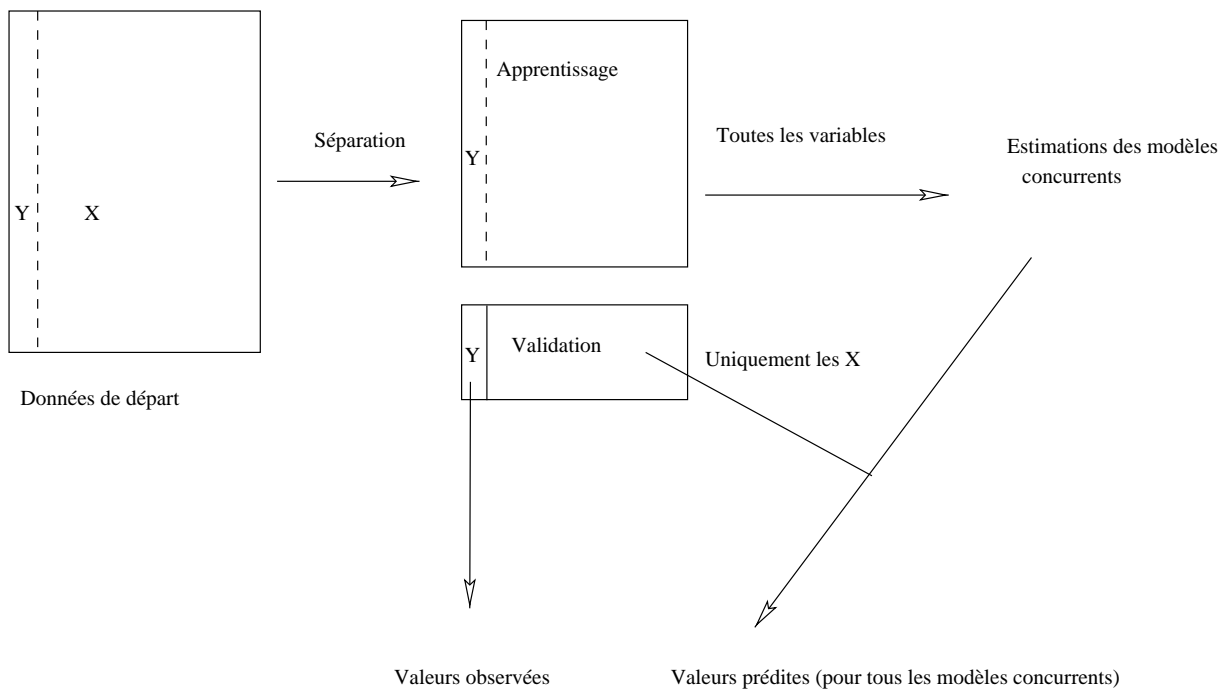


FIGURE 3.2 – Procédure d'apprentissage/validation.

Le tableau 3.1 compare les probabilités d'erreur estimées sur les données de l'exemple de la figure 3.1. La procédure qui utilise un seul échantillon pour estimer ces probabilités va sélectionner le modèle saturé, ce n'est pas le cas de la procédure Apprentissage-Validation qui fournit des estimations des taux d'erreurs plus précises et qui sélectionnera le modèle logistique.

	Saturé	Logistique
Sans AV	0	0.146
Avec AV	0.244	0.160

TABLE 3.1 – Pourcentages de mal classés des modèles saturés et logistique de l'exemple de la Figure 3.1 avec et sans la procédure apprentissage-validation (les deux échantillons sont de même taille).

Cette procédure semble la plus indiquée pour choisir un modèle. Il faut néanmoins la nuancer car elle requiert beaucoup de données

- dans l'échantillon d'apprentissage pour estimer convenablement les paramètres du modèle et ainsi ne pas trop pénaliser les modèles avec beaucoup de variables dont les coefficients seront moins bien estimés ;
- dans l'échantillon de validation pour bien évaluer la probabilité d'erreur des règles.

De plus il n'existe pas de règle pour choisir les tailles des deux échantillons.

3.1.5 Validation croisée

Lorsque l'on n'a pas assez de données pour l'apprentissage/validation, on peut avoir recours à une procédure de validation croisée. Le principe est de "moyenner" le pourcentage de mal classés à l'aide de plusieurs découpages de l'échantillon. Plus précisément, on divise l'échantillon initial en K sous échantillons E_k de même taille et on effectue K procédures apprentissage-validation pour lesquelles :

- l'échantillon test sera constitué d'une division E_k ;
- l'échantillon d'apprentissage sera constitué de l'ensemble des autres divisions $E - E_k$ (voir figure 3.3).

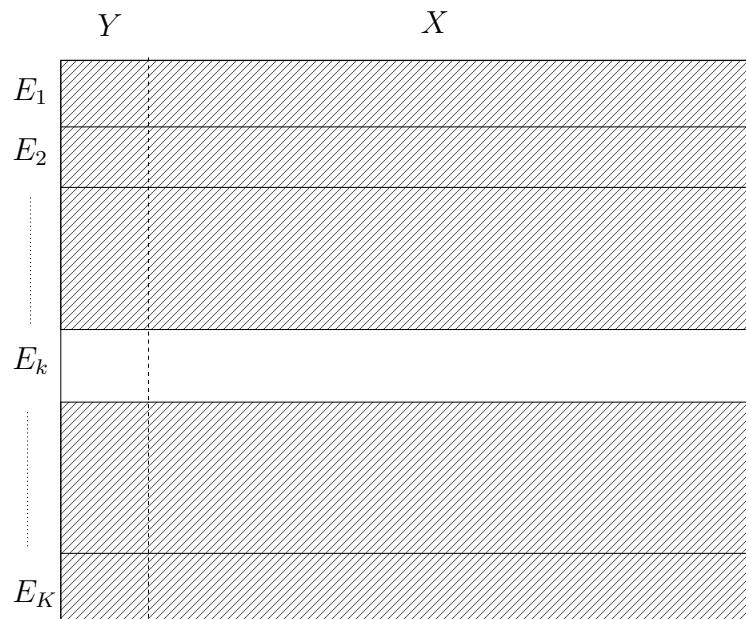


FIGURE 3.3 – Découpage de l'échantillon pour la validation croisée. L'échantillon d'apprentissage correspond à la partie hachurée.

On obtient ainsi une prévision pour chaque individu de la division E_k et une fois les K procédures apprentissage-validation effectuées, on a une prévision pour tous les individus de l'échantillon. Il suffit alors de comparer ces prévisions aux valeurs observées pour obtenir une estimation de la probabilité d'erreur de la règle. Le modèle retenu sera le modèle qui conduit à l'estimation minimale.

Bien entendu le choix du nombre K parties n'est pas anodin.

- Plus K est faible, plus la capacité de prévision sera évaluée dans de nombreux cas puisque le nombre d'observations dans la validation sera élevé, mais moins l'estimation sera précise puisque moins de données seront utilisées pour estimer les paramètres du modèle ;
- Au contraire, un K élevé conduit à peu d'observations dans la validation et donc à une plus grande variance dans l'estimation de la probabilité d'erreur.

Remarque

Sous R, la librairie `boot` permet d'estimer le pourcentage de mal classées par validation croisée. Si, par exemple, on considère le modèle composé des 6 variables explicatives sur les données du cancer de la prostate, on obtient :

```
> modele <- glm(Y~.,data=donnees,family=binomial)
> library(boot)
> cout <- fonction(Y_obs,prevision_prob){
```

```
+         return(mean(abs(Y_obs-prevision_prob)>0.5))}
> cv.glm(donnees,modele,cout)$delta[1]
1
0.3396226
```

3.1.6 Sélection automatique

Les procédures que nous venons d'étudier permettent de sélectionner un modèle à partir d'une famille de modèles donnée. Une autre approche de la sélection de modèle consiste à chercher parmi les variables $\mathbf{X}_1, \dots, \mathbf{X}_p$, celles qui "expliquent le mieux" Y . Par exemple, pour la régression logistique, nous pourrions nous poser le problème de chercher le meilleur sous-ensemble des p variables explicatives pour un critère C donné (AIC , BIC ...). Le nombre de sous ensembles de p variables étant 2^p , nous serions en présence de 2^p modèles logistiques possibles, c'est-à-dire 2^p modèles différents. Bien entendu, nous sélectionnerions le modèle qui optimiserait le critère C . Cependant, dans de nombreuses situations, p est grand et par conséquent le nombre de modèles considérés est "très grand". Les algorithmes d'optimisation du critère C deviennent très coûteux en temps de calcul. On préfère alors souvent utiliser des méthodes de recherche *pas à pas*.

Recherche pas à pas, méthode ascendante (forward selection)

A chaque pas, une variable est ajoutée au modèle.

- Si la méthode ascendante utilise un test de déviance, nous rajoutons la variable \mathbf{X}_j dont la valeur p (probabilité critique) associée à la statistique de test de déviance qui compare les 2 modèles est minimale. Nous nous arrêtons lorsque toutes les variables sont intégrées ou lorsque la valeur p est plus grande qu'une valeur seuil.
- Si la méthode ascendante utilise un critère de choix, nous ajoutons la variable \mathbf{X}_j dont l'ajout au modèle conduit à l'optimisation la plus grande du critère de choix. Nous nous arrêtons lorsque toutes les variables sont intégrées ou lorsque qu'aucune variable ne permet l'optimisation du critère de choix (voir aussi Figure 3.4).

Recherche pas à pas, méthode descendante (backward selection)

A la première étape toutes les variables sont intégrées au modèle.

- Si la méthode descendante utilise un test de déviance, nous éliminons ensuite la variable \mathbf{X}_j dont la valeur p associée à la statistique de test de déviance est la plus grande. Nous nous arrêtons lorsque toutes les variables sont retirées du modèle ou lorsque la valeur p est plus petite qu'une valeur seuil.
- Si la méthode descendante utilise un critère de choix, nous retirons la variable \mathbf{X}_j dont le retrait du modèle conduit à l'augmentation la plus grande du critère de choix. Nous nous arrêtons lorsque toutes les variables sont retirées ou lorsque qu'aucune variable ne permet l'augmentation du critère de choix.

Recherche pas à pas, méthode progressive (stepwise selection)

Idem que l'ascendante, sauf que l'on peut éliminer des variables déjà introduites. En effet, il peut arriver que des variables introduites au début de l'algorithme ne soient plus significatives après introduction de nouvelles variables. Remarquons qu'en général la variable "constante" est toujours présente dans le modèle.

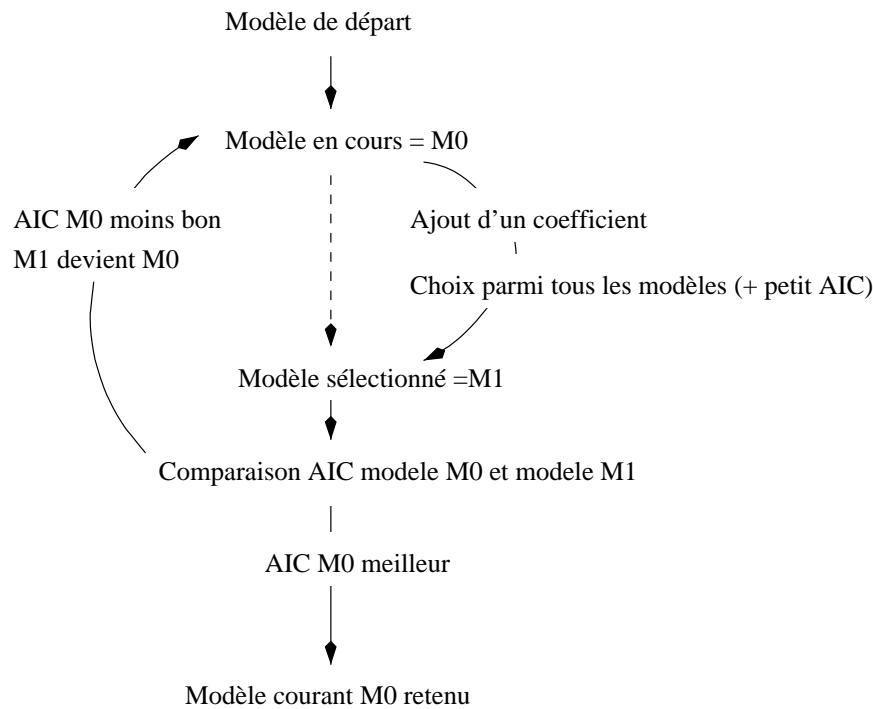


FIGURE 3.4 – Technique ascendante utilisant l'AIC.

Exemple 3.4

Reprenons l'exemple des données du cancer de la prostate. Nous allons sélectionner des modèles par les différentes approches pas à pas.

1. **Méthode ascendante** : le modèle initial est constitué uniquement de la variable `age`.

```

> model_age<-glm(Y~age,data=donnees,family=binomial)
> model_asc<-step(model_age,direction="forward",scope=list(upper=
  formula("Y~(age+acide+rayonx+taille+grade+log.acid.)"))))
> model_asc
  
```

```

Call: glm(formula = Y ~ age + rayonx + taille + log.acid., family = binomial,
  data = donnees)
  
```

Coefficients:

(Intercept)	age	rayonx1	taille1	log.acid.
2.65636	-0.06523	2.08995	1.75652	2.34941

Degrees of Freedom: 52 Total (i.e. Null); 48 Residual

Null Deviance: 70.25

Residual Deviance: 47.68 AIC: 57.68

2. **Méthode descendante** : le modèle initial est ici constitué de toutes les variables (sans interactions).

```

> modelcomplet<-glm(Y~.,data=donnees,family=binomial)
> model_des<-step(modelcomplet,direction="backward")
> model_des
  
```

```

Call: glm(formula = Y ~ acide + rayonx + taille + log.acid., family = binomial,
  data = donnees)
  
```

Coefficients:

(Intercept)	acide	rayonx1	taille1	log.acid.
9.067	-9.862	2.093	1.591	10.410

Degrees of Freedom: 52 Total (i.e. Null); 48 Residual

Null Deviance: 70.25

Residual Deviance: 46.43 AIC: 56.43

3. **Méthode progressive** : le modèle initial est ici constitué de toutes les variables (sans interactions).

```
> model_pro<-step(modelcomplet,direction="both")
> model_pro
```

```
Call: glm(formula = Y ~ acide + rayonx + taille + log.acid., family = binomial,
          data = donnees)
```

Coefficients:

(Intercept)	acide	rayonx1	taille1	log.acid.
9.067	-9.862	2.093	1.591	10.410

Degrees of Freedom: 52 Total (i.e. Null); 48 Residual

Null Deviance: 70.25

Residual Deviance: 46.43 AIC: 56.43

On peut également mettre des variables d'interactions parmi les variables candidates.

```
> model_pro1<-step(modelcomplet,direction="both",scope=list(upper=formula("Y~(age+acide+
  rayonx+taille+grade+log.acid.)^2"),lower=formula("Y~1")))
> model_pro1
```

```
Call: glm(formula = Y ~ acide + rayonx + taille + grade + log.acid. + taille:grade +
  taille:log.acid. + acide:grade, family = binomial,data = donnees)
```

Coefficients:

(Intercept)	acide	rayonx1	taille1
49.385	-49.186	3.135	-2.635
grade1	log.acid.	taille1:grade1	taille1:log.acid.
1.227	53.329	-14.264	-21.719
acide:grade1			
17.629			

Degrees of Freedom: 52 Total (i.e. Null); 44 Residual

Null Deviance: 70.25

Residual Deviance: 26.47 AIC: 44.47

Nous voyons sur cet exemple que suivant le choix de la méthode pas à pas et du modèle initial, les modèles sélectionnés diffèrent. La sélection d'un seul modèle peut s'effectuer en deux temps :

1. On sélectionne un nombre faible (entre 5 et 10 par exemple) de modèles candidats via des algorithmes pas à pas ;
2. On choisit le modèle qui minimise un critère de choix (AIC, BIC, ou minimisation de la probabilité d'erreur).

Une fois le modèle choisi, il est nécessaire de mener une étude plus approfondie de ce dernier qui permettra de le "valider" ou de l'affiner (suppression de points aberrants, analyse des résidus...).

3.2 Validation du modèle

3.2.1 Test d'adéquation de la déviance

Ce test permet de mesurer l'ajustement d'un modèle. L'idée est simple. Nous avons vu que le modèle saturé était le meilleur modèle en terme de qualité d'ajustement. Pour mesurer l'ajustement d'un modèle M_β , nous allons le comparer au modèle saturé en effectuant un test de rapport de vraisemblance (appelé déviance ici). Ce test n'est valable qu'en présence de données répétées.

Généralement, on écrit les hypothèses d'un test entre modèle emboîtés en terme de nullité de certains coefficients (ceux du grand modèle qui ne sont pas dans le petit). Il est difficile de présenter dans un cadre général une telle écriture puisque l'on a pas d'écriture générale de modèle saturé. C'est pourquoi, on commettra l'abus d'écrire les hypothèses sous cette forme

- H_0 : le modèle considéré à M_β de dimension p est adéquat ;
- H_1 : le modèle saturé adéquat.

Cependant, dans la plupart des cas, on pourra écrire les hypothèses de ce test en terme de nullité de certains coefficients du modèle saturé

Exemple 3.5

Considérons l'exemple où on dispose de deux variables explicatives X_1 et X_2 à deux modalités 0 et 1. On souhaite tester le modèle M_β

$$\text{logit } p_\beta(x) = \beta_0 + \beta_1 \mathbf{1}_{x_1=1} + \beta_2 \mathbf{1}_{x_2=1}$$

contre le modèle saturé qui (sous réserve que tous les croisements entre X_1 et X_2 soient observés) va ici s'écrire

$$\text{logit } p_{sat}(x) = \gamma_0 + \gamma_1 \mathbf{1}_{x_1=1} + \gamma_2 \mathbf{1}_{x_2=1} + \gamma_3 \mathbf{1}_{x_1=1} \mathbf{1}_{x_2=1}.$$

Sur cet exemple, les hypothèses du test de deviance vont s'écrire $H_0 : \gamma_3 = 0$ contre $H_1 : \gamma_3 \neq 0$.

Le test de la déviance compare le modèle saturé au modèle considéré au moyen de la déviance. Nous savons que

- si la déviance est grande, alors le modèle considéré est loin du modèle saturé et que par conséquent il n'ajuste pas bien les données ;
- Par contre si la déviance est proche de 0, le modèle considéré sera adéquat.

Pour quantifier cette notion de "proche de 0" et de "grande déviance", la loi de la déviance sous H_0 (le modèle considéré est le vrai modèle) va nous être utile. En effet si H_0 est vraie, le modèle considéré est vrai par définition. La déviance sera répartie sur \mathbb{R}^+ mais avec plus de chance d'être proche de 0. Par contre si H_0 n'est pas vraie la déviance sera répartie sur \mathbb{R}^+ mais avec plus de chance d'être éloignée de 0. Il nous faut donc connaître la loi de la déviance sous H_0 .

En présence de données répétées, si le nombre de répétitions n_t de chaque point du design tend vers ∞ , alors sous H_0 la déviance $D_{M_\beta} = 2(\mathcal{L}_{sat} - \mathcal{L}_n(\hat{\beta}))$ converge en loi vers un χ_{T-p}^2 où p est la dimension de M_β . Ainsi, on niveau α , on rejettera H_0 si la valeur observée de D_{M_β} est supérieure au quantile d'ordre $1 - \alpha$ de la loi χ_{T-p}^2 (voir Figure 3.5).

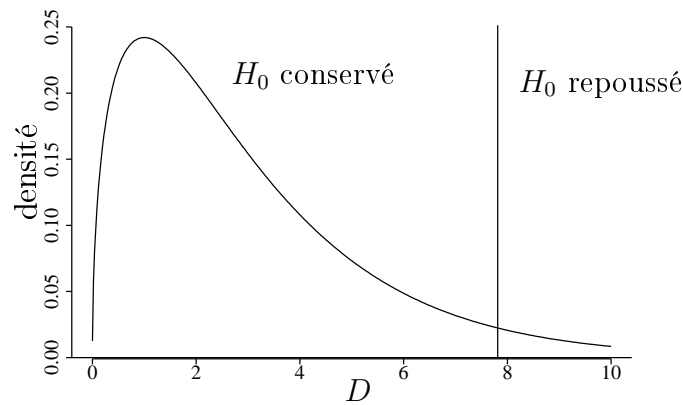


FIGURE 3.5 – Test de déviance, la droite verticale représente le seuil de rejet $D_c = q_{1-\alpha}(T - p)$.

Remarque

La validité de la loi et donc du test n'est qu'asymptotique, il est nécessaire d'avoir un peu de recul quant aux conclusions. Ce test ne peut être utilisé qu'en présence de données répétées. En effet, l'approximation de la loi de la déviance par une loi du χ^2 est d'autant plus valable lorsque le nombre de répétitions aux points du design est grand. En présence de données individuelles (aucune répétition sur les points du design), D_{M_β} ne suit pas une loi du χ^2 : le test d'adéquation d'Hosmer Lemeshow est alors conseillé.

3.2.2 Test d'Hosmer Lemeshow

Ce test permet de vérifier l'adéquation d'un modèle M_β en présence de données individuelles. L'approche consiste à se rapprocher du cas de données répétées en créant ces répétitions. Le test s'effectue de la manière suivante (voir Hosmer & Lemeshow (2000), chapitre 5 pour plus de précisions).

1. Les probabilités $\hat{p}_\beta(x_i)$ sont ordonnées par ordre croissant ($\hat{p}_\beta(x_i)$ est la probabilité $\mathbf{P}_\beta(Y = 1|X = x_i)$ estimée par le modèle) ;
2. Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand). On note
 - m_k les effectifs du groupe k ;
 - o_k le nombre de succès ($Y = 1$) observé dans le groupe k ;
 - μ_k la moyenne des $\hat{p}_\beta(x_i)$ dans le groupe k .

La statistique de test est alors

$$C^2 = \sum_{k=1}^K \frac{(o_k - m_k \mu_k)^2}{m_k \mu_k (1 - \mu_k)}.$$

Le test se conduit de manière identique au test de déviance, la statistique C^2 suivant approximativement sous H_0 un χ^2 à $K - 1$ degrés de liberté.

3.2.3 Analyse des résidus

On se donne M_β un modèle logistique. Pour simplifier les notations on écrira $p_i = p_\beta(x_i)$ et $\hat{p}_i = \hat{p}_\beta(x_i)$.

Les différents types de résidus

A l'image de la régression plusieurs types de résidus sont proposés par les logiciels. Le premier, le plus simple à calculer est tout simplement $Y_i - \hat{p}_i$. Ces résidus sont appelés *résidus bruts*. Ils permettent de mesurer l'ajustement du modèle sur chaque observation. Ces résidus n'ayant pas la même variance, ils sont difficiles à comparer. En effet, on rappelle que $\mathbf{V}_\beta(Y|X = x_i) = p_i(1 - p_i)$. Par conséquent, la variance de tels résidus risquent d'être élevées pour des valeurs de p_i proches de 1/2. Un moyen de pallier à cette difficulté est de considérer les *résidus de Pearson*

$$X_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}. \quad (3.1)$$

Par définition on standardise les résidus par la variance théorique de Y_i .

Remarque

En présence de données répétées ces résidus s'écrivent

$$X_t = \frac{Y_t - n_t \hat{p}_t}{\sqrt{\hat{n}_t \hat{p}_t (1 - \hat{p}_t)}}.$$

Pour n_t grand, $\varepsilon_t \sim \mathcal{N}(0, 1)$ si le modèle est valide et la statistique $\sum_{t=1}^T X_t^2 \sim \chi_{T-p}^2$ où p est la dimension du modèle. Cette statistique est appelée Statistique de Pearson, d'où le nom de résidu de Pearson. Cette statistique peut être utilisée pour construire un test d'adequation du modèle en question en présence de données répétées.

Cependant, comme \hat{p}_i est aléatoire, il est évident que $\mathbf{V}_\beta(Y_i - p_i) \neq \mathbf{V}_\beta(Y_i - \hat{p}_i)$. En effet, en notant

$$\begin{cases} \varepsilon_i = Y_i - p_i \\ \hat{\varepsilon}_i = Y_i - \hat{p}_i \end{cases}$$

on a

Hypothèses	Réalité
$\mathbf{E}(\varepsilon_i) = 0$	$\mathbf{E}(\hat{\varepsilon}_i) \simeq 0$
$\mathbf{V}(\varepsilon_i) = p_i(1 - p_i)$	$\mathbf{V}(\hat{\varepsilon}_i) \simeq p_i(1 - p_i)(1 - h_{ii})$

où h_{ii} est l'élément de la $i^{\text{ème}}$ ligne et de la $i^{\text{ème}}$ colonne de la matrice $H = \mathbf{X}(\mathbf{X}'\mathbf{W}_{\hat{\beta}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_{\hat{\beta}}$.

Il est par conséquent intéressant de considérer la version standardisée des résidus de Pearson

$$\frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)(1 - h_{ii})}}.$$

Ces résidus seront en effet plus facile à analyser (leur distribution étant "presque" centrée réduite).

Les *résidus de déviance* sont définis par

$$rd_i = \text{signe}(Y_i - \hat{p}_i) \sqrt{2\mathcal{L}_1(Y_i, \hat{\beta})} = \text{signe}(Y_i - \hat{p}_i) \sqrt{2Y_i \log \hat{p}_i + (1 - Y_i) \log(1 - \hat{p}_i)}.$$

Là encore pour tenir compte de la variabilité ces résidus sont standardisés :

$$\frac{rd_i}{\sqrt{1 - h_{ii}}}$$

Ces deux types de résidus de déviance sont ceux qui sont en général conseillés.

Remarque

1. En présence de données individuelles, nous n'avons pas d'informations précises sur la loi de ces résidus. On sait juste qu'ils sont approximativement d'espérance nulle et de variance 1.
2. En présence de données répétées, les résidus de déviance s'écrivent

$$rd_t = \sqrt{2 \left[y_t \log \frac{\bar{y}_t}{\hat{p}_t} + (n_t - y_t) \log \frac{n_t - y_t}{n_t - \hat{p}_t} \right]}$$

on retrouve bien que la déviance est égale à $\sum_{t=1}^T rd_t^2$. Par conséquent, lorsque les nombres de répétitions n_t sont grands, les résidus de déviance suivent approximativement une loi $\mathcal{N}(0, 1)$. On peut ainsi les analyser de la même manière quand dans le modèle linéaire.

3. Tout comme dans le modèle linéaire, on peut également étudier une version studentisé de ces résidus en estimant les quantités \hat{p}_i par validation croisée.

Examen des résidus

Index plot Pour le modèle logistique les résidus de déviance sont souvent préférés. De nombreuses études expérimentales ont montré qu'ils approchent mieux la loi normale que les résidus de Pearson. Pour cette raison ces résidus prennent généralement des valeurs qui varient entre -2 et 2. Nous pourrions construire un *index plot* pour détecter des valeurs aberrantes. Ce graphique ordonne les résidus en fonction du numéro de leur observation. Les points pour lesquels on observe un résidu élevé (hors de $[-2, 2]$ par exemple) devront faire l'objet d'une étude approfondie.

Exemple 3.6

Voici un exemple de calcul des résidus avec R sur les données du cancer de la prostate.

```
> model<-glm(Y~rayonx+taille+grade+log.acid.+taille:grade+grade:log.acid.,data=donnees,
              family=binomial)
> res_dev <- residuals(model) #residus de deviance
> res_pear <- residuals(model,type="pearson") #residus de Pearson
> res_dev_stand <- rstandard(model) #residu de deviance standardises
> H <- influence(model)$hat #diagonale de la hat matrix
> res_pear_stand <- res_pear/sqrt(1-H) #residu de Pearson standardises

> plot(rstudent(model),type="p",cex=0.5,ylab="Résidus studentisés par VC")
> abline(h=c(-2,2))
```

Graphique prédiction linéaire/résidus Ce graphique qui représente $X'\hat{\beta}$ en abscisse et $\hat{\varepsilon}$ en ordonné permet de détecter les valeurs aberrantes mais aussi les structurations suspectes. Si une structuration suspecte apparaît, il sera peut être adéquat d'ajouter une nouvelle variable afin de prendre en compte cette structuration. Dans le cas des données individuelles ce type de graphique donne toujours des structurations (Figure 3.7) et n'est donc pas à conseiller.

```
> plot(predict(model),rstudent(model),type="p",cex=0.5,xlab="prévision linéaire",
        ylab="Résidus studentisés par VC")
```

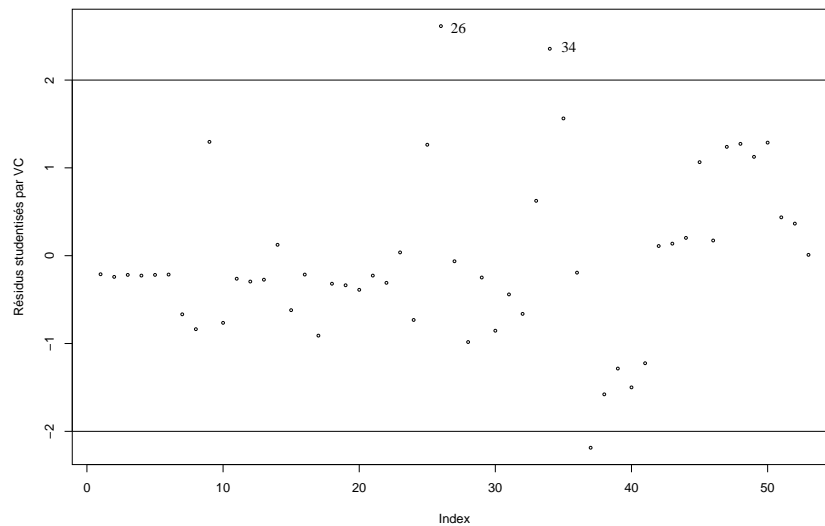


FIGURE 3.6 – Index plot des résidus de déviance studentisés.

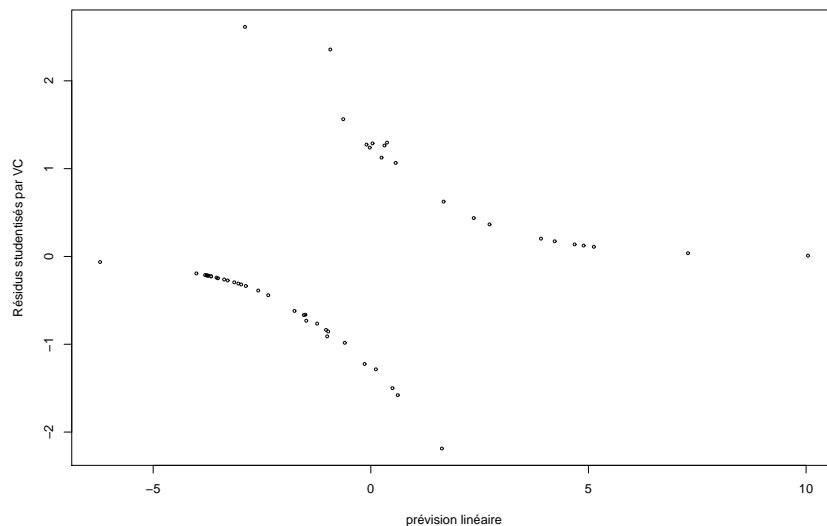


FIGURE 3.7 – Graphique prédiction/résidus pour un modèle logistique

Résidus partiels Les résidus partiels sont définis par

$$\hat{\varepsilon}_{.j}^P = \frac{Y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)} + \hat{\beta}_j X_{.j}$$

L'analyse consiste à tracer pour toutes les variables j les points avec en abscisse la variable j et en ordonnée les résidus partiels. Si le tracé est linéaire alors tout est normal. Si par contre une tendance non linéaire se dégage, il faut remplacer la variable j par une fonction de celle ci donnant la même tendance que celle observée.

Exemple 3.7

Nous reprenons l'exemple du modèle précédent et traçons sur la figure 3.8 les résidus partiels pour la variable `log.acid..`

```
> residpartiels<-resid(model,type="partial")
> prov<-loess(residpartiels[,"log.acid."]~donnees$log.acid)
> ordre<-order(donnees$log.acid.)
> plot(donnees$log.acid.,residpartiels[,"log.acid."],type="p",cex=0.5,xlab="",ylab="")
```

```
> matlines(donnees$log.acid. [ordre] ,predict (prov) [ordre])
> abline(lsfilt(donnees$log.acid.,residpartiels[,"log.acid."]),lty=2)
```

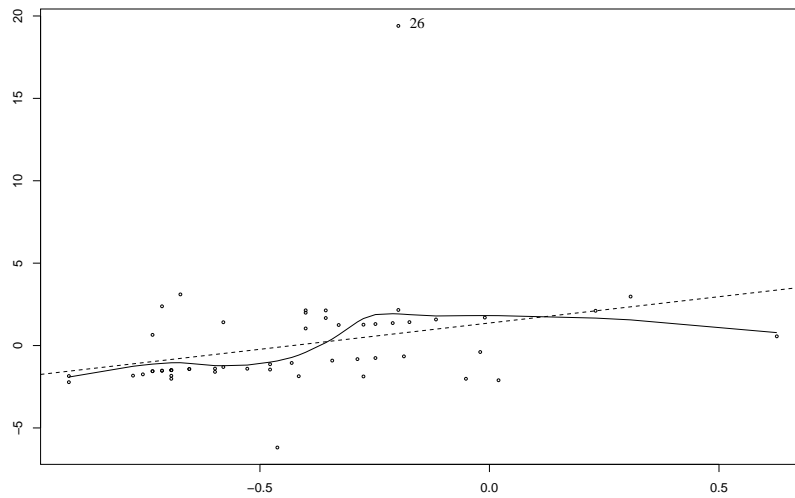


FIGURE 3.8 – Résidus partiels pour la variable `log.acid.`, le trait continu représente le résumé lissé des données par l’estimateur `loess`, le trait discontinu représente l’estimateur linéaire par moindres carrés.

Même si l’ajustement n’est pas “parfaitement linéaire”, la figure 3.8 montre qu’aucune transformation n’est nécessaire, les résidus partiels étant répartis le long de la droite ajustée.

Exemple 3.8

Étudions ici les résidus partiels de la variable `age` pour l’exemple sur les données de panne du TP 1 (voir exemple 2.6).

```
> model <- glm(panne~.,data=donnees,family=binomial)
> residpartiel <- residuals(model,type="partial")
> plot(donnees$age,residpartiel[,"age"],cex=0.5)
> est <- loess(residpartiel[,"age"]~donnees$age)
> ordre <- order(donnees$age)
> matlines(donnees$age[ordre] ,predict(est) [ordre])
> abline(lsfilt(donnees$age,residpartiel[,"age"]),lty=2)
```

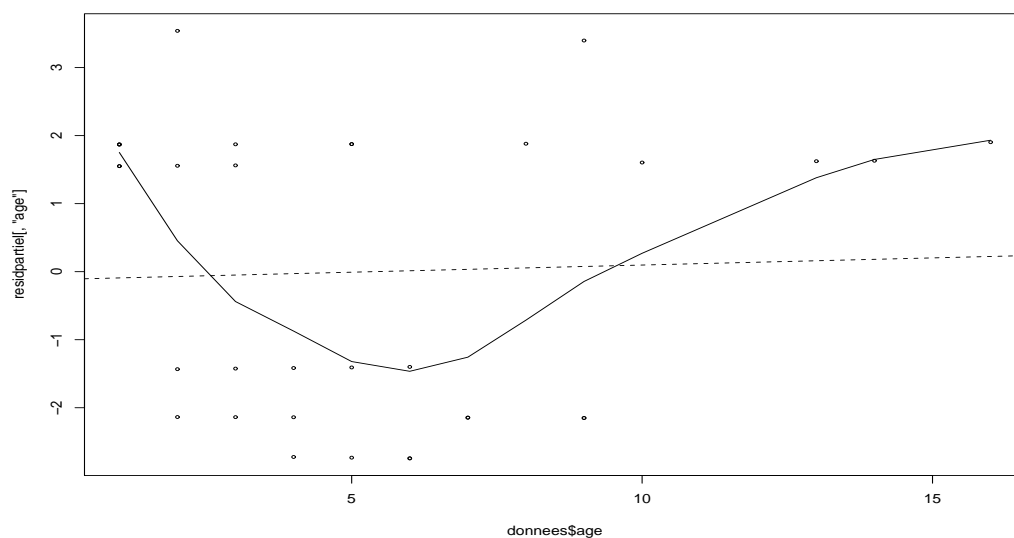


FIGURE 3.9 – Résidus partiels pour la variable `age`.

La figure 3.9 suggère de prendre en compte la variable age^2 dans le modèle.

Mallows (1986) propose d'utiliser les résidus partiels augmentés qui dans certaines situations permettent de mieux dégager cette tendance. Les résidus partiels augmentés pour la $j^{\text{ème}}$ variable nécessitent un nouveau modèle logistique identique mis à part le fait qu'une variable explicative supplémentaire est ajoutée : $\mathbf{X}_{p+1} = \mathbf{X}_j^2$ la $j^{\text{ème}}$ variable élevée au carré. Le nouveau vecteur de coefficient β du modèle est estimé et les résidus partiels sont alors définis comme

$$\hat{\varepsilon}_{.j}^{PA} = \frac{Y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)} + \hat{\beta}_j X_{.j} + \hat{\beta}_{p+1} X_{.j}^2.$$

L'analyse des diagrammes est identique à ceux des résidus partiels. Pour une analyse plus complète sur l'utilisation des résidus, on pourra se reporter au chapitre 5 de l'ouvrage de Collet (2003).

3.2.4 Points leviers et points influents

Ces notions sont analogues à celles du modèle linéaire (voir Cornillon & Matzner-Løber (2007), chapitre 4).

Points leviers

Par définition les points leviers sont les points du design qui déterminent très fortement leur propre estimation. Nous avons vu que l'algorithme d'estimation des paramètres effectuée à chaque étape une régression linéaire et s'arrête lorsque le processus devient stationnaire :

$$\hat{\beta} = (\mathbf{X}'W_{\hat{\beta}}\mathbf{X})^{-1}\mathbf{X}'W_{\hat{\beta}}z,$$

et la prédiction linéaire est

$$\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'W_{\hat{\beta}}\mathbf{X})^{-1}\mathbf{X}'W_{\hat{\beta}}z = Hz,$$

où H est une matrice de projection selon la métrique $W_{\hat{\beta}}$. Comme nous transformons $\mathbf{X}\hat{\beta}$ par une fonction monotone, des $\mathbf{X}\hat{\beta}$ extrêmes entraînent des valeurs de \hat{p} extrêmes. Nous allons donc utiliser la même méthode de diagnostic que celle de la régression simple avec une nouvelle matrice de projection H . Pour la $i^{\text{ème}}$ prédiction linéaire nous avons

$$[\mathbf{X}\hat{\beta}]_i = H_{ii}z_i + \sum_{j \neq i} H_{ij}z_j.$$

Si H_{ii} est grand relativement aux H_{ij} , $j \neq i$ alors la $i^{\text{ème}}$ observation contribue fortement à la construction de $[\mathbf{X}\hat{\beta}]_i$. On dira que le "poinds" de l'observation i sur sa propre estimation vaut H_{ii} .

Comme H est un projecteur nous savons que $0 \leq H_{ii} \leq 1$. Nous avons les cas extrêmes suivants :

- si $H_{ii} = 1$, \hat{p}_i est entièrement déterminé par Y_i car $H_{ij} = 0$ pour tout j .
- si $H_{ii} = 0$, Y_i n'a pas d'influence sur \hat{p}_i .

La trace d'un projecteur étant égale à la dimension du sous espace dans lequel on projette, on a $tr(H) = \sum_i H_{ii} = p + 1$. Donc en moyenne H_{ii} vaut $(p + 1)/n$. Pour dire que la valeur de H_{ii} contribue trop fortement à la construction de \hat{p}_i , il faut un seuil au delà duquel le point est un

point levier. Par habitude, si $H_{ii} > 2p/n$ ou si $H_{ii} > 3p/n$ alors le $i^{\text{ème}}$ point est déclaré comme un point levier.

En pratique un tracé de H_{ii} est effectué et l'on cherche les points dont le H_{ii} est supérieur à $3(p+1)/n$ ou $2(p+1)/n$. Ces points sont leviers et leur valeur influe fortement sur leur propre prévision.

Pour le modèle de l'exemple 3.6, on a

```
> p<-length(model$coefficients)
> n<-nrow(donnees)
> plot(influence(model)$hat,type="h",ylab="hii")
> seuil1<-3*p/n
> abline(h=seuil1,col=1,lty=2)
> seuil2<-2*p/n
> abline(h=seuil2,col=1,lty=3)
```

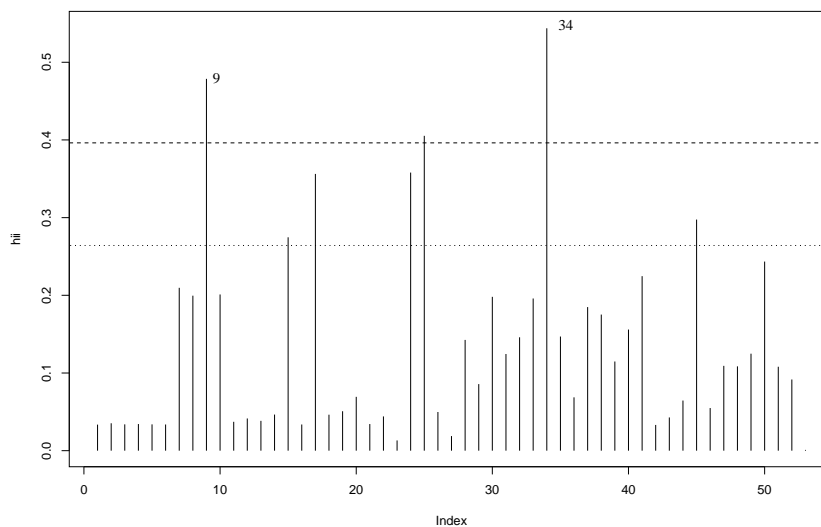


FIGURE 3.10 – Points leviers.

Points influents

Les points influents sont des points qui influent sur le modèle de telle sorte que si on les enlève, alors l'estimation des coefficients sera fortement changée. La mesure la plus classique d'influence est la distance de Cook. Il s'agit d'une distance entre le coefficient estimé avec toutes les observations et celui estimé avec toutes les observations sauf une. La distance de Cook pour l'individu i est définie par

$$D_i = \frac{1}{p+1} (\hat{\beta}_{(i)} - \hat{\beta})' \mathbf{X}' W_{\hat{\beta}} \mathbf{X} (\hat{\beta}_{(i)} - \hat{\beta}) \approx \frac{r_{Pi}^2 H_{ii}}{(p+1)(1-H_{ii})^2},$$

où r_{Pi} est le résidu de Pearson pour le $i^{\text{ème}}$ individu.

Les distances de Cook sont généralement représentées comme sur la figure 3.11. Si une distance se révèle grande par rapport aux autres, alors ce point sera considéré comme influent. Il convient alors de comprendre pourquoi il est influent, soit

- il est levier ;
- il est aberrant ;
- (les deux !)

Dans tous les cas il convient de comprendre si une erreur de mesure, une différence dans la population des individus est à l'origine de ce phénomène. Eventuellement pour obtenir des conclusions robustes il sera bon de refaire l'analyse sans ce(s) point(s).

Toujours pour le modèle de l'exemple 3.6, on représente les distances de Cook avec la commande

```
> plot(cooks.distance(model), type="h", ylab="Distance de Cook")
```

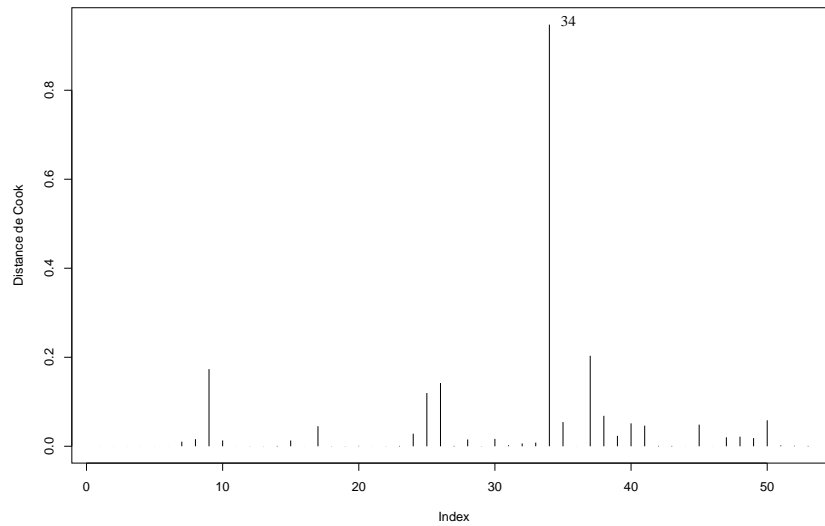


FIGURE 3.11 – Distances de Cook.

Chapitre 4

Modèle logistique multi-classes

Nous traitons dans ce chapitre le cas où la variable à expliquer Y prend plus de deux modalités. Pour simplifier les notations, nous supposons que Y peut prendre K valeurs $1, \dots, K$. Nous cherchons donc à expliquer Y par le vecteur $X = (1, \mathbf{X}_1, \dots, \mathbf{X}_p)'$ des variables explicatives (X_j étant qualitatives, quantitatives ou pouvant représenter des interactions). Nous distinguons deux cas :

- les modalités de Y sont ordonnées : il existe une hiérarchie naturelle entre elles. Par exemple le degré de satisfaction relativement à un produit, le degré d'adhésion à une opinion... En biostatistique, il peut s'agir d'un diagnostic sur l'état de santé (très bonne, bonne, moyenne, mauvais santé), sur le stade d'évolution d'une maladie, ou encore sur la taille ou la nature d'une tumeur (tumeur absente, bénigne, ou maligne). On parle dans ce cas de *modèle polytomique ordonné*;
- il n'existe pas de relation d'ordre sur les modalités de Y , la variable à expliquer est purement nominale : accord pour un prêt (oui, non, examen du dossier). On parle dans ce cas de *modèle polytomique nominal* où de *modèle polytomique multinomial*.

Tout comme dans le cas binaire, nous cherchons ici à modéliser la loi de $Y|X = x$. Si on suppose que Y prend ses valeurs dans $\{1, \dots, K\}$, le problème va consister à mettre une forme paramétrique sur les probabilités $\pi_j = \mathbf{P}(Y = j|X = x), j = 1, \dots, K$. Que les données soient individuelles $\{(y_i, x_i), i = 1, \dots, n\}$ ou répétées $\{(y_{it}, x_t), i = 1, \dots, n_t, t = 1, \dots, T\}$, on supposera que les observations sont indépendantes. Dans ce chapitre, nous n'aborderons pas en détails les méthodes d'estimations ainsi que les comportements asymptotiques des estimateurs. Une fois les conditions d'identifiabilité des modèles vérifiées, l'estimation des paramètres se fera par maximum de vraisemblance. Sous des conditions du même type que celles présentées dans le cas binaire, on admettra les résultats classiques sur le comportement asymptotique des estimateurs : variance asymptotique se déduisant de la matrice d'information de Fisher, normalité asymptotique de l'estimateur du maximum de vraisemblance, tests de Wald, du rapport de vraisemblance et du score...

4.1 Le modèle saturé ou modèle multinomial

Tout comme dans le cas binaire, ce modèle présente un intérêt essentiellement dans le cas de données répétées. Ici encore, le modèle saturé ne met pas de contrainte sur la forme des probabilités π_j . Il est donc défini par l'ensemble des probabilités

$$\pi_{jt} = \mathbf{P}(Y_{it} = j|X = x_t), i = 1, \dots, n_t; t = 1, \dots, T, j = 1, \dots, K - 1$$

(j varie de 1 à $K - 1$ car $\forall t, \sum_{j=1}^K \pi_{jt} = 1$). Le modèle saturé est donc de dimension $T(K - 1)$. Les estimateurs du maximum de vraisemblance sont donnés par

$$\hat{\pi}_{jt} = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{1}_{Y_{it}=j},$$

c'est-à-dire par la proportion des Y_{it} prenant la valeur j au point x_t . On note $\pi(t) = (\pi_{1t}, \dots, \pi_{(K-1)t})'$.

Proposition 4.1

1. $\hat{\pi}(t)$ est un estimateur sans biais de $\pi(t)$.
2. $\mathbf{V}(\hat{\pi}(t)) = \frac{1}{n_t} \Sigma_t$ où $\Sigma_t(j, j) = \pi_{jt}(1 - \pi_{jt})$ et $\Sigma_t(j, \ell) = -\pi_{jt}\pi_{\ell t}$ si $j \neq \ell$.
3. Si $n_t \rightarrow \infty$ alors

$$\sqrt{n_t}(\hat{\pi}(t) - \pi(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_t).$$

Les points 1 et 2 se déduisent directement du fait que $\forall t = 1, \dots, T$ le vecteur aléatoire

$$\left(\sum_{i=1}^{n_t} \mathbf{1}_{Y_{it}=1}, \dots, \sum_{i=1}^{n_t} \mathbf{1}_{Y_{it}=K-1} \right)'$$

suit une loi multinomiale $\mathcal{M}(n_t, \pi(t))$. On comprend bien pourquoi ce modèle présente un intérêt uniquement en présence de données répétées. En effet, en présence de données individuelles ($n_t = 1$), chaque paramètre est estimé à partir d'une seule observation et on obtient de très faibles performances pour les estimateurs.

4.2 Modèle polytomique ordonné

4.2.1 Cas binaire

Nous revenons d'abord au cas où Y est binaire (0 ou 1) et supposons, sans perte de généralité, que nous sommes en présence d'une seule variable explicative X . On introduit ϵ une variable aléatoire centrée et une variable latente (non observée) $Y^* = \tilde{\beta}_0 + \beta_1 x + \epsilon$ telle que $Y|X = x$ vaut 1 lorsque la variable latente Y^* est grande (supérieure à un seuil s) et 0 sinon. Nous obtenons :

$$\mathbf{P}(Y = 1|X = x) = \mathbf{P}(\tilde{\beta}_0 + \beta_1 x + \epsilon > s) = \mathbf{P}(-\epsilon < -s + \tilde{\beta}_0 + \beta_1 x) = F(\beta_0 + \beta_1 x)$$

où F est la fonction de répartition de la variable $-\epsilon$ et $\beta_0 = -s + \tilde{\beta}_0$. Pour finir de spécifier le modèle, il reste à choisir la fonction de répartition F . Si on choisit

$$F(x) = \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{1 + \exp(x)}, \quad (4.1)$$

on obtient le modèle logistique étudié dans les chapitres précédents. Si F est la fonction de répartition associée à la loi normale centrée réduite, nous obtenons alors le modèle **probit** (voir section 1.3 et figure 4.1).

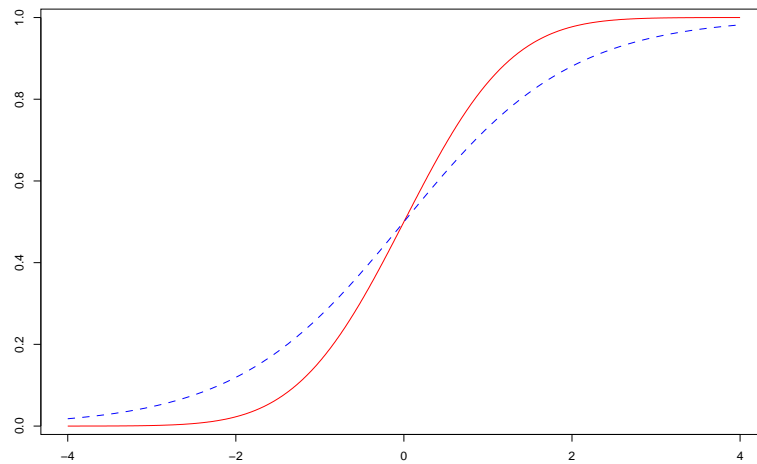


FIGURE 4.1 – Fonctions de répartition des lois normale (trait plein) et logistique (tirets).

4.2.2 Généralisation

Le modèle polytomique ordonné peut être présenté comme une généralisation du modèle dichotomique présenté dans la partie précédente, avec cette fois Y prenant K modalités ordonnées. On se place toujours dans le cas d'une seule variable explicative X . On introduit non plus un seul seuil, mais plusieurs seuils $\alpha_1, \dots, \alpha_{K-1}$ déterministes tels que :

$$(Y|X = x) = \begin{cases} 1 & \text{si } Y^* < \alpha_1 \\ j & \text{si } \alpha_{j-1} \leq Y^* < \alpha_j, \quad j = 2, \dots, K-1 \\ k & \text{si } Y^* \geq \alpha_{K-1} \end{cases} \quad \text{avec } Y^* = \beta_1 x + \epsilon.$$

Le choix de la fonction de répartition logistique (4.1) conduit au modèle :

$$\mathbf{P}_\beta(Y \leq j|X = x) = F(\alpha_j - \beta_1 x), \quad j = 1, \dots, K-1$$

ou encore

$$\text{logit } p_\beta^j(x) = \text{logit } \mathbf{P}_\beta(Y \leq j|X = x) = \alpha_j - \beta_1 x, \quad j = 1, \dots, K-1. \quad (4.2)$$

Si on est en présence de p variables explicatives, le modèle devient

$$\text{logit } p_\beta^j(x) = \alpha_j - \beta_1 \mathbf{x}_1 - \dots - \beta_p \mathbf{x}_p, \quad j = 1, \dots, K-1, \quad (4.3)$$

ou encore

$$p_\beta^j(x) = \frac{\exp(\alpha_j - \beta_1 \mathbf{x}_1 - \dots - \beta_p \mathbf{x}_p)}{1 + \exp(\alpha_j - \beta_1 \mathbf{x}_1 - \dots - \beta_p \mathbf{x}_p)}.$$

Dans ce modèle, seule la constante diffère suivant les différents niveaux de Y . Ce modèle nécessite l'estimation de $p+K-1$ coefficients (p pentes et $K-1$ constantes car $\sum_{j=1}^K \mathbf{P}_\beta(Y = j|X = x) = 1$).

Remarque

Suivant le logiciel les coefficients estimés peuvent différer. La procédure **LOGISTIC** de SAS estime par exemple les pentes $b_j = -\beta_j$. Sous R les fonctions **polr**, **lmr** et **vglm** des bibliothèques **MASS**, **Design** et **VGAM** permettent de construire des modèles logistiques pour expliquer une variable qualitative ordinaire. Il est important de consulter l'aide de la fonction afin de connaître la signification des coefficients estimés.

Exemple 4.1

La fonction **polr** de la librairie MASS utilise un modèle de la forme (4.2) et (4.3). Reprenons le jeu de données du TP2.

On pose à 195 étudiants la question : si vous trouvez un portefeuille dans la rue contenant de l'argent et des papiers :

- vous gardez tout (réponse 1) ;
- vous gardez l'argent et rendez le portefeuille (réponse 2) ;
- vous rendez tout (réponse 3).

On construit alors la variable **WALLET** telle que

- **WALLET**=1 si l'étudiant répond 1 ;
- **WALLET**=2 si l'étudiant répond 2 ;
- **WALLET**=3 si l'étudiant répond 3 ;

Pour chaque étudiant, on note :

- Le sexe (variable **MALE**=1 si homme, 0 si femme) ;
- la nature des études suivies (variable **BUSINESS**= 1 pour les écoles de commerce, 0 pour les autres écoles) ;
- l'existence de punitions passées (variable **PUNISH**=1 si puni seulement à l'école primaire, 2 si puni seulement à l'école primaire et secondaire et 3 si puni seulement à l'école primaire, secondaire et supérieur) ;
- l'explication ou pas par les parents des punitions reçues dans l'enfance (variable **EXPLAIN**=1 si les parents expliquaient, 0 sinon).

On cherche à expliquer la variable **WALLET** par les autres variables. On note $Y = \text{WALLET}$, $X_1 = \text{MALE}$, $X_2 = \text{BUSINESS}$, $X_3 = \text{PUNISH} = 1$ et $X_4 = \text{EXPLAIN}$. Le modèle s'écrit

$$\text{logit } p_{\beta}^j = \alpha_j - \beta_1 \mathbf{1}_{x_1=1} - \beta_2 \mathbf{1}_{x_2=1} - \beta_3 \mathbf{1}_{x_3=2} - \beta_4 \mathbf{1}_{x_3=3} - \beta_5 \mathbf{1}_{x_4=1}.$$

On obtient les estimations avec **polr**.

```
> library(MASS)
> model <- polr(wallet~.,data=donnees)
> model
Call:
polr(formula = wallet ~ ., data = donnees)

Coefficients:
  male1  business1  punish2  punish3  explain1
-1.0598227 -0.7388746 -0.6276423 -1.4030892  1.0518775

Intercepts:
  1|2      2|3
-2.5678520 -0.7890143

Residual Deviance: 307.3349
AIC: 321.3349
```

Tout comme dans le cas binaire, on peut tester la nullité d'un sous ensemble de coefficients à l'aide des statistiques de Wald, du rapport de vraisemblance et du score. Par exemple, on obtient la probabilité critique du test du rapport de vraisemblance pour le test $H_0 : \beta_1 = \dots = \beta_5 = 0$ contre $H_1 : \exists j \in \{1, \dots, 5\}, \beta_j \neq 0$ avec les commandes

```
> model0 <- polr(wallet~1,data=donnees)
> statRV <- -2*(logLik(model0)-logLik(model))
```

```
> 1-pchisq(statRV,df=length(coef(model))-length(coef(model0)))
[1] 1.589732e-08
attr(,"df")
[1] 2
attr(,"class")
[1] "logLik"
```

4.2.3 L'égalité des pentes

Dans le modèle (4.3), les coefficients des variables explicatives sont identiques quel que soit le niveau de Y (on dit qu'il y a égalité des pentes) tandis que les constantes diffèrent. Ainsi, dans ce modèle, quelle que soit la modalité j considérée, une variable explicative donnée a la même influence sur $\mathbf{P}_\beta(Y \leq j|X = x)$. Dans le cas où l'on dispose d'une seule variable explicative X , le modèle peut être représenté par la figure 4.2.

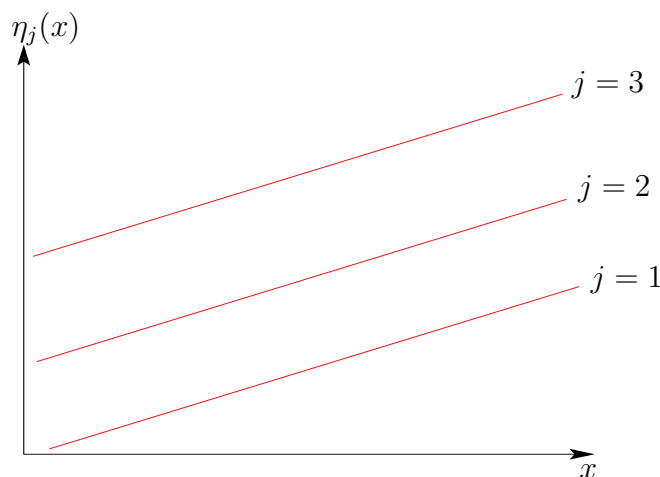


FIGURE 4.2 – Représentation du modèle logit $p_\beta^j(x) = \alpha_j - \beta_1 x = \eta_j(x)$.

Si on envisage des valeurs différentes pour les paramètres de pentes β_m ($m = 1, \dots, p$) alors les droites de la figure 4.2 vont se couper. Il est facile de voir que ceci remet en cause le caractère ordonné des modalités de la variable expliquée. En effet, supposons qu'au delà d'une valeur x_0 , la première droite ($j = 1$) se situe au-dessus de la seconde ($j = 2$), on a alors :

$$\forall x > x_0, \mathbf{P}_\beta(Y \leq 1|X = x) > \mathbf{P}_\beta(Y \leq 2|X = x).$$

Un tel résultat est évidemment gênant : il remet en cause la modélisation retenue, qui distribue les différentes modalités de Y sur un même axe ordonné. Il faut dans ce cas se tourner vers un modèle plus général (voir section 4.3).

Dans le cas où la variable Y est ordonnée, on définit l'odds d'un individu x_i relativement à l'évènement $\{Y \leq j\}$ par

$$\text{odds}(x_i) = \frac{\mathbf{P}_\beta(Y \leq j|X = x)}{1 - \mathbf{P}_\beta(Y \leq j|X = x)} = \exp(\alpha_j - x_i' \beta).$$

L'odds ratio entre deux individus x_i et $x_{i'}$ relativement à l'évènement $\{Y \leq j\}$ s'écrit donc

$$\text{OR}(x_i, x_{i'}) = \exp((x_{i'} - x_i)' \beta). \quad (4.4)$$

Cet odds ratio ne dépend pas de la modalité j , on dit que l'hypothèse des seuils aléatoires se traduit par une "hypothèse de proportionnalité des odds ratio".

4.2.4 Le test d'égalité des pentes

On se pose la question de vérifier si la modélisation en terme de seuil aléatoire est “raisonnable” vis à vis de nos données. Nous avons vu dans la partie précédente que cette hypothèse se traduit par un hypothèse d'égalité des pentes ou de proportionnalité des odds ratio. Un moyen de vérifier si cette hypothèse est raisonnable consiste à tester l'hypothèse d'égalité des pentes. Cette approche consiste tout simplement à comparer le modèle en question (avec égalité des pentes) à un modèle où on lève l'égalité des pentes. Il s'agit de considérer les hypothèses

$$H_0 : \beta_m^1 = \dots = \beta_m^{K-1}, \forall m = 1, \dots, p$$

contre

$$H_1 : \exists m \in \{1, \dots, p\}, \exists (j, \ell) \in (\{1, \dots, K-1\})^2 \text{ tels que } \beta_m^j \neq \beta_m^\ell$$

pour le modèle

$$\text{logit } p_\beta^j(x) = \alpha_j - \sum_{m=1}^p \beta_m^j x_m, j = 1, \dots, k-1 \quad (4.5)$$

Sous H_0 (égalité des pentes), nous retrouvons le modèle polytomique ordinal. Ce test peut être mené à l'aide des statistiques de Wald, du rapport de vraisemblance et du score. Par exemple, la PROC LOGISTIC de SAS effectue le test du score (“Score Test for Proportional Odds Assumption”). Si H_0 est vérifiée, la pente de la log-vraisemblance pour l'estimateur du maximum de vraisemblance contraint $\hat{\gamma} = (\hat{\alpha}_1, \dots, \hat{\alpha}_k, \hat{\beta}_1, \dots, \hat{\beta}_p)$ doit être proche de 0. On utilise que sous H_0 le statistique du score

$$\nabla \mathcal{L}_n(\hat{\gamma}) \hat{\mathcal{I}}_{H_0}^{-1} \nabla \mathcal{L}_n(\hat{\gamma})$$

converge en loi vers un $\chi_{p(K-2)}^2$ ($\hat{\mathcal{I}}_{H_0}$ est un estimateur de la matrice d'information de Fisher du modèle). Le nombre de degrés de liberté s'obtient en faisant la différence entre la dimension du modèle (4.5) ($K-1 + p(K-1)$) et celle du modèle sous H_0 ($K-1 + p$)

La probabilité critique du test d'égalité des pentes pour la statistique de rapport de vraisemblance s'obtient facilement sur R. Pour le modèle prenant en compte uniquement les variables Male et EXPLAIN de l'exemple 4.1, on obtient cette probabilité critique avec les commandes

```
> library(VGAM)
> model1 <- vglm(wallet~male+explain,data=donnees,cumulative(parallel=TRUE)) #modele sous H0
[1] "head(extra$orig.w)"
NULL
> model2 <- vglm(wallet~male+explain,data=donnees,cumulative(parallel=FALSE))
#modele sans egalite des pentes
[1] "head(extra$orig.w)"
NULL
> statRV <- -2*(logLik(model1)-logLik(model2))
> 1-pchisq(statRV,df=length(coef(model2))-length(coef(model1)))
[1] 0.3702765
```

Un exemple d'interprétation des coefficients

Tout comme pour le modèle dichotomique, les coefficients s'interprètent en terme d'odds ratio. Considérons le cas où l'on cherche à expliquer la variable Y qui correspond au type de mention obtenue au BAC (3 modalités : “AB”, “B”, “TB”) par la variable X qui correspond à la moyenne en maths au cours des deux premiers trimestres. Supposons que le coefficient β estimé pour le modèle

$$\text{logit } p_\beta^j = \alpha_j - \beta x$$

soit égal à $\log 2$ et que pour une note fixée x , le modèle fournisse les estimations :

$$\mathbf{P}_{\hat{\beta}}(Y = "AB" | X = x) = \frac{1}{4}, \quad \mathbf{P}_{\hat{\beta}}(Y = "B" | X = x) = \frac{1}{2} \quad \text{et} \quad \mathbf{P}_{\hat{\beta}}(Y = "TB" | X = x) = \frac{1}{4}.$$

Dans ce cas, on déduit de (4.4) les tableaux suivants :

	Proba	odds $\{Y = j\}$	odds $\{Y \leq j\}$
AB	1/4	1/3	1/3
B	1/2	1	3
TB	1/4	1/3	$+\infty$

TABLE 4.1 – Odds Ratio pour l'individu x .

	odds $\{Y \leq j\}$	odds $\{Y = j\}$	Proba
AB	2/3	2/3	2/5
B	6	16/19	16/35
TB	$+\infty$	1/6	1/7

TABLE 4.2 – Odds Ratio pour l'individu $x + 1$.

En effet, la formule (4.4) entraîne que les odds relativement aux événements $\{Y \leq j\}$ sont multipliés par $\exp \beta$ lorsque la variable x est augmentée d'une unité. Ainsi, la première colonne du tableau 4.2 s'obtient en multipliant la dernière colonne du tableau 4.1 par 2. On pourra par exemple dire que si pour la note x , j'ai 1 mention AB pour trois autres mentions, alors pour la note $x + 1$ j'ai 2 mentions AB pour trois autres mentions.

4.3 Le modèle polytomique nominal

Dans le cas du modèle polytomique ordonné, le caractère ordinal de la variable expliquée permettait d'introduire une variable latente, des seuils, et d'estimer des probabilités cumulées. Lorsque la variable à expliquer est purement nominale, cette démarche n'est plus possible.

4.3.1 Le modèle

Désignons par $\{1, \dots, K\}$ les modalités (non ordonnées) de Y et par $X = (1, \mathbf{X}_1, \dots, \mathbf{X}_p)$ les p variables explicatives. Tout comme pour le modèle dichotomique, nous allons chercher à modéliser les probabilités $\mathbf{P}(Y = j | X = x)$ pour $j = 1, \dots, K$. Comme $\sum_{j=1}^K \mathbf{P}(Y = j | X = x) = 1$, il suffit de modéliser les probabilités $\mathbf{P}(Y = j | X = x)$ pour $j = 1, \dots, K - 1$ par exemple. Nous prenons ici le groupe K comme groupe témoin et définissons le modèle multinomial par

$$\log \frac{p_{\beta}^j(x)}{p_{\beta}^K(x)} = \log \frac{\mathbf{P}_{\beta}(Y = j | X = x)}{\mathbf{P}_{\beta}(Y = K | X = x)} = \beta_0^j + \beta_1^j \mathbf{x}_1 + \dots + \beta_p^j \mathbf{x}_p = x' \beta^j, \quad j = 1, \dots, K - 1.$$

On a alors

$$p_{\beta}^j(x) = \frac{\exp(x' \beta^j)}{1 + \sum_{k=1}^{K-1} \exp(x' \beta^k)} \quad (4.6)$$

avec pour convention $\beta^K = 0$.

Remarque

- Si $K = 2$, on retombe sur le modèle logistique dans le cas binaire.
- L'inconvénient majeur de ce modèle est la multiplication des paramètres à estimer (un vecteur de paramètres β^j pour $K - 1$ modalités de Y) qui rend les résultats plus délicats à interpréter.
- Bien entendu, le choix du groupe de référence à une importance sur la valeur des paramètres estimés mais pas sur le modèle : quel que soit le groupe de référence choisi, les probabilités estimés $\mathbf{P}_{\beta}(Y = j | X = x)$ seront bien entendu identiques.
- Sous R, les fonctions **multinom** et **vglm** des bibliothèques **nnet** et **VGAM** permettent d'ajuster un modèle polytomique nominal. Sous SAS, on utilise la procédure **CATMOD** ou la procédure **LOGISTIC** avec l'option **ling=glogit**.

4.3.2 Estimation et interprétation des paramètres

Les paramètres $\beta^j, j = 1, \dots, k - 1$ sont estimés par maximum de vraisemblance. Pour une observation (x, y) , on désigne par y_1, \dots, y_K un codage disjonctif complet de y , *i.e.*, $y_j = 1$ si $y = j$, 0 sinon. La vraisemblance s'écrit

$$L(y, \beta) = p_\beta^1(x)^{y_1} \dots p_\beta^K(x)^{y_K}$$

où $p_j(x) = \mathbf{P}(Y = j|X = x)$ est défini par (4.6). La vraisemblance suit donc une loi multinomiale $\mathcal{M}(1, p_1(x), \dots, p_k(x))$. C'est pour cela que ce modèle est également appelé "modèle polytomique multinomial". Les estimateurs du maximum de vraisemblance s'obtiennent une fois de plus en annulant les dérivées partielles par rapport aux différents paramètres de la vraisemblance de l'échantillon. Comme pour le cas dichotomique, il n'y a pas de solutions explicites pour les estimateurs et on a recours à des méthodes numériques pour les calculer. Il n'y a pas de réelles nouveautés par rapport au cas binaire, l'algorithme est simplement plus délicat à écrire à cause de la multiplication du nombre de paramètres. Le lecteur pourra consulter l'ouvrage d'Hosmer & Lemeshow (2000) pour plus de détails.

Les odds ratio n'apparaissent généralement pas dans les sorties logiciels pour le modèle multinomial : il faut donc les calculer à la main en prenant garde au codage particulier des variables explicatives qualitatives. On rappelle que pour un individu x , l'odds d'un évènement $Y = j$ est égal au rapport $\mathbf{P}(Y = j|X = x)/\mathbf{P}(Y \neq j|X = x)$. Dans le cas du modèle multinomial, on définit l'odds d'un évènement j_1 contre un évènement j_2 par

$$\text{odds}(x, Y = j_1 \text{ vs } Y = j_2) = \frac{\mathbf{P}_\beta(Y = j_1|X = x)}{\mathbf{P}_\beta(Y = j_2|X = x)} = \exp((\beta^{j_1} - \beta^{j_2})'x).$$

Et pour deux individus x_i et $x_{i'}$, on définit alors l'odds ratio par

$$\begin{aligned} \text{OR}(x_i, x_{i'}, Y = j_1 \text{ vs } Y = j_2) &= \frac{\mathbf{P}_\beta(Y = j_1|X = x_i)/\mathbf{P}_\beta(Y = j_2|X = x_i)}{\mathbf{P}_\beta(Y = j_1|X = x_{i'})/\mathbf{P}_\beta(Y = j_2|X = x_{i'})} \\ &= \exp((\beta^{j_1} - \beta^{j_2})'(x_i - x_{i'})). \end{aligned}$$

Ainsi si les deux individus x_i et $x_{i'}$ ne diffèrent que d'une unité pour la variable ℓ , on a

$$\text{OR}(x_i, x_{i'}, Y = j_1 \text{ vs } Y = j_2) = \exp(\beta_\ell^{j_1} - \beta_\ell^{j_2}).$$

Exemple 4.2

Nous reprenons le problème de l'exemple 4.1 et considérons le modèle polytomique nominal

$$\log \frac{p_\beta^j(x)}{p_\beta^3(x)} = \beta_0^j + \beta_1^j \mathbf{1}_{x_1=1} + \beta_2^j \mathbf{1}_{x_2=1} + \beta_3^j \mathbf{1}_{x_3=2} + \beta_4^j \mathbf{1}_{x_3=3} + \beta_5^j \mathbf{1}_{x_4=1}, \quad j = 1, 2.$$

On obtient les estimations avec la fonction **multinom**

```
> library(mnet)
> model3 <- multinom(wallet~., data=donnees)
# weights: 21 (12 variable)
initial value 214.229396
iter 10 value 151.045327
final value 150.780162
converged
```

```
> model3
Call:
multinom(formula = wallet ~ ., data = donnees)

Coefficients:
(Intercept)      male1  business1  punish2  punish3  explain1
2    1.299394 -0.09558412 -0.7635377 -0.8959272 -1.788003  0.7957086
3    2.406209 -1.26720191 -1.1791095 -1.1450957 -2.141172  1.5935325

Residual Deviance: 301.5603
AIC: 325.5603
```

On remarque que la fonction **multinom** n'ajuste pas directement le modèle écrit précédemment puisqu'elle prend comme modalité de référence la première modalité de Y . Il faut transformer la valeur des estimations pour obtenir les $\hat{\beta}^j$. La fonction **vglm** prend quant à elle la troisième modalité de Y comme modalité de référence :

```
> model4 <- vglm(wallet~.,data=donnees,multinomial)
[1] "head(extra$orig.w)"
NULL
> model4
Call:
vglm(formula = wallet ~ ., family = multinomial, data = donnees)

Coefficients:
(Intercept):1 (Intercept):2      male1:1      male1:2  business1:1
-2.4062095    -1.1068174    1.2672025    1.1716184    1.1791118
 business1:2  punish2:1  punish2:2  punish3:1  punish3:2
 0.4155709    1.1450946    0.2491692    2.1411709    0.3531734
 explain1:1  explain1:2
-1.5935357   -0.7978215

Degrees of Freedom: 390 Total; 378 Residual
Residual Deviance: 301.5603
Log-likelihood: -150.7802
```

Tout comme pour les autres modèles, on peut tester la nullité d'un sous ensemble de coefficients à l'aide des statistiques de Wald, du rapport de vraisemblance et du score. Par exemple, on obtient la probabilité critique du test du rapport de vraisemblance pour le test $H_0 : \beta_1^j = \dots = \beta_5^j = 0$ contre $H_1 : \exists(j, k) \in \{1, 2\} \times \{1, \dots, 5\}, \beta_k^j \neq 0$ avec les commandes

```
> model5 <- vglm(wallet~1,data=donnees,multinomial)
[1] "head(extra$orig.w)"
NULL
> statRV <- -2*(logLik(model5)-logLik(model4))
> 1-pchisq(statRV,df=length(coef(model4))-length(coef(model5)))
[1] 2.087963e-07
```

Exemple 4.3

Nous terminons cette partie par un exemple de calcul d'odds ratio sur R. On considère le problème d'expliquer Y à trois niveaux 1,2 et 3 par X une variable qualitative à trois modalités a, b et c . Les données sont obtenues par les commandes :

```
> Y <- factor(c(rep(3,5),rep(2,5),rep(1,5)))
> set.seed(189)
> X <- factor(sample(c(rep("a",5),rep("b",5),rep("c",5))))
> donnees <- data.frame(X,Y)
```

On utilise la fonction **multinom** pour ajuster un modèle logistique multinomial :

```
> library(nnet)
> model <- multinom(Y~X,data=donnees)
```

Les coefficients du modèle sont obtenus par :

```
> beta <- coef(model)
> beta
      (Intercept)          Xb          Xc
2  1.0979772850 -10.5259443 -0.4045297
3 -0.0005551724  -0.4047684  0.6939763
```

Nous remarquons que le niveau de référence de Y est ici 1. On veut calculer l'odds ratio $OR(b, c, Y = 2 \text{ vs } Y = 3)$. On calcule d'abord $odds(b, Y = 2 \text{ vs } Y = 3)$ et $odds(c, Y = 2 \text{ vs } Y = 3)$

```
> oddb23 <- exp(beta[1,1]+beta[1,2]-beta[2,1]-beta[2,2])
> oddc23 <- exp(beta[1,1]+beta[1,3]-beta[2,1]-beta[2,3])
```

On déduit l'odds ratio cherché

```
> oddb23/oddc23
[1] 0.0001206436
```

On aurait pu le calculer directement

```
> exp(beta[1,2]-beta[2,2]-beta[1,3]+beta[2,3])
[1] 0.0001206436
```

Annexes

A.1 Rappels sur la méthode du maximum de vraisemblance

Pour plus de précisions, le lecteur pourra se reporter au chapitre 6 de l'ouvrage de Lejeune (2004) ainsi qu'au polycopié de Cadre (2010).

On considère X_1, \dots, X_n un n -échantillon i.i.d. dont la loi mère admet pour densité $f_\theta(x)$, θ étant le paramètre à estimer. On désigne par $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ un estimateur de θ .

Théorème A.1 (Inégalité de Cramer-Rao)

On suppose que $\theta \in \mathbb{R}$ et que $\hat{\theta}$ est sans biais. Sous certaines conditions de régularité, on a

$$\mathbf{V}_\theta(\hat{\theta}) \geq \frac{1}{nI(\theta)},$$

où $I(\theta)$ est l'*information de Fisher* :

$$I(\theta) = \mathbf{E} \left[\left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right)^2 \right].$$

Si un estimateur sans biais atteint la borne de Cramer-Rao, on dit qu'il est **efficace**.

Supposons maintenant que le paramètre θ à estimer est de dimension supérieure à 1. On introduit la *matrice d'information de Fisher* $\mathcal{I}(\theta)$ symétrique d'ordre k dont l'élément en position (i, j) est :

$$\mathbf{E} \left[\frac{\partial}{\partial \theta_i} \ln f(X, \theta) \frac{\partial}{\partial \theta_j} \ln f(X, \theta) \right] = -\mathbf{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(X, \theta) \right].$$

On montre alors que pour tout estimateur sans biais T et pour tout $u \in \mathbb{R}^k$

$$\mathbf{V}(u'\hat{\theta}) \geq u' \frac{[\mathcal{I}(\theta)]^{-1}}{n} u,$$

où $\mathbf{V}(u'\hat{\theta})$ représente la variance de la combinaison linéaire $u'\hat{\theta}$.

Définition A.1

On appelle *fonction de vraisemblance* de θ pour une réalisation donnée x_1, \dots, x_n de l'échantillon, la fonction de θ :

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i).$$

Définition A.2

On appelle *estimation du maximum de vraisemblance* une valeur $\hat{\theta}$, s'il en existe une, telle que :

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

Une telle solution dépend de x_1, \dots, x_n ($\hat{\theta} = h(x_1, \dots, x_n)$). La statistique $\hat{\theta} = h(X_1, \dots, X_n)$ est appelée *estimateur du maximum de vraisemblance (EMV)*.

Théorème A.2

Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance défini ci dessus. Sous certaines conditions de régularité, on a :

- $\hat{\theta}$ converge presque sûrement vers θ (il est donc asymptotiquement sans biais) ;
- $\hat{\theta}$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, [\mathcal{I}(\theta)]^{-1}).$$

La matrice de variance-covariance de $\hat{\theta}$ se “rapproche” de $\frac{1}{n}[\mathcal{I}(\theta)]^{-1}$ lorsque $n \rightarrow \infty$. On dit que l'estimateur du maximum de vraisemblance est asymptotiquement efficace.

A.2 Echantillonnage Rétrospectif

Une clinique cherche à mesurer l'effet du tabac sur le cancer du poumon. Elle prélève parmi ses patients un échantillon composé de $n_1 = 250$ personnes atteintes par le cancer et $n_0 = 250$ personnes ne présentant pas la maladie. Les résultats (simulés !) de l'étude sont présentés dans le tableau suivant :

	Fumeur	Non fumeur
Non malade	48	202
Malade	208	42

TABLE A.3 – Résultats de l'enquête.

Le statisticien responsable de l'étude réalise un modèle logistique. Les sorties sur R sont :

```
Call: glm(formula = Y ~ X, family = binomial)
```

Coefficients:

```
(Intercept) Xnon_fumeur
      1.466      -2.773
```

Degrees of Freedom: 499 Total (i.e. Null); 498 Residual

Null Deviance: 692.3

Residual Deviance: 499.9 AIC: 503.9

On note Y la variable qui prend pour valeur 1 si l'individu est malade, 0 sinon et X la variable explicative (fumeur ou non fumeur).

1. Ecrire le modèle logistique. Quelle est la probabilité $\mathbf{P}(Y = 1|X = \text{fumeur})$ estimée par ce modèle ?
2. Ce résultat vous paraît-il surprenant ?

Pour un individu (X, Y) , on définit S la variable indicatrice d'appartenance à l'échantillon de l'individu, *i.e.*,

$$S = \begin{cases} 1 & \text{si l'individu est dans l'échantillon} \\ 0 & \text{sinon.} \end{cases}$$

On définit également :

- $\tau_1 = \mathbf{P}(S = 1|Y = 1, X = x) = \mathbf{P}(S = 1|Y = 1)$: taux de sondage dans le groupe 1 (malade).
- $\tau_0 = \mathbf{P}(S = 1|Y = 0, X = x) = \mathbf{P}(S = 1|Y = 0)$: taux de sondage dans le groupe 0 (malade).
- $p_0(x) = \mathbf{P}(Y = 1|X = x, S = 1)$.
- $p(x) = \mathbf{P}(Y = 1|X = x)$.

3. Montrer à l'aide du théorème de Bayes que :

$$\text{logit } p_0(x) = \log \left(\frac{\tau_1}{\tau_0} \right) + \text{logit } p(x).$$

4. On définit

- $\pi_1 = \mathbf{P}(Y = 1)$ la probabilité à priori d'appartenance au groupe 1 ;
 - $\pi_0 = \mathbf{P}(Y = 0)$ la probabilité à priori d'appartenance au groupe 0 ;
- Exprimer τ_1 en fonction de n_1/n (la proportion d'individus du groupe 1 dans l'échantillon), π_1 et $\mathbf{P}(S = 1)$ (la probabilité qu'un individu soit dans l'échantillon).

5. Des études préalables ont montré que les probabilités π_1 et π_0 pouvaient être estimées par 0.005 et 0.995. En déduire une estimation de $\mathbf{P}(Y = 1|X = \text{fumeur})$.

Commentaires

Cette exercice nous montre que la manière de constituer l'échantillon d'apprentissage $(X_1, Y_1), \dots, (X_n, Y_n)$ est un point important qui ne doit pas être ignoré. On distingue essentiellement deux schémas d'échantillonnage :

Le schéma de mélange : tous les individus ont la même probabilité d'être sélectionnés dans l'échantillon. Les observations sont tirées aléatoirement sans distinction des groupes 0 et 1. Dans ce cas, les proportions d'individus par groupe sont sensiblement identiques dans l'échantillon et dans la population. On peut montrer que les estimateurs du maximum de vraisemblance des π_i sont les quantités n_i/n .

Le schéma rétrospectif : les proportions d'individus par groupe ne sont pas les mêmes dans l'échantillon et dans la population. Ce schéma d'échantillonnage est souvent utilisé lorsque les probabilités π_i (probabilité d'appartenance au groupe i) sont très différentes les unes des autres. Dans de tels cas, le schéma de mélange conduit à travailler avec des effectifs trop petits dans certains groupes, alors que le schéma rétrospectif permet de travailler avec des effectifs comparables. Par exemple, en diagnostic médical, on a souvent des problèmes de discrimination entre deux groupes où l'un des groupes (1 par exemple) est associé à une maladie et l'autre est caractéristique de l'absence de cette maladie. Dans de telles situations, on a bien sûr π_1 beaucoup plus petit que π_0 . L'usage consiste alors à étudier deux échantillons de taille à peu près équivalente ($n_1 \sim n_0$), le premier étant celui des malades, le second celui des individus sains (voir exercice précédent).

Pour un tel schéma, le modèle logistique appliqué directement sur les données ne nous fournit pas une estimation de $\mathbf{P}(Y = 1|X = x)$ mais de $\mathbf{P}(Y = 1|X = x, S = 1)$. Une propriété remarquable du modèle logistique est que l'effet de ce changement de proportion entre la population et l'échantillon n'intervient dans l'expression de la probabilité $\mathbf{P}(Y = 1|X = x)$ qu'au niveau de la constante β_0 :

$$\text{logit } \mathbf{P}(Y = 1|X = x) = \text{logit } \mathbf{P}(Y = 1|X = x, S = 1) - \log \left(\frac{\tau_1}{\tau_0} \right).$$

En pratique : lorsque l'échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ est constitué selon un schéma rétrospectif il faut :

1. Construire le modèle logistique sur cet échantillon. On obtient alors une estimation de

$$\text{logit } \hat{\mathbf{P}}(Y = 1|X = x, S = 1) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

2. On en déduit alors l'estimation de la probabilité recherchée

$$\text{logit } \hat{\mathbf{P}}(Y = 1|X = x) = \left(\beta_0 - \log \frac{\tau_1}{\tau_0} \right) + \beta_1 x_1 + \dots + \beta_p x_p.$$

Un tel schéma nécessite bien entendu une connaissance a priori des probabilités π_1 et π_0 .

A.3 Exercices

Exercice A.1

On se place dans le cas où les données sont présentées sous forme binomiale (répétitions au point de design x_t). On note :

- T le nombre de points de design différents $\{x_1, \dots, x_T\}$;
- n_t le nombre de répétitions au point x_t ;
- s_t le nombre de succès au point x_t : $s_t = \sum_{i=1}^{n_t} y_{it}$;
- $\bar{y}_t = \frac{s_t}{n_t}$ le nombre moyen de succès au point x_t .

x_i	nb succès	nb échecs	moyenne succès
x_1	s_1	$n_1 - s_1$	\bar{y}_1
\vdots	\vdots	\vdots	\vdots
x_t	s_t	$n_t - s_t$	\bar{y}_t
\vdots	\vdots	\vdots	\vdots
x_T	s_T	$n_T - s_T$	\bar{y}_T

TABLE A.4 – Données répétées.

Montrer que la log-vraisemblance s'écrit :

$$\begin{aligned} \mathcal{L}(\beta) &= \sum_{t=1}^T \log \binom{n_t}{s_t} + \sum_{t=1}^T n_t \{ \bar{y}_t \log(p(x_t)) + (1 - \bar{y}_t) \log(1 - p(x_t)) \} \\ &= \sum_{t=1}^T \log \binom{n_t}{s_t} + \sum_{i=1}^n n_t \left\{ \bar{y}_t \log \left(\frac{p(x_t)}{1 - p(x_t)} \right) + \log(1 - p(x_t)) \right\} \end{aligned}$$

où $p(x_t) = \mathbf{P}(Y = 1 | X = x_t)$.

Exercice A.2 (Nombre de paramètres identifiables)

On se place dans le cas de deux variables explicatives A et B de type facteur admettant respectivement deux (a_1, a_2) et trois (b_1, b_2, b_3) niveaux. On dispose de 10 individus, les données sont :

$$\begin{pmatrix} a_1 & b_3 \\ a_2 & b_2 \\ a_2 & b_1 \\ a_1 & b_2 \\ a_1 & b_1 \\ a_2 & b_3 \\ a_1 & b_3 \\ a_2 & b_1 \\ a_2 & b_2 \\ a_1 & b_3 \end{pmatrix}, \quad Y = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}.$$

1. Rappeler brièvement l'algorithme d'estimation des paramètres par la méthode du maximum de vraisemblance pour le modèle logistique.
2. Ecrire la matrice de codage disjonctif complet des variables explicatives.
3. Discuter du rang de cette matrice et en déduire le nombre de paramètres identifiables de manière unique du modèle logistique pour ce jeu de données.

4. Identifier ces paramètres sur les sorties R suivantes :

```
A<-factor(c("a1","a2","a2","a1","a1","a2","a1","a2","a2","a1"))
B<-factor(c("b3","b2","b1","b2","b1","b3","b3","b1","b2","b3"))
donnees<-data.frame(A,B)
Y<-factor(c(1,0,1,1,0,0,0,1,1,0))
model<-glm(Y~A+B,data=donnees,family=binomial)
model

Call:  glm(formula = Y ~ A + B, family = binomial, data = donnees)

Coefficients:
(Intercept)      Aa2          Bb2          Bb3
 5.690e-01    1.882e-01   -6.310e-16   -1.716e+00

Degrees of Freedom: 9 Total (i.e. Null); 6 Residual
Null Deviance:      13.86
Residual Deviance: 12.12      AIC: 20.12
```

5. Refaire les questions 2, 3 et 4 en considérant les variables d'interactions.

```
model1<-glm(Y~A+B+A:B,data=donnees,family=binomial)
model1

Call:  glm(formula = Y ~ A + B + A:B, family = binomial, data = donnees)

Coefficients:
(Intercept)      Aa2          Bb2          Bb3      Aa2:Bb2      Aa2:Bb3
 -19.57         39.13         39.13         18.87        -58.70        -58.01

Degrees of Freedom: 9 Total (i.e. Null); 4 Residual
Null Deviance:      13.86
Residual Deviance: 6.592      AIC: 18.59
```

6. On considère le modèle model2 dont les sorties R sont

```
model2<-glm(Y~A:B-1,data=donnees,family=binomial)
model2

Call:  glm(formula = Y ~ A:B - 1, family = binomial, data = donnees)

Coefficients:
  Aa1:Bb1  Aa2:Bb1  Aa1:Bb2  Aa2:Bb2  Aa1:Bb3  Aa2:Bb3
-1.957e+01  1.957e+01  1.957e+01 -2.748e-16 -6.931e-01 -1.957e+01

Degrees of Freedom: 10 Total (i.e. Null); 4 Residual
Null Deviance:      13.86
Residual Deviance: 6.592      AIC: 18.59
```

Identifier les paramètres estimés à l'aide de la matrice de codage disjonctif complet étudiée à la question précédente. Montrer que les modèles model1 et model2 sont équivalents (on pourra montrer que les probabilités $\mathbf{P}(Y = 1|X = x)$ estimées par les deux modèles sont les mêmes pour tout x).

Exercice A.3

Le traitement du cancer de la prostate change si le cancer a atteint ou non les nœuds lymphatiques entourant la prostate. Pour éviter une investigation lourde (ouverture de la cavité abdominale) un

certain nombre de variables sont considérées comme explicative de la variable Y binaire : $Y = 0$ le cancer n'a pas atteint le réseau lymphatique, $Y = 1$ le cancer a atteint le réseau lymphatique. Le but de cette étude est donc d'essayer d'expliquer Y par les variables suivantes.

- âge du patient au moment du diagnostic **age**
- le niveau d'acide phosphatase sérique **acide**, que l'on appellera par la suite niveau d'acidité
- Le résultat d'une analyse par rayon X, 0= négatif, 1=positif **rayonx**
- La taille de la tumeur, 0=petite, 1=grande, **taille**
- L'état pathologique de la tumeur déterminée par biopsie, 0=moyen, 1=grave, **grade**
- Le logarithme népérien du niveau d'acidité **log.acid**

	age	acide	rayonx	taille	grade	log.acid.
1	66	0.48	0	0	0	-0.73396918
2	68	0.56	0	0	0	-0.57981850
3	66	0.50	0	0	0	-0.69314718
4	56	0.52	0	0	0	-0.65392647
5	58	0.50	0	0	0	-0.69314718
6	60	0.49	0	0	0	-0.71334989
7	65	0.46	1	0	0	-0.77652879
8	60	0.62	1	0	0	-0.47803580
9	50	0.56	0	0	1	-0.57981850
10	49	0.55	1	0	0	-0.59783700

TABLE A.5 – Les 10 premiers individus.

On a construit 10 modèles logistiques dont les sorties R sont présentées dans le tableau A.7. On souhaite donner une prévision pour la probabilité $\mathbf{P}(Y = 1|X = x)$ pour cinq nouveaux individus. Reporter dans le tableau suivant les probabilités estimées par les différents modèles.

Individu \ Modèle	Modèle									
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
(61,0.60,1,0,1,-0.51)										
(49,0.86,0,0,1,-0.15)										
(67,0.72,1,0,1,-0.33)										
(51,0.95,1,1,1,-0.05)										

TABLE A.6 – Probabilités estimés par les différents modèles.

MODELE 1	MODELE 2
<pre>glm(formula = Y ~ age + acide, ...)</pre>	<pre>glm(formula = Y ~ log.acid. + rayonx, ...)</pre>
Coefficients: (Intercept) age acide 1.32811 -0.05639 2.14004	Coefficients: (Intercept) log.acid. rayonx1 -0.3308 2.0363 2.1020
MODELE 3	MODELE 4
<pre>glm(formula = Y ~ taille + grade, ...)</pre>	<pre>glm(formula = Y ~ taille + grade + taille:grade, ...)</pre>
Coefficients: (Intercept) taille1 grade1 -1.6008 1.4253 0.7293	Coefficients: (Intercept) taille1 grade1 taille1:grade1 -2.251 2.588 2.657 -2.860
MODELE 5	MODELE 6
<pre>glm(formula = Y ~ taille:grade - 1, ...)</pre>	<pre>glm(formula = Y ~ age + acide + age:acide, ...)</pre>
Coefficients: taille0:grade0 taille1:grade0 -2.2513 0.3365 taille0:grade1 taille1:grade1 0.4055 0.1335	Coefficients: (Intercept) age acide age:acide -2.61203 0.00675 7.93956 -0.09282
MODELE 7	MODELE 8
<pre>glm(formula = Y ~ taille:grade - 1, ...)</pre>	<pre>glm(formula = Y ~ taille:age, ...)</pre>
Coefficients: taille0:grade0 taille1:grade0 -2.2513 0.3365 taille0:grade1 taille1:grade1 0.4055 0.1335	Coefficients: (Intercept) taille0:age taille1:age 2.79408 -0.07157 -0.04351
MODELE 9	MODELE 10
<pre>glm(formula = Y ~ taille:log.acid. + rayonx:grade - 1, ...)</pre>	<pre>glm(formula = Y ~ taille:(age + grade), ...)</pre>
Coefficients: taille0:log.acid. taille1:log.acid. 3.3406 0.9356 rayonx0:grade0 rayonx1:grade0 -0.6776 1.2771 rayonx0:grade1 rayonx1:grade1 0.1129 2.3963	Coefficients: (Intercept) taille0:age taille1:age 3.17016 -0.09228 -0.04758 taille0:grade1 taille1:grade1 2.79235 -0.23813

TABLE A.7 – Les 10 modèles.

A.4 Correction

Exercice A.1

La variable $S|X = x_t$ suit une loi binomiale $\text{Bin}(n_t, p(x_t))$. Dès lors la vraisemblance s'écrit :

$$\prod_{i=1}^n P(S = s_t | X = x_t) = \prod_{t=1}^n \binom{n_t}{s_t} p(x_t)^{s_t} (1 - p(x_t))^{n_t - s_t}.$$

On obtient donc pour log-vraisemblance :

$$\mathcal{L}(\beta) = \sum_{t=1}^T \log \binom{n_t}{s_t} + \sum_{t=1}^T [n_t \log p(x_t) + (n_t - s_t) \log(1 - p(x_t))].$$

On pourra consulter le livre de Collet (2003) (page 59) pour plus de détails.

Exercice A.2

1. La procédure d'estimation par MV consiste à "mettre à jour" les coefficients via une régression pondérée : $\hat{\beta}^{k+1} = (\mathbf{X}'W^k\mathbf{X})^{-1}\mathbf{X}'W^kZ^k$. La première difficulté consiste à définir \mathbf{X} dans cet exemple. \mathbf{X} est une matrice qualitative, il est impossible de faire des opérations dessus, on a donc recours à un codage disjonctif complet.
2. La matrice de codage s'écrit

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

La matrice X utilisée pour la régression vaut :

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

3. Le nombre de paramètres estimés par le modèle est égal à la dimension de $\mathbf{X}'\mathbf{X}$, c'est-à-dire au nombre de colonnes de \mathbf{X} . Pour effectuer la régression, nous devons inverser $\mathbf{X}'\mathbf{X}$ (on considère que la matrice des poids W vaut l'identité). Les vecteurs colonnes de \mathbf{X} ne sont clairement pas linéairement indépendants (voir deuxième et troisième colonne), la matrice \mathbf{X} n'est donc pas de plein rang, le modèle est surparamétré. Il est évident que :

- la connaissance de la 2^{ème} colonne implique celle de la 3^{ème} ;
- la connaissance des 4^{ème} et 5^{ème} colonne implique celle de la 6^{ème}.

On peut donc “supprimer” des colonnes à \mathbf{X} sans “perdre” d’information. Le logiciel R supprime la première colonne correspondant chaque modalité, ce qui revient à considérer la matrice :

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Cette nouvelle matrice est de plein rang (égal à 4), le modèle est constitué de 4 paramètres identifiables de manière unique.

4. Le modèle est ainsi défini par :

x	b_1	b_2	b_3
a_1	β_0	$\beta_0 + Bb_2$	$\beta_0 + Bb_3$
a_2	$\beta_0 + Aa_2$	$\beta_0 + Aa_2 + Bb_2$	$\beta_0 + Aa_2 + Bb_3$

TABLE A.8 – Valeur de logit $p(x)$ pour les différentes classes.

5. Dans le cas d’interaction la matrice de codage “totale” ou “surparamétrée” va s’écrire

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

On peut montrer que le rang de cette matrice est $m_1 m_2 = 6$. Dans sa procédure d’estimation,

R considère la matrice :

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Le modèle est défini par :

x	b_1	b_2	b_3
a_1	β_0	$\beta_0 + Bb_2$	$\beta_0 + Bb_3$
a_2	$\beta_0 + Aa_2$	$\beta_0 + Aa_2 + Bb_2 + Aa_2 : Bb_2$	$\beta_0 + Aa_2 + Bb_3 + Aa_2 : Bb_3$

TABLE A.9 – Valeur de logit $p(x)$ pour les différentes classes.

6. En terme de matrice de codage, le modèle model2 ne considère que les variables d'interactions :

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Il est défini par :

x	b_1	b_2	b_3
a_1	$Aa_1 : Bb_1$	$Aa_1 : Bb_2$	$Aa_3 : Bb_3$
a_2	$Aa_2 : Bb_1$	$Aa_2 : Bb_2$	$Aa_2 : Bb_3$

TABLE A.10 – Valeur de logit $p(x)$ pour les différentes classes.

Pour montrer que ces modèles sont équivalents, il suffit de montrer que les logit sont identiques pour chaque classe dans les modèles model2 et model3.

Exercice A.3

Correction sous R. Les commandes sont :

```
data<-read.csv("~/COURS/GLM/TP/cancerprostate.csv",sep=";")
donnees<-data[,names(data)!="Y"]
donnees[,"rayonx"]<-factor(donnees[,"rayonx"])
donnees[,"taille"]<-factor(donnees[,"taille"])
```

```

donnees[, "grade"] <- factor(donnees[, "grade"])

Y <- factor(data[, names(data) == "Y"])

model1 <- glm(Y ~ age + acide, data = donnees, family = binomial) #2 continues
model2 <- glm(Y ~ log.acid. + rayonx, data = donnees, family = binomial) #1 cont, 1 quali
model3 <- glm(Y ~ taille + grade, data = donnees, family = binomial) #2 quali
model4 <- glm(Y ~ taille + grade + taille:grade, data = donnees, family = binomial) #2 quali + inter
model5 <- glm(Y ~ taille:grade - 1, data = donnees, family = binomial) # = modele 4
model6 <- glm(Y ~ age + acide + age:acide, data = donnees, family = binomial) # 2 conti + inter
model7 <- glm(Y ~ taille:grade - 1, data = donnees, family = binomial) #1 inter continue
model8 <- glm(Y ~ taille:age, data = donnees, family = binomial) # 1 inter qualit/conti
model9 <- glm(Y ~ taille:log.acid. + rayonx:grade - 1, data = donnees, family = binomial)
model10 <- glm(Y ~ taille:(age + grade), data = donnees, family = binomial)

X_test <- data.frame(matrix(c(61, 0.60, 1, 0, 1, -0.51, 49, 0.86, 0, 0, 1, -0.15, 67, 0.72, 1, 0, 1,
-0.33, 51, 0.95, 1, 1, 1, -0.05), ncol = 6, byrow = T))
names(X_test) <- names(donnees)

X_test[, "rayonx"] <- factor(X_test[, "rayonx"])
X_test[, "taille"] <- factor(X_test[, "taille"])
X_test[, "grade"] <- factor(X_test[, "grade"])

PRED <- matrix(0, ncol = 10, nrow = 4)
for (i in 1:10) {
  PRED[, i] <- eval(parse(text = paste("predict(model", i, ", newdata = X_test, type = 'response')", sep = "")))}

```

Le tableau cherché est la matrice PRED.

```

round(PRED, digits = 2)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 0.30 0.68 0.29 0.60 0.60 0.30 0.60 0.17 0.67 0.58
[2,] 0.60 0.35 0.29 0.60 0.60 0.65 0.60 0.33 0.40 0.81
[3,] 0.29 0.75 0.29 0.60 0.60 0.28 0.60 0.12 0.78 0.45
[4,] 0.62 0.84 0.64 0.53 0.53 0.69 0.53 0.64 0.91 0.62

```

Bibliographie

- Albert A. & Anderson D. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.
- Antoniadis A., Berruyer J. & Carmona R. (1992). *Régression non linéaire et applications*. Economica.
- Cadre B. (2010). Statistique, cours de m1. Poly. de cours, 40 pages, <http://w3.bretagne.ens-cachan.fr/math/people/benoit.cadre/>.
- Collet D. (2003). *Modelling Binary Data*. Chapman & Hall.
- Cornillon P. & Matzner-Løber E. (2007). *Régression : Théorie et Application*. Springer.
- Droesbeke J., Lejeune M. & Saporta J. (2007). *Modèles Statistiques pour Données Qualitatives*. Technip.
- Guyon X. (2005). Le modèle linéaire et ses généralisations. Poly. de cours, 83 pages, <http://samos.univ-paris1.fr/-Xavier-Guyon->.
- Hosmer D. & Lemeshow S. (2000). *Applied Logistic Regression*. Wiley.
- Lejeune M. (2004). *Statistique, la Théorie et ses Applications*. Springer.
- Perreti-Watel P. (2002). Régression sur variables catégorielles. Poly. de cours, 72 pages.
- Pommeret D. (2008). Régression sur variables catégorielles et sur variables de comptage. Poly. de cours, 100 pages.