Modèle linéaire généralisé
Guide d'utilisation

les auteurs

Le 14 octobre 2004

Table des matières

Liste des tableaux

Table des figures

1 Modèle multinomial nominal : catégories non ordonnées

1.1 Introduction : exemple 'carcasses'

Dans de nombreux pays de l'Union Européenne, la teneur en viande maigre d'une carcasse de porc est évaluée à partir d'épaisseurs tissulaires mesurées en différents sites sur la carcasse. Les méthodes d'évaluation de la teneur en viande maigre ne tiennent en revanche pas compte du génotype, exposant ainsi la prédiction à un biais entre les différents génotypes. Une manière de réduire le biais passe par la construction d'un modèle de prédiction du génotype par les épaisseurs tissulaires.

	Type génétique			
muscle	g1 : P	g2 : LWP	g3 : PAL	g4 : LWLF
44,5	0	37	10	17
50,0	0	34	14	5
52,0	1	54	23	8
54,0	3	52	23	6
56,0	3	54	34	7
58,0	3	44	20	1
60,0	4	23	12	2
62,0	7	23	8	0
66,5	19	25	6	0

TAB. 1 – *données 'carcasses' : effectifs observés*

La **variable à expliquer** est ici le **génotype de la carcasse**, variable **qualitative nominale**, dont les quatre modalités sont les types génétiques notées g1 (P : Piétrain), g2 (LWP : LargeWhite x Piétrain), g3 (PAL : Pen-ar-Lan), g4 (LWLF : LargeWhite x Landrace Français). La **variable explicative** est une **épaisseur de muscle** (en mm), elle a été découpée en 9 classes de manière à pouvoir regrouper les données dans un petit tableau (Tableau 1) et à faciliter la saisie des données au lecteur désirant reproduire l'analyse avec le logiciel de son choix[1].

L'**unité d'échantillonnage** est ici la **carcasse**, sur laquelle on observe l'une des modalités g1, g2, g3 ou g4 (le génotype prend l'une de ces modalités et une seule). On appelle réponse le type génétique observé. On dispose d'autant de réponses que de carcasses (ici, 582). Les **réponses** sont **indépendantes les unes des autres**[2]. Chaque réponse est représentée par un quadruplet du type (1,0,0,0) pour les carcasses présentant la modalité g1, (0,1,0,0) pour celles ayant la modalité g2, (0,0,1,0) pour celles ayant la modalité g3 et (0,0,0,1) pour les carcasses du type g4.

Dans le tableau de données (Tableau 1), les réponses sont présentées de manière groupée, par classe d'épaisseur de muscle (le centre de chaque classe est reporté dans la colonne muscle). Par exemple, pour la première classe d'épaisseur de muscle (centre de classe = 44,5), on a observé 0 carcasse de type g1, 37 de type g2, 10 de type g3 et 17 de type g4. Le tableau de données se présente donc sous la forme d'un tableau de contingence. On cherche à modéliser le génotype et plus particulièrement les probabilités pour une carcasse de présenter les différentes modalités g1, g2, g3 et g4 en fonction de l'épaisseur de muscle.

1.2 Où le modèle linéaire généralisé est incontournable

L'utilisateur qui ne connaîtrait pas le modèle linéaire généralisé pourrait être tenté de réaliser quatre régressions linéaires simples pour modéliser les proportions associées aux différentes modalités du génotype (Tableau 2) en fonction de l'épaisseur de muscle par quatre droites. Une telle analyse pose au moins deux problèmes, tous les deux détectables sur le graphe des résidus en fonction des probabilités prédites, réalisé pour le type génétique g1 (Figure 1). D'une part certaines probabilités prédites sont négatives ! (de l'ordre de -0.10 ici, elles pourraient être supérieures à 1 pour d'autres jeux de données). D'autre part, la structuration des résidus en fonction des probabilités prédites est très marquée, avec des résidus positifs, puis négatifs, et enfin positifs. Cette structuration vient de la forme de la liaison entre probabilité et épaisseur de muscle, qui est loin d'être linéaire (Figure 2).

1. Le découpage en classe de l'épaisseur de muscle est réalisé à des fins pédagogiques. Dans une analyse réelle, l'épaisseur de muscle n'aurait pas été découpée en classes, mais utilisée telle quelle, sous forme de variable continue.

2. le génotype d'une carcasse est indépendant du génotype d'une autre carcasse.

muscle	Type génétique			
	g1 : P	g2 : LWP	g3 : PAL	g4 : LWLF
44,5	0	57,81	15,63	26,56
50,0	0	64,15	26,42	9,43
52,0	1,16	62,79	26,74	9,30
54,0	3,57	61,90	27,38	7,14
56,0	3,06	55,10	34,69	7,14
58,0	4,41	64,71	29,41	1,47
60,0	9,76	56,10	29,27	4,88
62,0	18,42	60,53	21,05	0
66,5	38,00	50,00	12,00	0

TAB. 2 – *données 'carcasses' : proportions observées*

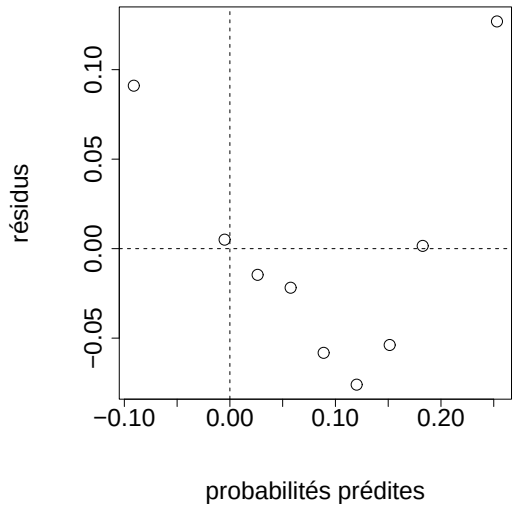

FIG. 1 – *Ce qu'il ne faut pas faire : une régression linéaire simple par type génétique. Résidus en fonction des probabilités prédites, pour le type génétique g1.*

L'utilisateur qui se bornerait à présenter (et ne regarderait que) les paramètres estimés d'un tel modèle passerait à côté de ces problèmes. Par contre l'examen du graphique des proportions observées en fonction de l'épaisseur de muscle, ou du graphique des résidus en fonction des probabilités prédites lui ferait inévitablement suspecter des problèmes. *Moralité : rien ne vaut des représentations graphiques des données brutes et des résidus !*

1.3 Modèle multinomial nominal

La probabilité associée à chacun des types génétiques est dépendante de l'épaisseur de muscle. On souhaite prédire le génotype d'une carcasse à partir de l'épaisseur de muscle, donc estimer la probabilité associée à chaque type génétique pour chaque classe d'épaisseur, et son intervalle de confiance. L'utilisation du modèle pour prédire le génotype d'une nouvelle carcasse suppose implicitement que cette carcasse est issue d'une population dont la composition est similaire à celle de l'étude (7% de type génétique g1, 59% de type génétique

g2, 26% de type génétique g3, 8% de type génétique g4).

On souhaite modéliser la façon dont varient les probabilités π_{g_1}, π_{g_2}, π_{g_3}, π_{g_4}, associées à chacune des modalités g1, g2, g3, g4 en fonction de l'épaisseur de muscle. On pourrait penser réaliser trois modélisations – une modélisation associée à chacun des types génétiques g1, g2, g3 – puis en déduire la probabilité associée au type génétique g4 (la somme des probabilités étant égale à un). Ceci pourrait se faire à l'aide de trois modèles logistiques (modèles linéaires généralisés, loi binomiale, lien logit).

Cependant, il est préférable de n'utiliser qu'un seul modèle, pour obtenir un modèle plus *simple*[3], donc plus précis, et pour pouvoir imposer la contrainte $\pi_{g_1} + \pi_{g_2} + \pi_{g_3} + \pi_{g_4} = 1$. Pour ce faire, on utilisera la loi multinomiale qui est une extension de la loi binomiale au cas de plus de deux modalités, et une généralisation du lien logit, pour définir le modèle utilisé. On utilisera la fonction de lien définie par[4] $\log(\frac{\pi_j}{\pi_J})$ où J désigne l'une des modalités de la variable à expliquer. Cette écriture correspond à une généralisation du logit défini précédemment par $\log(\frac{\pi_j}{1-\pi_j})$ au cas d'une variable à plus de 2 modalités et pour laquelle on choisit une modalité de référence, la modalité J. Selon les auteurs, ce modèle est appelé 'modèle multinomial', 'modèle multinomial pour variable à catégories non ordonnées', 'modèle logit généralisé' [6], 'baseline-category logit model' [2] [1].

Notons (n_1, n_2, n_3, n_4) le quadruplet des nombres de carcasses associées aux modalités g1, g2, g3, g4, pour une classe d'épaisseur de muscle donnée. Si les observations peuvent être supposées indépendantes les unes des autres (c'est le cas) et si les carcasses ont toutes les mêmes probabilités π_{g_1}, π_{g_2}, π_{g_3}, π_{g_4} (c'est le cas si la population est homogène – cette hypothèse peut être remise en cause si les proportions varient selon les régions par exemple–), alors le quadruplet (n_1, n_2, n_3, n_4) suit une loi multinomiale de paramètres $(n, \boldsymbol{p})$, avec $n = n_1 + n_2 + n_3 + n_4$ et $\boldsymbol{p} = (\pi_{g_1}, \pi_{g_2}, \pi_{g_3}, \pi_{g_4})$.

Si l'on note x la variable explicative 'épaisseur de muscle' et si on choisit la modalité g_4 comme modalité de référence, le modèle s'écrit :

$$\begin{cases} \log\left(\frac{\pi_{g1}}{\pi_{g4}}\right) = \theta_{g1} + \beta_{g1}\, x \\ \log\left(\frac{\pi_{g2}}{\pi_{g4}}\right) = \theta_{g2} + \beta_{g2}\, x \\ \log\left(\frac{\pi_{g3}}{\pi_{g4}}\right) = \theta_{g3} + \beta_{g3}\, x \end{cases}$$

ou encore[5]

$$\log\left(\frac{\pi_j}{\pi_{g4}}\right) = \theta_j + \beta_j\, x \qquad j = g1, g2, g3 \tag{1}$$

avec $\pi_{g1} + \pi_{g2} + \pi_{g3} + \pi_{g4} = 1$.

Ce modèle suppose un effet[6] linéaire de l'épaisseur de muscle sur le logit généralisé des probabilités associées aux différentes modalités.

3. ayant un plus petit nombre de paramètres (plus parcimonieux).

4. notation : $\log(x)$ signifie $\ln(x)$ et représente le logarithme népérien de x (réciproque de la fonction exponentielle).

5. On peut aussi poser le modèle : $\log(\frac{\pi_j}{1-\pi_j}) = \theta_j + \beta_j\, x \quad j = g1, g2, g3, g4$. Il s'agit d'un modèle équivalent défini avec une autre fonction de lien; dans le premier cas, la modalité g4 est prise comme référence, alors que dans le second cas, toutes les autres modalités servent de référence, aucune n'étant privilégiée. Ces modèles ont des paramétrisations différentes, correspondant à des contraintes d'identifiabilité différentes (nullité du paramètre associé à la dernière modalité ou nullité de la somme des paramètres associés aux différentes modalités). Les déviances, estimations, résidus, nombre de paramètres sont identiques, seule l'interprétation des paramètres est différente.

6. c'est un abus de langage : il serait plus juste de dire que le logit généralisé est lié linéairement à l'épaisseur de muscle. Le terme effet sera employé dans la suite. Il doit être compris au sens statistique, mais pas au sens d'une interprétation causale.

1.4 Mise en œuvre

1.4.1 Mise en forme des données

La mise en œuvre a été réalisée sous R (version 1.9.1). La syntaxe est analogue sous Splus (sous réserve de remplacer le signe d'affectation "=" par les deux caractères "<-" ou le caractère "_").

R

R est un système d'analyse statistique et graphique (c'est un dialecte du langage S), qui est distribué librement et dont le développement et la distribution sont assurés par plusieurs statisticiens rassemblés dans le *R Development Core Team*. Chacun peut sans difficulté télécharger un exécutable précompilé pour Windows (entre autres), distribué par le site internet du *Comprehensive R Archive Network* (CRAN)[7]. L'utilisateur intéressé trouvera au travers de l'aide en ligne ainsi qu'en faisant une recherche sur internet un grand nombre de documents d'introduction au logiciel. R est gratuit, très facile à installer, met à la disposition des utilisateurs une grande variété de méthodes, l'utilisateur débutant pourra facilement réaliser quelques analyses simples, et l'utilisateur confirmé des analyses plus sophistiquées.

R

L'estimation d'un modèle multinomial nominal peut être réalisée à l'aide de la fonction *multinom* de la bibliothèque nnet (Venables et Ripley) [9]. Lorsque les modalités sont ordonnées, on peut utiliser la fonction *polr* de la bibliothèque MASS (Venables et Ripley) [9], ou la fonction *lrm* de la bibliothèque Design (F. Harrell) [5], pour ajuster un modèle multinomial ordinal. Seule la bibliothèque VGAM (T. Yee) [10], actuellement en cours de réalisation, diffusée par son auteur[8], permet d'ajuster un grand nombre de modèles, pour catégories ordonnées ou non, au moyen d'une unique fonction, la fonction *vglm*. Cette fonction présente plusieurs avantages : modèles définis de façon simple et homogène, existence de fonctions génériques, du type coef, fitted, predict, anova (certaines sont prévues mais non encore disponibles). La mise en œuvre est donc présentée avec la fonction vglm de la bibliothèque VGAM.

mise en forme des données : création du dataframe d1.

```
> g1=c(0,0,1,3,3,3,4,7,19)
> g2=c(37,34,54,52,54,44,23,23,25)
> g3=c(10,14,23,23,34,20,12,8,6)
> g4=c(17,5,8,6,7,1,2,0,0)
> x=c(44.5,50,52,54,56,58,60,62,66.5)

> d1=data.frame(g1=g1,g2=g2,g3=g3,g4=g4)
```

1.4.2 Examen des données

Les proportions associées à chaque modalité sont représentées en fonction de l'épaisseur de muscle (Figure 2 gauche). Afin de vérifier *a priori* la linéarité de l'effet épaisseur de muscle, les logit des proportions de réponse g1, g2, g3, g4 sont représentés en fonction de l'épaisseur. Les logit n'étant pas calculables pour des proportions égales à 0 ou 1, les 'logit empiriques'[9] sont calculés.

```
> S=d1$g1+d1$g2+d1$g3+d1$g4
```

7. http://cran.r-project.org
8. http://www.stat.auckland.ac.nz/yee
9. les *logit empiriques* (logitE) sont ici définis par rapport à l'ensemble des modalités g1, g2, g3, g4 et non par rapport à la modalité g4. Ils sont obtenus en ajoutant 0.5 aux effectifs du numérateur et du dénominateur :

$$\text{logitE}(\pi_j) = \log\left(\frac{n_j+0.5}{n-n_j+0.5}\right) = \log\left(\frac{\pi_j+0.5/n}{1-\pi_j+0.5/n}\right), \quad \text{pour } j = g_1, g_2, g_3, g_4.$$

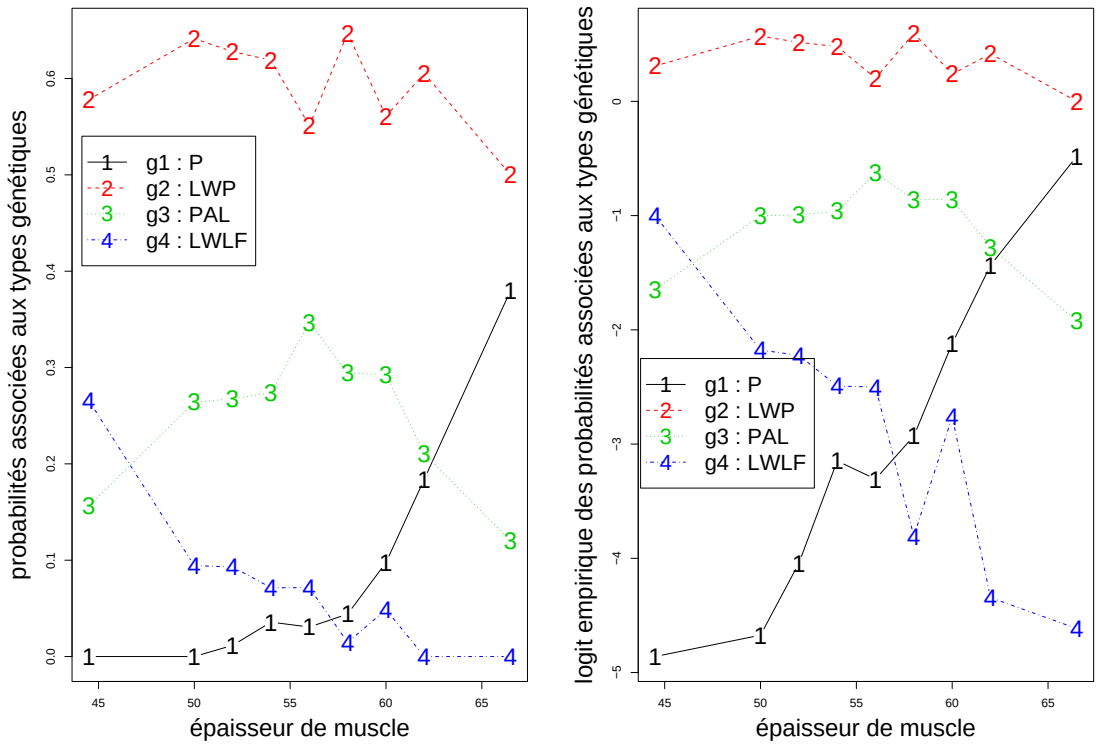

FIG. 2 – *Exemple 'carcasse' : vérification* a priori *de la linéarité de l'effet épaisseur de muscle : proportions en fonction de l'épaisseur de muscle, et logit empirique en fonction de l'épaisseur de muscle.*

```
> matplot(x,d1/S,type="b",lty=1:4,xlab="épaisseur de muscle",          # Figure 1 gauche
+ ylab="probabilités associées aux modalités")
> legend(locator(1),c("50% piétrain","25% piétrain","autres","0% piétrain"),
+ lty=1:4,col=1:4,pch=paste(1:4))

> attach(d1)
> g1.logit=log((g1+0.5)/(S-g1+0.5))
> g2.logit=log((g2+0.5)/(S-g2+0.5))
> g3.logit=log((g3+0.5)/(S-g3+0.5))
> g4.logit=log((g4+0.5)/(S-g4+0.5))
> detach()

> d1.logit=data.frame(g1=g1.logit,g2=g2.logit,g3=g3.logit,g4=g4.logit)

> matplot(x,d1.logit,type="b",lty=1:4,xlab="épaisseur de muscle",      # Figure 1 droite
+ ylab="logit empirique des probabilités associées aux modalités")
> legend(locator(1),c("50% piétrain","25% piétrain","autres","0% piétrain"),
+ lty=1:4,col=1:4,pch=paste(1:4))
```

1.4.3 Estimation

La bibliothèque VGAM doit être installée une fois pour toutes (install.packages). Par contre, la bibliothèque doit être chargée (library) dans chaque session R, avant toute utilisation de la fonction *vglm*.

```
> install.packages("VGAM", CRAN="http://www.stat.auckland.ac.nz/~yee")
> library(VGAM)
```

La fonction *vglm* peut être utilisée. Par défaut, la dernière modalité est la référence. On choisit la modalité g2 (25% piétrain) comme référence, c'est à la fois le type génétique le plus fréquent, et un type génétique qui ne parait pas lié à une épaisseur de muscle particulière (Figure 2). Le modèle estimé est le suivant :

$$\log\left(\frac{\pi_j}{\pi_{g2}}\right) = \theta_j + \beta_j\, x \qquad\qquad j = g1, g3, g4 \tag{2}$$

```
> d1$muscle=x
> vglm1=vglm(cbind(g1,g3,g4,g2)~poly(muscle,1),multinomial(),data=d1)
> summary(vglm1)


Call:
vglm(formula = cbind(g1, g3, g4, g2) ~ poly(muscle, 1), family = multinomial(),
    data = d1)

Pearson Residuals:
                        Min       1Q    Median       3Q    Max
log(mu[,1]/mu[,4]) -0.61206 -0.37818 -0.063911 0.056675 1.0117
log(mu[,2]/mu[,4]) -1.39640 -0.62111  0.010270 0.257046 1.5770
log(mu[,3]/mu[,4]) -0.90221 -0.67057 -0.099538 0.157039 1.1901

Coefficients:
                     Value Std. Error   t value
(Intercept):1     -2.936592    0.27770 -10.57470
(Intercept):2     -0.831294    0.09900  -8.39683
(Intercept):3     -2.550911    0.22514 -11.33018
poly(muscle, 1):1  5.091388    0.71740   7.09705
poly(muscle, 1):2  0.096227    0.34382   0.27987
poly(muscle, 1):3 -3.069671    0.59582  -5.15197

Number of linear predictors:  3

Names of linear predictors:
log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]), log(mu[,3]/mu[,4])

Dispersion Parameter for multinomial family:   1

Residual Deviance: 14.58408 on 21 degrees of freedom

Log-likelihood: -555.6784 on 21 degrees of freedom

Number of Iterations: 4
```

Après avoir rappelé la commande utilisée, la fonction summary affiche un résumé des résidus de Pearson (médiane, minimum, maximum, quartile inférieur, quartile supérieur), les estimations des paramètres, l'écart-type et la statistique de student associée (une valeur supérieure à deux en valeur absolue indique un paramètre significativement différent de zéro), enfin la déviance (ou déviance résiduelle) et le nombre de degrés de liberté résiduels.

$\widehat{\theta}_{g1} = -2.94$ $(t = -10.6)$ $\widehat{\theta}_{g3} = -0.83$ $(t = -8.4)$ $\widehat{\theta}_{g4} = -2.55$ $(t = -11.3)$
$\widehat{\beta}_{g1} = 5.09$ $(t = 7.1)$ $\widehat{\beta}_{g3} = 0.10$ $(t = 0.3)$ $\widehat{\beta}_{g4} = -3.07$ $(t = -5.2)$

Les paramètres θ_{g1}, θ_{g3}, θ_{g4} sont significativement négatifs : cela signifie que les probabilités associées aux types génétiques g1, g3, g4 sont globalement[10] inférieures à la probabilité associée au génotype g2 (modalité de référence).

Le paramètre β_{g1} est significativement positif : cela signifie que la probabilité associée à la modalité g1 ("50% piétrain") augmente avec l'épaisseur de muscle, plus fortement que la probabilité associée à la modalité de référence. Etant donné que la probabilité associée à la modalité de référence, 25% piétrain, ne varie pas avec l'épaisseur de muscle, on peut en conclure que la probabilité associée à la modalité g1 augmente significativement avec l'épaisseur de muscle. Le paramètre β_{g4} est significativement négatif : la probabilité associée à la modalité g4 ("0% piétrain") diminue significativement avec l'épaisseur de muscle. Le paramètre β_{g3} n'est pas significatif : la probabilité associée à la modalité "autres" est indépendante de l'épaisseur de muscle. La même analyse, conduite avec les types génétiques dans l'ordre g1, g4, g2, g3, montre que la probabilité associée à la modalité g2 ("25% piétrain") ne dépend pas non plus de l'épaisseur de muscle (attention à choisir comme modalité de référence, une modalité qui ne dépend pas de l'épaisseur de muscle, ici la modalité g3, c'est à dire les types génétiques "autres". Cela facilite l'interprétation des paramètres β_j).

1.4.4 Test d'explication

On ajuste le modèle nul, celui qui suppose que les probabilités associées aux différents types génétiques sont constantes, indépendantes de l'épaisseur de muscle. La comparaison, par test de sous modèle, entre le modèle candidat et le modèle nul, permet de tester l'effet de l'épaisseur de muscle.

```
> vglm0=vglm(cbind(g1,g3,g4,g2)~1,multinomial(),data=d1)
> summary(vglm0)


Call:
vglm(formula = cbind(g1, g3, g4, g2) ~ 1, family = multinomial(), data = d1)
....
Residual Deviance: 117.5274 on 24 degrees of freedom

Log-likelihood: -607.15 on 24 degrees of freedom

Number of Iterations: 5
```

La fonction *anova.vglm* n'étant pas encore opérationelle (elle est prévue), on en écrit une pour comparer deux modèles emboîtés. Le test de sous modèle montre que l'épaisseur de muscle apporte une explication significative.

```
> anova.vglm=function(vgam1,vgam2)
+ {stat=deviance(vgam2)-deviance(vgam1)
+  df=df.residual(vgam2)-df.residual(vgam1)
+  c(stat=stat/df,p=1-pchisq(stat,df)) }
```

10. en moyenne pour toutes les classes d'épaisseur de muscle.

```
> anova(vglm1,vglm0)

    stat        p
34.31444  0.00000
```

1.4.5 Validation du modèle

La représentation des proportions observées et des probabilités estimées montre que l'ajustement parait correct (Figure 3 gauche). Les intervalles de confiance à 95% illustrent la précision des prédictions, qui est meilleure aux bornes de l'intervalle [0,1], et proportionelle à l'effectif par classe d'épaisseur de muscle.

La fonction *fitted* calcule les probabilités estimées selon l'expression ci-dessous (démonstration en annexe). Attention, pour retrouver les probabilités estimées manuellement, le lecteur devra à nouveau estimer le modèle en remplacant 'poly(muscle)' par 'muscle' et utiliser les paramètres ainsi obtenus sans transformation de la variable 'muscle' (cf. p. 13) . En effet, la fonction *poly* calcule des polynômes orthogonaux, qui sont en fait orthonormés; en termes plus simples, le régresseur 'muscle' est centré et réduit, avant l'ajustement (moyenne nulle et variance unité).

$$\widehat{\pi}_j = \frac{\exp(\widehat{\theta}_j + \widehat{\beta}_j\, x)}{1 + \sum_{h=g_1,g_3,g_4} \exp(\widehat{\theta}_h + \widehat{\beta}_h\, x)} = \frac{\exp(\widehat{\theta}_j + \widehat{\beta}_j\, x)}{\sum_{h=g_1,g_2,g_3,g_4} \exp(\widehat{\theta}_h + \widehat{\beta}_h\, x)} \tag{3}$$

```
> matplot(x,fitted(vglm1),type="l",lty=c(1,3,4,2),col=c(1,3,4,2),      # Figure2 gauche
+ xlab="classe d'epaisseur de muscle",ylab="proba associee aux genotypes")
> attach(d1)
> matpoints(x,cbind(g1,g3,g4,g2)/S,type="p",col=c(1,3,4,2),pch=c("1","3","4","2"),
+ xlab="classe d'epaisseur de muscle",ylab="proba associee aux genotypes")
> detach()
> legend(locator(1),c("50% piétrain","25% piétrain","autres","0% piétrain"),
+ lty=1:4,col=1:4,pch=paste(1:4))
```

```
> fitted(vglm1)

           g1        g3          g4        g2
1 0.001274205 0.2147758 0.261199505 0.5227505
2 0.006589793 0.2581140 0.124479185 0.6108170
3 0.011635624 0.2680739 0.092356239 0.6279342
4 0.020281065 0.2748404 0.067642385 0.6372361
5 0.034884641 0.2780668 0.048889323 0.6381592
6 0.059073043 0.2769678 0.034787300 0.6291719
7 0.097978608 0.2702068 0.024244557 0.6075701
8 0.157816316 0.2560011 0.016409165 0.5697734
9 0.388302280 0.1908496 0.005739455 0.4151087
```

Pour chaque classe d'épaisseur de muscle, le modèle fournit les probabilités estimées, pour une carcasse, d'appartenir à l'un ou l'autre des types génétiques. Ainsi, pour une carcasse dont l'épaisseur de muscle est inférieure à 49mm, la probabilité d'être du type génétique g1 est de 0.00127, du type génétique g2 de 0.55275, du type génétique g3 de 0.21478 et du type génétique g4 de 0.26120. Autrement dit, pour une épaisseur de

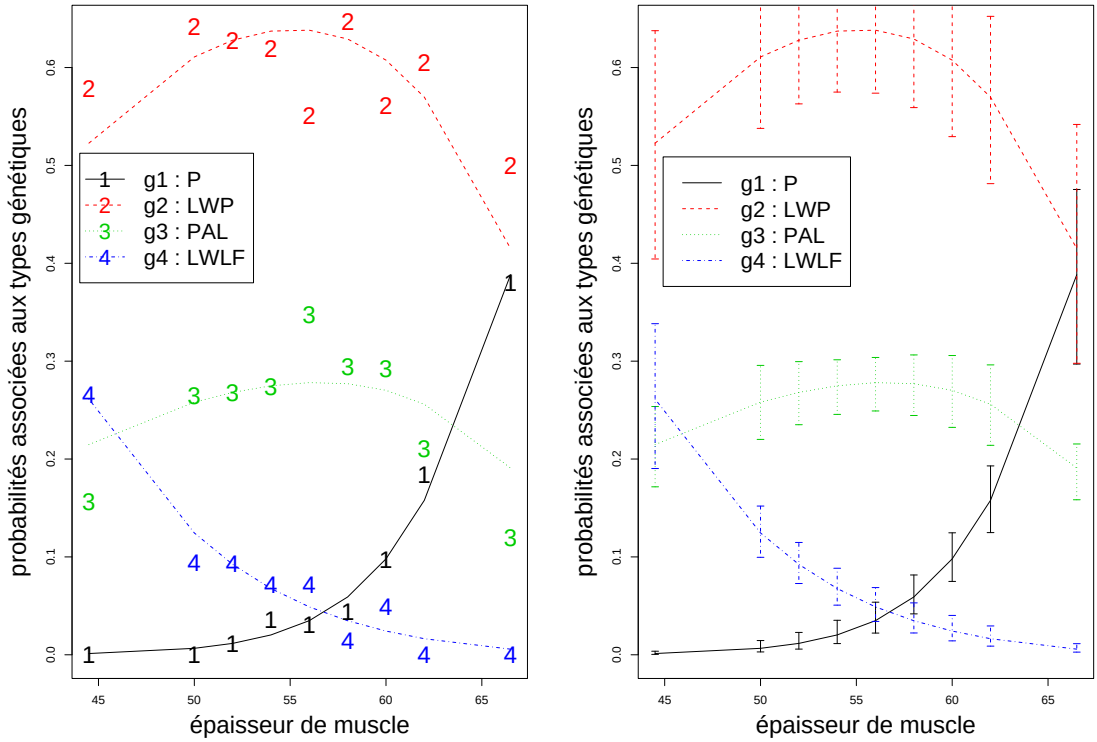

FIG. 3 – *Exemple 'carcasse' : proportions associées aux différents modalités, en fonction de l'épaisseur de muscle, probabilité estimée, et intervalle de confiance à 95%.*

muscle inférieure à 49mm, on a une probabilité d'environ 50% d'être en présence d'une carcasse de type g2 pour environ 25% que ce soit une carcasse de type g3 ou g4. Et très peu de chance que ce soit une carcasse de type g1 (mais ce n'est pas impossible[11]).

1.5 Pour aller plus loin

1.5.1 Intervalles de confiance pour les probabilités estimées

La fonction *IC.vglm*, retourne les probabilités prédites et les bornes inférieures et supérieures des intervalles de confiance. Elle utilise la fonction *predict* qui retourne les prédictions dans l'échelle du prédicteur linéaire pour les modalités 1 à $J-1$, et les écarts-type associés; on ajoute une colonne de 0 à la matrice des prédictions et à celle des écarts-type pour la modalité J ($\beta_J = 0$ et écart-type$(\beta_J) = 0$); on calcule un intervalle de confiance à 95% pour les prédictions dans l'échelle du prédicteur linéaire (les paramètres β_j étant approximativement distribués selon des lois normales, le prédicteur linéaire, qui est une combinaison linéaire des coefficients est également approximativement distribué selon une loi normale, ce qui justifie le mode de calcul). Ces intervalles de confiance sont symétriques. On applique ensuite la fonction réciproque du lien (fonction *lien.inverse*) aux bornes des intervalles de confiance calculés dans l'échelle du prédicteur linéaire, pour obtenir des intervalles de confiance dans l'échelle des probabilités. Ces intervalles de confiance sont disymétriques.

```
> lien.inverse=function(mat)
+ {exp(mat)/(apply(exp(mat),1,sum)) }
```

11. Dans l'échantillon de départ, on n'a aucune carcasse de ce type: effectif=0

```
> IC.vglm=function(vglm1)
+ pred1=predict(vglm1,se.fit=T)
+  pred1$fitted.values=cbind(pred1$fitted.values,0)
+  pred1$se.fit=cbind(pred1$se.fit,0)
+  min1=pred1$fitted.values-1.96*pred1$se.fit
+  max1=pred1$fitted.values+1.96*pred1$se.fit
+  list(pred=lien.inv(pred1$fitted.values),min=lien.inv(min1),max=lien.inv(max1))

> pred=IC.vglm(vglm1)
```

On utilise la fonction *errbar*[12] de la bibliothèque Hmisc (F. Harrell) pour représenter les intervalles de confiance (Figure 3 droite). La fonction a été légèrement modifiée pour le choix des couleurs et types de lignes (ajout des arguments col et lty passés aux fonctions segments). Une alternative à l'utilisation de la fonction errbar est l'utilisation directe de la fonction segments.

```
> library(Hmisc)

> matplot(x,fitted(vglm1),type="l",lty=c(1,3,4,2),col=c(1,3,4,2),      # Figure 2 droite
+ xlab="classe d'epaisseur de muscle",ylab="proba associee aux genotypes")
attach(d1)
> errbar(x,pred$pred[,1],pred$max[,1],pred$min[,1],add=T,lty=3,col=1)
> errbar(x,pred$pred[,2],pred$max[,2],pred$min[,2],add=T,lty=2,col=3)
> errbar(x,pred$pred[,3],pred$max[,3],pred$min[,3],add=T,lty=3,col=4)
> errbar(x,pred$pred[,4],pred$max[,4],pred$min[,4],add=T,lty=4,col=2)
> legend(locator(1),c("50% piétrain","25% piétrain","autres","0% piétrain"),
+ lty=1:4,col=1:4)
```

Enfin, une représentation des résidus de Pearson[13] en fonction du régresseur (Figure 4) (ou en fonction des probabilités prédites) confirme l'adéquation du modèle : les résidus sont répartis aléatoirement autour de zéro, avec une variance constante, sans structure particulière. Dans une prochaine version de la bibliothèque VGAM, on devrait pouvoir calculer les résidus de Pearson via la commande *residuals(vglm1, type="pearson")*.

```
> pred1=fitted(vglm1)[,paste("g",1:4,sep="")]
> res1=(d1/S-pred1)/(sqrt(pred1*(1-pred1)/S))
> matplot(x,res1,type="p",lty=c(1,3,4,2),col=c(1,3,4,2),              # Figure 3 gauche
+ xlab="épaisseur de muscle",ylab="proba associee aux genotypes",
+ cex=2,cex.lab=2,ylim=c(-2,2))
> abline(h=0)
> abline(h=c(-2,2),lty=2)
> matplot(pred1,res1,type="p",lty=c(1,3,4,2),col=c(1,3,4,2),          # Figure 3 droite
+ xlab="probabilités prédites",ylab="résidus de Pearson",
+ cex=2,cex.lab=2,ylim=c(-2,2))
> abline(h=0)
> abline(h=c(-2,2),lty=2)
```

12. A partir de la version 2.0.0 de R, la fonction errbar fait partie des fonctions de base. L'appel à la librairie Hmisc n'est plus nécessaire.

13. les résidus de Pearson sont des résidus standardisés :

$$\frac{\widetilde{\pi}_j - \widehat{\pi}_j}{\sqrt{\widehat{\pi}_j\,(1-\widehat{\pi}_j)/n}}$$

où $\widehat{\pi}_j$ sont les probabilités estimées et $\widetilde{\pi}_j$ sont les proportions observées (Tableau 2).

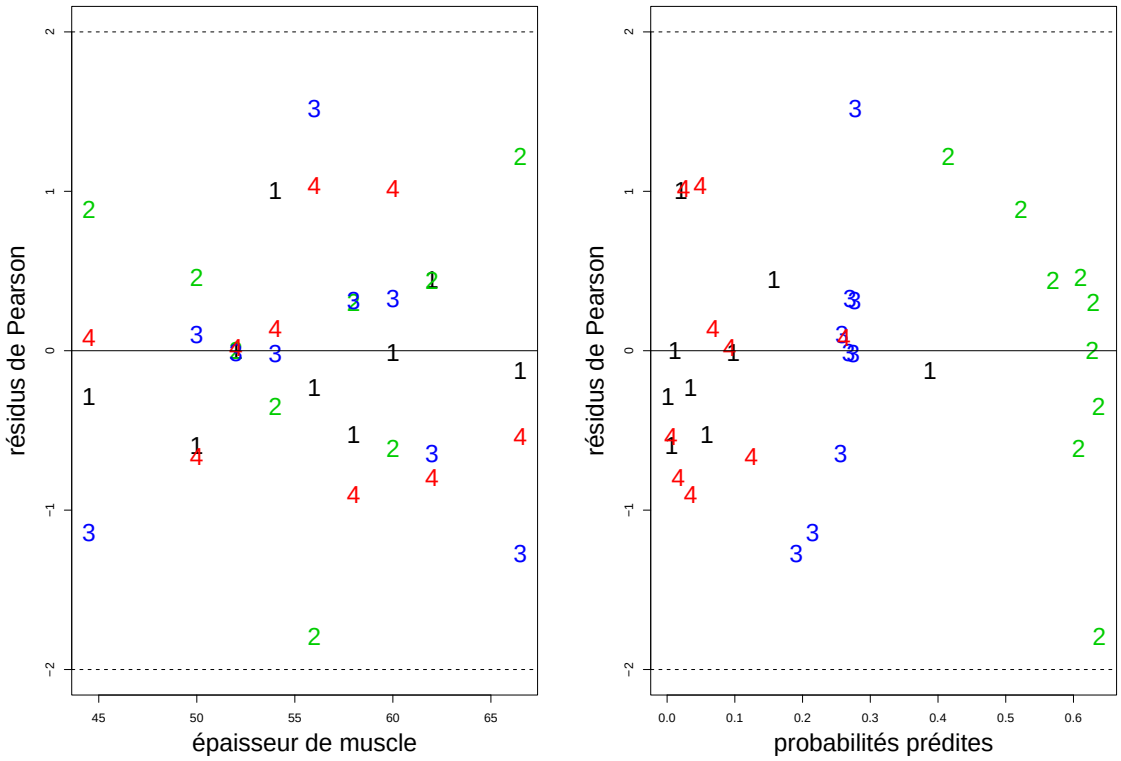

FIG. 4 – *Exemple 'carcasse' : résidus de Pearson en fonction de l'épaisseur de muscle.*

1.5.2 Interprétation : cotes et rapports de cotes

Les notions de *cote*, *rapport de cotes*, couramment utilisées en épidémiologie et dans les présentations anglo-saxonnes de la régression logistique, sont rappelées dans l'encart ci-dessous. Elles ne sont pas indispensables pour interpréter les résultats. D'ailleurs, le terme *risque* se justifie plus lorsqu'il s'agit d'une maladie, que dans le cas présent. Cependant, les résultats peuvent être présentés en ces termes.

Définitions

risque	p	probabilité associée à un évènement.
cote (odds)	$\pi/1-\pi$	rapport entre la probabilité de l'évènement et de l'évènement contraire.
rapport de cotes (odds ratio)	$\frac{\pi/(1-\pi)\,\vert X=x_1}{\pi/(1-\pi)\,\vert X=x_2}$	rapport des cotes calculées pour deux valeurs différentes des régresseurs.
risque relatif	$\frac{\pi\vert X=x1}{\pi\vert X=x2}$	rapport des risques pour deux valeurs différentes des régresseurs.

Définitions

Quelques points de repère, pour assimiler les échelles des risques, des cotes, et des log(cotes) :

risque π	cote $\pi/1-\pi$	log(cote) $\log(\pi/1-\pi)$
0.00	0.00	$-\infty$
0.01	0.01	-4.6
0.10	0.11	-2.2
0.50	1.00	0.0
0.90	9.00	2.2
0.99	99.00	4.6
1.00	∞	∞

Le rapport de cotes (parfois appelé de manière abusive "risque relatif estimé") est une approximation du risque relatif d'autant meilleure que p est petit, ce qui est le cas lorsqu'on étudie les facteurs de risque d'une maladie rare. En effet, la cote et le risque sont très voisins dès que le risque est faible.

$\widehat{\theta}_{g1} = -2.94\ (t=-10.6)$ $\quad \widehat{\theta}_{g3} = -0.83\ (t=-8.4)$ $\quad \widehat{\theta}_{g4} = -2.55\ (t=-11.3)$
$\widehat{\beta}_{g1} = 5.09\ (t=7.1)$ $\quad \widehat{\beta}_{g3} = 0.10\ (t=0.3)$ $\quad \widehat{\beta}_{g4} = -3.07\ (t=-5.2)$

Les paramètres θ_{g1}, θ_{g3}, θ_{g4}, sont les logarithmes des cotes associés aux modalités g1, g3, g4, pour une épaisseur de muscle égale à la moyenne (ceci parce que le modèle a été écrit 'poly(muscle)', si bien que le vecteur 'muscle' a été centré). Pour une épaisseur de muscle moyenne, les cotes sont $\exp(-2.94) = 0.05$, $\exp(-0.83) = 0.44$, $\exp(-2.55) = 0.08$, pour les modalités g1, g3, g4, respectivement. Les probabilités sont donc respectivement $\exp(-2.94)/(1+\exp(-2.94)) = 0.05$, $\exp(-0.83)/(1+\exp(-0.83)) = 0.30$, $\exp(-2.55)/(1+\exp(-2.55)) = 0.07$, selon l'expression $p = \text{cote}/(1+\text{cote}) = \exp(\theta_j)/(1+\exp(\theta_j))$. L'échantillon de carcasses utilisées était composé de 7% de type génétique g1, de 26% de type génétique g3, de 8% de type génétique g4, et de 59% de type génétique g2.

Avant d'interpréter les paramètres β_j, c'est à dire l'effet de l'épaisseur de muscle, il est nécessaire de réajuster le même modèle, en remplaçant 'poly(muscle)' par 'muscle' dans l'écriture du modèle. En effet, poly implique l'utilisation de polynomes orthonormés, si bien que le vecteur 'muscle' a été centré mais aussi normalisé, préalablement à l'ajustement. En conséquence, les paramètres β_j représentent les variations associées à une augmentation de une unité du régresseur centré et normé, les coefficients obtenus ne peuvent pas être utilisés directement. (par contre, les tests de nullité des coefficients sont parfaitement valides).
Remarquons que ce nouvel ajustement n'est pas nécessaire dans une pratique statistique réelle, il est réalisé dans un but pédagogique uniquement, pour illustrer les notions de cotes et rapport de cotes. Le calcul des probabilités estimées et intervalles de confiance (Figure 3 droite) est suffisant pour interpréter le modèle ajusté. Dans tous les cas, sauf pour interpréter les pentes comme la variation correspondant à une augmentation de une unité du régresseur, il est vivement recommandé d'utiliser la fonction *poly* pour tout effet linéaire, cela permet d'obtenir des paramètres peu corrélés, et des termes constants (logarithmes de cotes) correspondant à une valeur du régresseur égale à la moyenne, plutôt qu'à une valeur du régresseur égale à 0, ce qui, dans certain cas, n'a absoluement aucun sens. En contrepartie, les pentes (ici les paramètres β_j), ont l'inconvénient de ne plus être interprétables directement, puisqu'elles correspondent à une augmentation du logarithme de la cote, pour une augmentation d'une unité du régresseur ... normalisé ! par contre, les termes constants (les paramètres θ_j) peuvent être interprétés.

```
> vglm1.1=vglm(cbind(g1,g3,g4,g2)~muscle,multinomial(),data=d1)
> summary(vglm1.1)


Call:
vglm(formula = cbind(g1, g3, g4, g2) ~ muscle, family = multinomial(),
    data = d1)
```

```
Pearson Residuals:
                        Min       1Q    Median       3Q    Max
log(mu[,1]/mu[,4]) -0.61206 -0.37818 -0.063911 0.056675 1.0117
log(mu[,2]/mu[,4]) -1.39640 -0.62111  0.010270 0.257046 1.5770
log(mu[,3]/mu[,4]) -0.90221 -0.67057 -0.099538 0.157039 1.1901

Coefficients:
                    Value Std. Error  t value
(Intercept):1 -18.0520610   2.342898 -7.70501
(Intercept):2  -1.1169759   1.009918 -1.10601
(Intercept):3   6.5624235   1.622510  4.04461
muscle:1        0.2704557   0.038108  7.09705
muscle:2        0.0051116   0.018264  0.27987
muscle:3       -0.1630616   0.031650 -5.15197

Number of linear predictors:  3

Names of linear predictors:
log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]), log(mu[,3]/mu[,4])

Dispersion Parameter for multinomial family:   1

Residual Deviance: 14.58408 on 21 degrees of freedom

Log-likelihood: -555.6784 on 21 degrees of freedom

Number of Iterations: 4
```

La cote associée au type génétique g1, pour une épaisseur de muscle égale à 50mm, vaut $\exp(-18+0.27\cdot 50)=0.01=1/90$, ce qui signifie qu'environ 1% des carcasses dont l'épaisseur de muscle est égale à 50mm correspondent au type génétique g1, autrement dit que, pour une carcasse de type génétique 1, on trouvera 90 carcasses de type génétiques g2, g3, ou g4. Pour un muscle 30% plus épais, c'est à dire dont l'épaisseur vaut 65mm (65/50=1.3), la cote vaut $\exp(-18+0.27\cdot 65)=0.64=1/1.6$, donc, pour une carcasse correspondant au type génétique 1, on trouvera presque 2 carcasses de type génétique g2, g3, ou g4.

1.6 Autres documents

Les exemples de Agresti [1] ont été mis en oeuvre sous Splus et R, et présentés par Thompson [8] dans un document très complet. On peut aussi citer Fahrmeir et Tutz [4] pour une présentation intéressante et détaillée des modèles multinomiaux pour variables nominales ou ordinales, pour ceux que le formalisme mathématique ne rebute pas; ainsi que Afsa [3] pour son document de travail qui présente les modèles logit polytomiques non ordonnés (ou modèles de choix discrets) avec un minimum de formalisme, mais de façon claire et concise.

1.7 Récapitulatif des commandes

```
vglm(cbind(g1,g3,g4,g2) ~ poly(muscle),multinomial(),data=d1)
```

Références

[1] Alan Agresti. *Categorical Data Analysis.* Wiley, 1990.

[2] Alan Agresti. *An Introduction to Categorical Data Analysis.* Wiley, New York, 1996.

[3] AFSA ESSAFI C. Les modèles logit non ordonnés : théorie et applications. Technical report, Institut national de la statistique et des études économiques, 2002. Série des Documents de Travail Méthodologie Statistique.

[4] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag, New York, 1991.

[5] F.E. Harrell. *Regression Modeling Strategies//with applications to linear models, logistic regression, and survival analysis.* Springer-Verlag, New York, 2001.

[6] SAS Institute Inc. *SAS/STAT User's Guide, Version 8.* Cary, NC, 1999.

[7] McCullagh and Nelder. *Generalized Linear Models.* Chapman and Hall, London, 1983.

[8] L.A. Thompson. S-plus (and r) manual to accompany agresti's categorical data analysis (2002), second edition. Technical report, University of Houston-Clear Lake, 2003. http://math.cl.uh.edu/ thompsonla/Splusdiscrete2.pdf.

[9] Venables and Ripley. *Modern Applied Statistics with Splus,* chapter 7. Statistics and computing, 1996.

[10] T.W. Yee. Vgam family functions for categorical data/beta version 0.5-15. Technical report, Department of Statistics, University of Auckland, New Zealand, 2004. http://www.stat.auckland.ac.nz/ yee.

Première partie

Annexes

démonstration

Montrons comment exprimer p_j en fonction des paramètres θ_j, β_j et de x, autrement dit comment s'exprime la réciproque de la fonction de lien *logit généralisé.*

Le modèle multinomial est défini par :

$$\log(\frac{\pi_j}{\pi_{g_2}}) = \theta_j + \beta_j x, \quad j = g_1, g_3, g_4 \qquad \text{avec} \sum_{h=g_1,g_2,g_3,g_4} \pi_h = 1$$

Donc

$$\pi_{g_2} = 1 - \sum_{h=g_1,g_3,g_4} \pi_h = 1 - \pi_{g_2} \sum_{h=g_1,g_3,g_4} \frac{\pi_h}{\pi_{g_2}} = 1 - \pi_{g_2} \sum_{h=g1,g_3,g_4} exp(\theta_h + \beta_h x)$$

Donc

$$\pi_{g_2} \left(1 + \sum_{h=g_1,g_3,g_4} exp(\theta_h + \beta_h x) \right) = 1$$

Donc

$$\pi_{g_2} = \frac{1}{1 + \sum_{h=g_1,g_3,g_4} exp(\theta_h + \beta_h x)}$$

De la définition du modèle multinomial, on déduit

$$\pi_j = \pi_{g_2} \, exp(\theta_j + \beta_j x)$$

D'où

$$\pi_j = \frac{exp(\theta_j + \beta_j x)}{1 + \sum_{h=g_1,g_3,g_4} exp(\theta_h + \beta_h x)}, \quad j = g_1, g_2, g_3, g_4$$

Cette expression permet de calculer les probabilités π_j pour toutes les modalités, y compris la modalité de référence. Les paramètres θ_j et β_j pour la modalité de référence sont fixés à 0. Comme $exp(\theta_j + \beta_j x) = 1$ pour la modalité de référence g_2, on peut aussi écrire

$$\pi_j = \frac{exp(\theta_j + \beta_j x)}{\sum_{h=g_1,g_2,g_3,g_4} exp(\theta_h + \beta_h x)}, \quad j = g_1, g_2, g_3, g_4$$

Cette dernière expression découle directement de l'expression du lien logit généralisé. Elle montre clairement que c'est grace au lien logit généralisé que les probabilités estimés sont comprises entre 0 et 1. C'est le fait de prendre l'exponentielle du prédicteur linéaire qui assure des probabilités positives. Et c'est la normalisation de chaque terme $\exp(\theta_j + \beta_j x)$ par la somme des $exp(\theta_h + \beta_h x)$ qui assure des probabilités inférieures à 1.

Index