Loi multinomiale

Binomiale : ex. malade Y=O, non malade Y=N

non groupées

	N	O
individus	1	0
	0	1
	⋮	⋮

groupées

	N	O
groupes	10	5
	8	7
	⋮	⋮

Multinomiale : ex. notation maladie

non groupées

	0	1	2	3
individus	0	1	0	0
	1	0	0	0
	⋮	⋮	⋮	⋮

groupées

	0	1	2	3
groupes	9	1	15	17
	5	0	10	13
	⋮	⋮	⋮	⋮

Le cas multinomial est une extension du cas binomial, à plus de 2 modalités (catégories). Les modalités sont toujours exclusives.

Les données peuvent être non groupées ou groupées. Si elles sont groupées, le tableau de données comporte autant de lignes que de groupes d'individus.

On distingue le cas des catégories ordonnées, et des catégories non ordonnées. Lorsqu'on note une maladie, les catégories sont ordonnées, alors que lorsqu'on observe l'occurence d'un type de maladie (rougeole, rubéole, roséole par exemple, en supposant qu'un individu ne peut pas contracter deux de ces maladies en même temps), les catégories sont non ordonnées.

Sources :

- Multivariate Statistical Modelling Based on Generalized Linear Models, Fahrmeir and Tutz.
- An Introduction to Categorical Data Analysis, Alan Agresti.

Les deux articles qui suivent sont intéressants pour qui veut faire le lien avec l'analyse des correspondances, et les modèles loglinéaires. Il est conseillé de les lire dans l'ordre où ils sont donnés.

- Modèle linéaire généralisé et analyse des correspondances, M. Tenenhaus, Y. Le Roux, C. Guimart, P.L. Gonzalez, R.S.A., 1993, XLI (2), 59-86.
- Modèles loglinéaires pour l'étude des associations entre captures par unité d'effort et variables hydroclimatiques, J. Badia, P. Prouzet, Aquat. Living Resour., 1996, 9, 217-223.

Cas particulier de la loi binomiale

Modèle logit (= Modèle logistique)

$$p_1 = P(Y = 1)$$

$$\log\left(\frac{p_1}{p_0}\right) = \log\left(\frac{p_1}{1-p_1}\right) = \alpha + \beta x$$

$$\iff p_1 = \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$$

$$\iff p_1 = F(\alpha+\beta x)$$

$$\iff F^{-1}(p_1) = \alpha + \beta x$$

F : fct de répartition de la loi logistique

F^{-1} : fct de lien logit : g.

On revoit le cas déjà vu de la loi binomiale, pour se familiariser avec certaines notations, et avec le fait que le choix d'une fonction de lien (par exemple le lien logit) équivaut au choix d'une distribution des seuils individuels (ex. : distribution logistique associée au lien logit). Voir chap. 2.9.

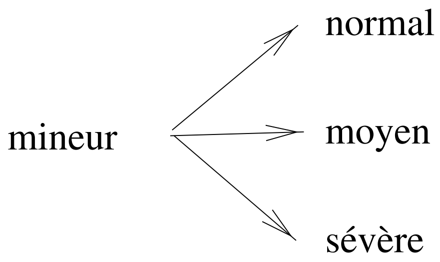
Remarque : autre notation équivalente, utilisée dans certains ouvrages :

$$P(Y = 1|x) = F(\alpha + \beta x)$$

Cette notation conditionnelle, plus lourde, est évitée ici, elle sera utilisée lorsqu'elle deviendra indispensable (transparent **??**).

Remarque formateur : le 1+ de $\frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$ vient du fait qu'on remplace p_0 par $1 - p_1$ ($p_0 + p_1 = 1$). Voir la démonstration transparent **??**; voir aussi chapitre 2, pour le cas binomial.

Exemple mineurs



nb d'années	normal	moyen	sévère
5.8	98	0	0
15.0	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46.0	12	6	10
51.5	4	2	5

L'exemple mineurs est classiquement utilisé pour présenter le cas multinomial :
on observe chez des mineurs travaillant dans des mines de charbon l'état d'avancement d'une maladie des poumons. La sévérité de la maladie est mesurée à l'aide d'une radio, et la réponse est qualitative. La radio peut se révéler normale, déceler une légère attaque du poumon, ou une attaque sévère. On observe donc une variable multinomiale à 3 modalités ordonnées. Le tableau de données contient, pour chaque classe d'ancienneté dans la mine, le nombre de mineurs présentant une radio normale, le nombre de mineurs présentant un début de maladie, et le nombre de mineurs présentant une maladie avancée. Le nombre d'années reporté dans le tableau est le centre de la classe correspondante.

On cherche à expliquer l'état d'avancement de la maladie en fonction du nombre d'années passées dans la mine.

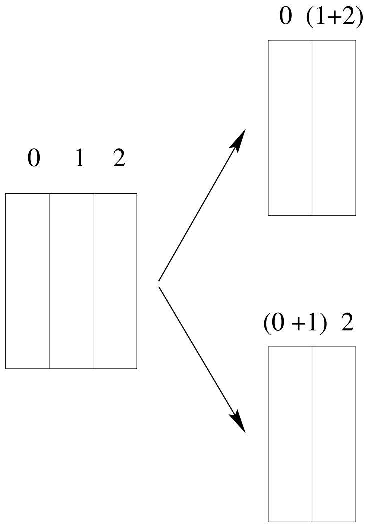
Cas multinomial

$q = 3$: nombre de modalités

$$p_0 = P(Y = 0)$$

$$p_1 = P(Y = 1)$$

$$p_2 = P(Y = 2)$$



$$p_0 + p_1 + p_2 = 1$$

Prenons le cas d'une variable à expliquer qui suit une loi multinomiale à 3 modalités 0, 1 et 2, et que l'on souhaite expliquer par une variable explicative X (par exemple le nombre d'années passées dans la mine, ou son logarithme).

On peut se ramener au cas de plusieurs modèles logit, en "découpant" le problème. On s'intéresserait dans un premier temps à $\log(\frac{p_1+p_2}{p_0})$, dans un deuxième temps, à $\log(\frac{p_2}{p_0+p_1})$.

On peut concrétiser à l'aide de l'exemple mineurs. $\frac{p_1+p_2}{p_0}$ représente le fait d'être malade (moyennement ou gravement) par rapport à non malade (il s'agit d'un rapport des probabilités, ou "odd"). Et $\frac{p_2}{p_0+p_1}$ compare la probabilité d'être très malade, à la probabilité de ne pas l'être, ou de l'être moyennement.

Modèle multinomial

ex. 3 modalités, 1 variable explicative

2 modèles :

$$\log(\frac{p_1+p_2}{p_0}) \;=\; \alpha_1+\beta_1\,x$$
$$\log(\frac{p_2}{p_0+p_1}) \;=\; \alpha_2+\beta_2\,x$$

1 modèle : $\boxed{\text{MIEUX}}$

$$\begin{cases} \log(\frac{p_1}{p_0}) = \alpha_1+\beta_1 x \\ \log(\frac{p_2}{p_0}) = \alpha_2+\beta_2 x \end{cases}$$

$$\Leftrightarrow \log(\frac{p_r}{p_0}) = \alpha_r+\beta_r\,x \qquad r=1,2$$

$$p_0+p_1+p_2=1$$

$\boxed{\text{Modèle Logit Multinomial}}$

On peut effectivement découper le problème en deux. Mais on peut faire mieux : on a tout intérêt à ne pas "découper" le modèle, à étudier toutes les données dans un même modèle, pour obtenir des écarts-type plus petits, donc un modèle plus puissant. (il faut remarquer que le nombre de découpages possibles augmente vite avec le nombre de classes).

On peut utiliser le modèle logit multinomial, ou "modèle logit multicatégories". Ce modèle ne tient pas compte de la nature ordonnée des catégories.

Remarque : avec les deux modèles binomiaux, on n'accède pas directement à p_1. L'inconvénient est résolu avec le modèle multinomial.

Remarque : c'est la modalité 0 qui est ici considérée comme modalité de référence. On peut choisir une autre modalité, selon l'exemple. Choisir une modalité de référence ou une autre revient à changer la paramétrisation du modèle. Les paramètres estimés ne prendront pas les mêmes valeurs, et ils s'interprèteront différemment. Par contre, la table d'analyse de déviance sera identique, de même que les probabilités estimées $\widehat{p_0}$, $\widehat{p_1}$, $\widehat{p_2}$.

Remarque : dans SAS, par défaut, c'est la dernière modalité qui sert de référence (les modalités sont ordonnées par ordre alphabétique). Quelle que soit la modalité choisie comme modalité de référence, on peut toujours retrouver $\log(\frac{p_a}{p_b})$ pour a et b quelconque. En effet,

$$\log(\frac{p_a}{p_b}) = \log(\frac{p_a/p_0}{p_b/p_0}) = \log(\frac{p_a}{p_0}) - \log(\frac{p_b}{p_0})$$

Catégories non ordonnées

Modèle Logit Multinomial, q modalités

$$\log(\frac{p_r}{p_0}) = \alpha_r + \beta_r x, \qquad r = 1, 2$$

$$\Leftrightarrow \; p_r = \frac{exp(\alpha_r + \beta_r x)}{1 + \sum_{s=1}^{q-1} exp(\alpha_s + \beta_s x)}$$

$$\Leftrightarrow \; p_r = F(\alpha_r + \beta_r x)$$

F : fct de répartition de la loi logistique

F^{-1} : fct de lien logit.

La réponse est dite nominale lorsque les catégories (ou modalités) sont non ordonnées. Ni l'ordre, ni *a fortiori* l'écart entre deux catégories n'ont de sens.

Le modèle logit multinomial s'applique au cas de catégories non ordonnées. Il présente une ordonnée à l'origine α_r, et une pente β_r, pour chacune des catégories, exceptée la catégorie de référence. La catégorie "de référence", notée 0, peut être la première, la dernière, ou n'importe quelle catégorie.

Remarque formateur :
Montrons que $\log(\frac{p_r}{p_0}) = \alpha_r + \beta_r x \Rightarrow p_r = \frac{exp(\alpha_r+\beta_r x)}{1+\sum_{s=1}^{q-1} exp(\alpha_s+\beta_s x)}$

$$\sum_{s=0}^{q-1} p_s = 1 \;\Rightarrow p_0 = 1 - \sum_{s=1}^{q-1} p_s$$

$$\Rightarrow p_0 = 1 - p_0 \sum_{s=1}^{q-1} \frac{p_s}{p_0}$$

$$\log(\tfrac{p_r}{p_0}) = \alpha_r + \beta_r x \;\Rightarrow p_0 = 1 - p_0 \sum_{s=1}^{q-1} exp(\alpha_s + \beta_s x)$$

$$\Rightarrow p_0 \left(1 + \sum_{s=1}^{q-1} exp(\alpha_s + \beta_s x)\right) = 1$$

$$\Rightarrow p_0 = \frac{1}{1+\sum_{s=1}^{q-1} exp(\alpha_s+\beta_s x)}$$

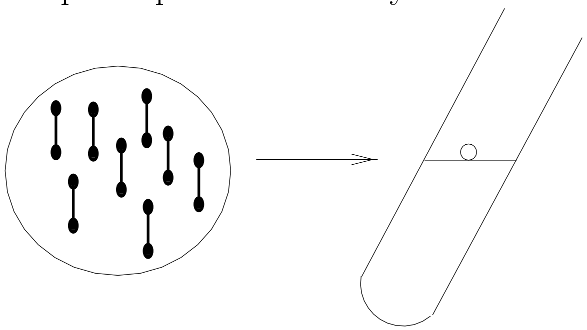
$$\log(\tfrac{p_r}{p_0}) = \alpha_r + \beta_r x \;\Rightarrow p_r = p_0 \; exp(\alpha_r + \beta_r x)$$

$$\Rightarrow \boxed{p_r = \frac{exp(\alpha_r + \beta_r x)}{1 + \sum_{s=1}^{q-1} exp(\alpha_s + \beta_s x)}}$$

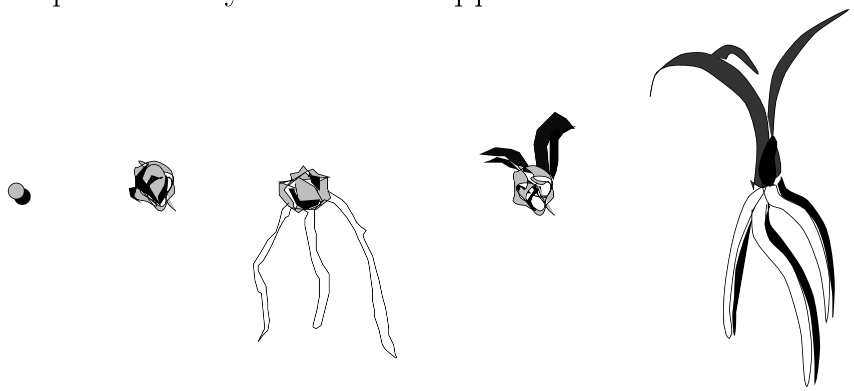
Exemple triticale

Culture in vitro de triticale, protocole

Étape 1 : pollen → embryons



Étape 2 : embryons → développement



L'objectif de cette étude est la production de plantes haploïdes par culture in vitro d'anthères.

On procède en 2 étapes. Sur un premier milieu, les grains de pollen contenus dans les anthères (partie de l'étamine) se développent en massifs de cellules appelés embryons. Les embryons sont ensuite repiqués sur un deuxième milieu, et subissent un développement plus ou moins complet. Suivant les cas, on n'obtient rien, ou un cal, ou des racines, une ébauche, ou encore une plantule complète.

C'est à la deuxième étape de la culture in vitro qu'on s'intéresse. Les 5 catégories (rien, cal, racines, ébauche, plantule) sont exclusives, et le total de leurs effectifs est égal au nombre d'embryons mis en culture. On a donc une variable réponse qualitative, que l'on considérera dans un premier temps comme non ordonnée, qui suit une loi multinomiale.

L'étude porte sur l'effet de la composition du milieu de culture, et du génotype de triticale, sur la variable réponse. Ainsi on étudie les 3 facteurs :

- Mbase : type de milieu de base, facteur à 2 modalités (synthétique, et "pomme de terre").
- Kin : addition ou non d'une hormone, la kinétine. Ce facteur a 2 modalités.
- Géno, facteur à 4 modalités correspondant à 4 génotypes de triticale.

Données brutes

exemple triticale

ligne	Mbase	Kin	Géno	plante	ébauche	racine	cal	rien	n_i
1	1	0	1	20	10	4	41	5	80
2	1	0	2	16	10	9	38	7	80
3	1	0	3	38	7	5	27	3	80
4	1	0	4	10	7	14	22	17	70
5	1	1	1	16	9	15	33	7	80
6	1	1	2	3	11	3	26	37	80
7	1	1	3	13	7	5	35	20	80
8	1	1	4	3	7	6	33	21	70
9	2	0	1	21	8	9	39	3	80
10	2	0	2	16	12	8	38	6	80
11	2	0	3	39	13	11	15	2	80
12	2	0	4	4	9	21	22	14	70
13	2	1	1	20	14	13	27	6	80
14	2	1	2	1	4	8	32	35	80
15	2	1	3	21	15	5	35	4	80
16	2	1	4	5	4	18	23	19	69

L'objectif du chercheur est de montrer l'effet de la kinétine sur la qualité de la régénération, mais les facteurs "milieu de base" et génotype sont connus pour être influents également.

Voici une première présentation des données brutes. Cette forme de présentation avec les différententes catégories (plante, ébauche, racine, cal et rien) en colonne est la plus naturelle. Elle met en évidence la nature multinomiale de la variable réponse. Les niveaux des 3 facteurs Mbase, Kin et Géno sont présentés. La première colonne contient le numéro de la ligne. On créera, à partir de cette première colonne, un facteur "ligne" que l'on utilisera dans l'analyse. Le facteur ligne a autant de modalités qu'il existe de combinaisons des facteurs étudiés (Mbase, Kin, Géno).

Ces données sont des données multinomiales groupées, avec des effectifs "suffisants". Les n_i varient entre 69 et 80, pour 5 modalités. n_i est le nombre d'embryons repiqués.

Modèle logit multinomial
astuce

$$\left\{\begin{array}{l}\text{lien logit}\\ \text{multinomiale}\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\text{lien log}\\ \text{poisson}\\ \text{conditionnellement à la marge ligne}\end{array}\right.$$

ex. : triticale :

Facteur ligne	Facteurs	5 catégories	
1 ⋮ 16		n_{ij}	$n_{i.}$

$\boxed{\text{Modèle à profil constant : Y' = ligne + catégorie}}$

Facteur ligne	Facteurs	5 catégories		
1		$\widehat{p_1}$	$\ldots$	$\widehat{p_5}$
⋮		$\widehat{p_1}$	$\ldots$	$\widehat{p_5}$
16			⋮	

$$\widehat{p_i} = \frac{n_{.j}}{n_{..}}$$

Lorsqu'on dispose d'un logiciel qui ne prévoit pas le GLM multinomial, (c'est le cas de Splus), on peut s'en tirer via une astuce. On démontre (voir "Statistical Modelling in GLIM" Aitkin p.231) que le modèle logit multinomial est statistiquement équivalent au modèle expliquant la variable comptage n_{ij}, avec lien log, loi de Poisson (modèle log-linéaire), conditionnel à la marge ligne $n_{i.}$.

On choisit l'exemple triticale pour illustrer. Voici une représentation du tableau des données initiales (elles ne sont pas réarrangées pour l'analyse, mais présentées de façon "naturelle", comme au transparent précédent).

Le modèle "à profil constant" suppose une même loi multinomiale, quels que soient les niveaux des facteurs, quelle que soit la ligne (il est parfois appelé modèle nul). Les probabilités estimées varient selon les catégories, mais elles ne varient pas selon les niveaux du facteur ligne. Pour obtenir de telles probabilités estimées, on inclut dans le modèle un effet ligne et un effet catégorie.

Il faut bien comprendre que l'utilisation de l'astuce revient à modéliser les effectifs n_{ij} (loi de Poisson). Les données fournies au modèle avec loi de Poisson sont les effectifs. On "travaille conditionnellement à la marge ligne", c'est à dire que l'on inclut dans le modèle le facteur ligne, ce qui permet de s'affranchir du fait que les effectifs varient d'une ligne à l'autre. En effet, les effectifs $n_{i.}$ (marge ligne) n'apportent pas d'information intéressante. On inclut également le facteur catégorie dans le modèle à profil constant, de manière à ce que les probabilités estimées varient d'une catégorie à l'autre.

On se ramène à l'explication d'une nouvelle variable Y', où Y' est la variable qui contient les comptages n_{ij} "déroulés". On considère les catégories comme les niveaux d'un facteur. On parlera de facteur, ou de pseudo-facteur catégorie, pour bien rappeler qu'à l'origine, catégorie n'est pas un facteur, mais une variable multinomiale. C'est l'astuce qui nous pousse à le considérer comme un facteur. Avec le modèle à profil constant, p_i est estimé par la fréquence de la catégorie, sur l'ensemble des observations : $\frac{n_{.j}}{n_{..}}$.

Modèle logit multinomial astuce

Modèles étudiés

$$\text{Y'} = \text{ligne} + \text{catégorie} \begin{array}{ll} + \text{Mbase} & .\ \text{catégorie} \\ + \text{Kin} & .\ \text{catégorie} \\ + \text{Mbase . Kin} & .\ \text{catégorie} \\ + \ldots & .\ \text{catégorie} \end{array}$$

Modèle saturé

Y' = ligne + catégorie + ligne : catégorie

5 catégories			
$\widehat{p_{11}}$	$\widehat{p_{12}}$	$\ldots$	$\widehat{p_{15}}$
$\widehat{p_{21}}$	$\widehat{p_{22}}$	$\ldots$	$\widehat{p_{25}}$
	$\vdots$		

$$\widehat{p_{ij}} = \frac{n_{ij}}{n_{i.}}$$

$$\begin{cases} \text{F}(p_r) = \text{Mbase+Kin+Mbase.Kin}, r = 2 \ldots 5 \\ \text{F : logit} \end{cases}$$

$$\Longleftrightarrow$$

$$\begin{cases} \text{F}(n_{ij}) = \text{ligne+cat+Mbase.cat+Kin.cat+Mbase.Kin.cat} \\ \text{F : log} \end{cases}$$

Les facteurs étudiés (Mbase, Kin, Géno) interviennent sous forme d'interaction avec le facteur catégorie. En effet, un traitement n'intervient pas de la même façon selon les catégories. L'effet kinétine défavorise la régénération de plantes, et favorise les modalités rien et cal (plus ou moins selon le génotype).

Avec le modèle saturé, toutes les probabilités estimées sont différentes les unes des autres, les probabilités estimées varient librement selon la ligne et la catégorie. Elles sont exactement égales aux proportions observées $\frac{n_{ij}}{n_{i.}}$.

r varie de 2 à 5, car la modalité 1 est la modalité de référence.

Astuce : démonstration

$$\left\{\begin{array}{l}\text{lien logit}\\ \text{multinomiale}\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\text{lien log}\\ \text{poisson}\\ \text{conditionnellement à la marge ligne}\end{array}\right.$$

Ceci est un commentaire inséré pour mémoire. *A priori*, il n'est pas destiné aux stagiaires, mais au formateur.

L'astuce est présentée dans "Statistical modelling in GLIM", M. AITKIN p.231.

Soient $N_1 \ldots N_r$ V.A. indépendantes $\sim \mathcal{P}(\lambda_i)$
Alors, $N = N_1 + \ldots + N_r \sim \mathcal{P}(\lambda = \sum_i \lambda_i)$

On montre (cf. la démonstration qui suit) que la distribution jointe des N_i, conditionnellement à la somme N est une distribution multinomiale :

$$P(N_1 = n_1, \ldots, N_r = n_r | N = n) = \frac{n!}{n_1! \ldots n_r!}\, p_1^{n_1} \ldots p_r^{n_r} \qquad \text{avec} \quad p_j = \frac{\lambda_j}{\lambda}.$$

Les paramètres du modèle poissonien sont : $\theta_j = \log(\frac{\lambda_j}{\lambda})$.

On veut ajuster un modèle de régression à chaque θ_j : $\theta_{ji} = \beta_j' x_i$, où j est un indice de catégorie, et i est un indice de ligne (on a un ensemble de paramètres β pour chaque niveau j du facteur catégorie).

Pour chacune des catégories j, le modèle poissonien s'écrit

$$\begin{aligned} \log\left(\frac{\lambda_{ji}}{\lambda_i}\right) &= \beta_j' x_i \quad \text{, soit} \\ \log \lambda_{ji} &= \log(\lambda_i) + \beta_j' x_i \\ &= \psi_i + \beta_j' x_i \end{aligned}$$

avec $\psi_i = \log(\lambda_i)$: paramètres de nuisance. Ces paramètres de nuisance sont le reflet des effectifs variables de la marge ligne; introduire ces paramètres de nuisance revient à travailler conditionnellement à la marge ligne.

Donc, un modèle multinomial logit peut être traité comme un modèle (appliqué aux effectifs) avec lien log, erreur poissonienne, et des paramètres de nuisance.

<u>Démo :</u>
Montrons que la loi jointe de variables aléatoires indépendantes qui suivent des lois de Poisson, conditionnellement à leur somme, est une loi multinomiale.

$$\begin{aligned} P(N_1 = n_1, \ldots, N_r = n_r) &= \textstyle\prod_{j=1}^r e^{-\lambda_j}.\frac{\lambda_j^{n_j}}{n_j!} \\ &= \textstyle\prod_j e^{-\lambda_j}.\prod \frac{\lambda_j^{n_j}}{n_j!} \\ &= e^{-\lambda}.\textstyle\prod \frac{\lambda_j^{n_j}}{n_j!} \\ P(N = n) &= e^{-\lambda}.\frac{\lambda^n}{n!} \end{aligned}$$

$$\begin{aligned} P(N_1 = n_1, \ldots, N_r = n_r | N = n) &= \frac{P(N_1=n_1,\ldots,N_r=n_r)}{P(N=n)} \\ &= \frac{\prod \frac{\lambda_j^{n_j}}{n_j!}}{\frac{\lambda^n}{n!}} \\ &= \prod \left(\frac{\lambda_j}{\lambda}\right)^{n_j} \frac{n!}{\prod n_j!} \\ &= p_1^{n_1} \ldots p_r^{n_r} \frac{n!}{n_1! \ldots n_r!} \\ &= \text{densité d'une loi multinomiale} \end{aligned}$$

Nouvelle présentation des données

exemple triticale

ligne	Mbase	Kin	Géno	Catégorie	$Y' = n_{ij}$
1	1	0	1	plante	20
2	1	0	2	plante	16
3	1	0	3	plante	38
4	1	0	4	plante	10
5	1	1	1	plante	16
6	1	1	2	plante	3
7	1	1	3	plante	13
8	1	1	4	plante	3
9	2	0	1	plante	21
...	...	...	...	...	...
16	2	1	4	plante	5
1	1	0	1	ébauche	10
...	...	...	...	...	...
16	2	1	4	ébauche	4
...	...	...	...	...	...
1	1	0	1	rien	5
...	...	...	...	...	...
16	2	1	4	rien	19

Une nouvelle présentation des données est adoptée, dans le but de mettre en œuvre l'astuce permettant d'étudier les cas multinomiaux au moyen d'une loi de Poisson, par l'intermédiaire d'un modèle équivalent, univarié.

On "déroule" la variable comptage n_{ij}, on appelle Y' cette nouvelle variable déroulée et on introduit le pseudo-facteur catégorie correspondant aux différents types de développement. La nouvelle variable Y' correspond aux données de comptage obtenues pour chaque ligne et chaque catégorie. On se ramène ainsi, par l'introduction du pseudo-facteur catégorie, à une seule colonne pour la variable à expliquer.

Le tableau présenté a 80 = 16 x 5 lignes.

Modèle à profil constant (ou modèle NUL)

exemple triticale

Y' =
ligne + catégorie

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
NULL			79	630.19	
ligne	15	4.17	64	626.01	0.997
cat	4	274.99	60	351.02	0.000

probabilités estimées

plante	ébauche	racine	cal	rien
0.20	0.12	0.12	0.39	0.17
0.20	0.12	0.12	0.39	0.17
0.20	0.12	0.12	0.39	0.17
		...		
0.20	0.12	0.12	0.39	0.17

Version Splus :
L'introduction du facteur ligne permet de travailler conditionnellement à la marge ligne. Cela permet de s'affranchir de la variabilité de cette marge, c'est à dire du nombre total d'embryons mis en culture dans chacune des conditions expérimentales. Dans cet exemple, le nombre total d'embryons par ligne est presque constant, ce qui explique que le facteur ligne n'ait pas d'effet.

Le pseudo-facteur catégorie est introduit dans le modèle. Le modèle “ligne+catégorie” suppose que la probabilité varie avec le type de catégorie, qu'elle ne varie pas avec le facteur ligne, ni *a fortiori* avec aucun des facteurs Mbase, Kin, Géno. Le facteur ligne absorbe la variabilité des effectifs marginaux (marge ligne), il ne reflète pas une variabilité des probabilités d'une ligne à l'autre. Ce modèle est un modèle à profil constant (voir le tableau des probabilités estimées), c'est un modèle NUL, dans la mesure où on n'a introduit dans le modèle aucun des facteurs Mbase, Kin, Géno, et dans la mesure où le 5-uplet des probabilités estimées est le même, quelle que soit la ligne.

La déviance résiduelle du modèle à profil constant est grande par rapport au quantile d'un $\chi^2(60)$. En effet, 351 >> qchisq(0.95,60)=79.08. Le modèle n'explique pas la variabilité des profils d'une ligne à l'autre. Il reste la totalité de cette variabilité à modéliser.

Remarque : le modèle nommé NULL de la table d'analyse de déviance (qui n'est pas du tout le modèle NUL ou à profil constant du titre), correspond à des effectifs n_{ij} constants, ne dépendant ni de la catégorie, ni de la ligne.

Modèle additif

exemple triticale

Y' =
ligne + cat +
Mbase.cat + Kin.cat + Géno.cat

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
NULL			79	630.19	
ligne	15	4.17	64	626.01	0.997
cat	4	274.99	60	351.02	0.000
Mbase.cat	4	12.78	56	338.24	0.012
Kin.cat	4	71.26	52	266.98	0.000
Géno.cat	12	176.04	40	90.94	0.000

Version Splus :
Pour étudier l'effet des facteurs Mbase, Kin et Géno sur la variable réponse, on introduit dans le modèle les termes d'interaction Mbase.cat, Kin.cat, Géno.cat, qui correspondent en fait aux effets principaux Mbase, Kin et Géno.

La déviance résiduelle a diminué, mais elle reste bien supérieure au quantile qchisq(0.95,40)=55.75. Une bonne part de la variabilité est "capturée" par le modèle additif, mais il reste une part de variabilité à expliquer.

Remarque : l'ordre des termes dans le modèle est important. Il faut introduire le facteur ligne d'abord, le facteur categorie ensuite, les autres termes après.

Modèle interactif d'ordre 2

exemple triticale

Y' =
ligne + cat +
Mbase.cat + Kin.cat + Géno.cat +
Mbase.Kin.cat+Mbase.Géno.cat+Kin.Géno.cat

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
NULL			79	630.19	
ligne	15	4.17	64	626.02	0.997
cat	4	274.99	60	351.02	0.000
Mbase.cat	4	12.78	56	338.24	0.012
Kin.cat	4	71.26	52	266.98	0.000
Géno.cat	12	176.04	40	90.93	0.000
Mbase.Kin.cat	4	1.76	36	89.18	0.780
Mbase.Géno.cat	12	17.25	24	71.92	0.140
Kin.Géno.cat	12	53.75	12	18.17	0.000

Version Splus :
Ce modèle "capture" la totalité de la variabilité. En effet, on accepte l'hypothèse d'adéquation du modèle, la déviance (résiduelle) n'est pas significative (18.17 est à comparer au quantile qchisq(0.95,12)=21.02).

On pourrait être tenté de retenir ce modèle. Toutefois, les tests de Chi2 associés à chacun des termes indiquent que les interactions Mbase.Kin et Mbase.Géno (i.e. les termes Mbase.Kin.cat et Mbase.Géno.cat) ne sont pas significatives. On envisage donc un sous-modèle ne contenant que l'interaction Kin.Géno.

Modèle avec un seul terme d'interaction (modèle retenu)

exemple triticale

Y' =
ligne + cat +
Mbase.cat + Kin.cat + Géno.cat +
Kin.Géno.cat

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
NULL			79	630.19	
ligne	15	4.17	64	626.02	0.997
cat	4	274.99	60	351.02	0.000
Mbase.cat	4	12.78	56	338.24	0.012
Kin.cat	4	71.26	52	266.98	0.000
Géno.cat	12	176.04	40	90.93	0.000
Kin.Géno.cat	12	53.29	28	37.64	0.000

Version Splus :
La déviance résiduelle demeure non significative. 37.64 est à comparer à qchisq(0.95,28)=41.33. On accepte l'hypothèse d'adéquation du modèle.

On peut aussi réaliser un test de sous-modèle, pour valider l'hypothèse d'adéquation de ce modèle (cela peut faire l'objet d'un exercice). $(37.64 - 18.17 = 19.47 \sim \chi^2(0.95, 28 - 12) = 26.30)$

Le modèle obtenu est plus parcimonieux que le précédent, il est adopté, on l'appellera modèle retenu.

Modèle à profil constant (ou modèle NUL)

exemple triticale

Y' =
ligne + catégorie

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	630.1957	0	.	.
ligne	626.0194	15	4.1763	0.9971
cat	351.0251	4	274.9942	0.0001

probabilité estimées

plante	ébauche	racine	cal	rien
0.20	0.12	0.12	0.39	0.17
0.20	0.12	0.12	0.39	0.17
0.20	0.12	0.12	0.39	0.17
		...		
0.20	0.12	0.12	0.39	0.17

Version SAS :
Les données triticales ont été analysées à l'aide de la procédure genmod.

L'introduction du facteur ligne permet de travailler conditionnellement à la marge ligne. Cela permet de s'affranchir de la variabilité de cette marge, c'est à dire du nombre total d'embryons mis en culture dans chacune des conditions expérimentales. Dans cet exemple, le nombre total d'embryons par ligne est presque constant, ce qui explique que le facteur ligne n'ait pas d'effet.

Le pseudo-facteur catégorie est introduit dans le modèle. Le modèle "ligne+catégorie" suppose que la probabilité varie avec le type de catégorie, qu'elle ne varie pas avec le facteur ligne, ni a fortiori avec aucun des facteurs Mbase, Kin, Géno. Le facteur ligne absorbe la variabilité des effectifs marginaux (marge ligne), il ne reflète pas une variabilité des probabilités d'une ligne à l'autre. Ce modèle est un modèle à profil constant (voir le tableau des probabilités estimées), c'est un modèle NUL, dans la mesure où on n'a introduit dans le modèle aucun des facteurs Mbase, Kin, Géno, et dans la mesure où le 5-uplet des probabilité estimées est le même, quelle que soit la ligne.

La déviance résiduelle du modèle à profil constant est grande par rapport au quantile d'un $\chi^2(60)$ (60=5x16 données, -1 pour l'effectif général, -15 pour les 16 lignes, -4 pour les 5 catégories). En effet, 351 >> qchisq(0.95,60)=79.08. Le modèle n'explique pas la variabilité des profils d'une ligne à l'autre. Il reste la totalité de cette variabilité à modéliser.

Modèle additif

exemple triticale

Y' =
ligne + cat +
Mbase(cat) + Kin(cat) + Géno(cat)

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	630.1957	0	.	.
ligne	626.0194	15	4.1763	0.9971
cat	351.0251	4	274.9942	0.0001
Mbase(cat)	338.2405	4	12.7846	0.0124
Kin(cat)	266.9794	4	71.2611	0.0001
Géno(cat)	90.9367	12	176.0427	0.0001

Version SAS :
Pour étudier l'effet des facteurs Mbase, Kin et Géno sur la variable réponse, on introduit dans le modèle les termes d'interaction Mbase.cat, Kin.cat, Géno.cat, qui correspondent en fait aux effets principaux Mbase, Kin et Géno (la notation Mbase.cat est équivalente à Mbase(cat)).

La déviance résiduelle a diminué, mais elle reste bien supérieure au quantile qchisq(0.95,40)=55.75 (le degré de liberté résiduel est donné par SAS, mais la sortie correspondante n'a pas été reportée sur ce transparent, il est égal à 40). Une bonne part de la variabilité est "capturée" par le modèle additif, mais il reste une part de variabilité à expliquer.

Remarque : l'ordre des termes dans le modèle est important. Il faut introduire le facteur ligne d'abord, le facteur categorie ensuite, les autres termes après.

Remarque : voir le transparent 13 "Version Splus", pour retrouver les degrés de liberté.

Modèle interactif d'ordre 2

exemple triticale

Y' =
ligne + cat +
Mbase(cat) + Kin(cat) + Géno(cat) +
Mbase*Kin(cat) + Mbase*Géno(cat) +
Kin*Géno(cat)

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	630.1957	0	.	.
ligne	626.0194	15	4.1763	0.9971
cat	351.0251	4	274.9942	0.0001
Mbase(cat)	338.2405	4	12.7846	0.0124
Kin(cat)	266.9794	4	71.2611	0.0001
Géno(cat)	90.9367	12	176.0427	0.0001
Mbase*Kin(cat)	89.1769	4	1.7598	0.7798
Mbase*Géno(cat)	71.9250	12	17.2519	0.1404
Kin*Géno(cat)	18.1753	12	53.7497	0.0001

Version SAS :
Ce modèle “capture” la totalité de la variabilité. En effet, on accepte l'hypothèse d'adéquation du modèle, la déviance (résiduelle) n'est pas significative (18.17 est à comparer au quantile qchisq(0.95,12)=21.02).

On pourrait être tenté de retenir ce modèle. Toutefois, les tests de Chi2 associés à chacun des termes indiquent que les interactions Mbase.Kin et Mbase.Géno (i.e. les termes Mbase*Kin(cat) et Mbase*Géno(cat)) ne sont pas significatives. On envisage donc un sous-modèle ne contenant que l'interaction Kin.Géno.

Modèle avec un seul terme d'interaction (modèle retenu)

exemple triticale

Y' =
ligne + cat +
Mbase(cat) + Kin(cat) + Géno(cat)
Kin*Géno(cat)

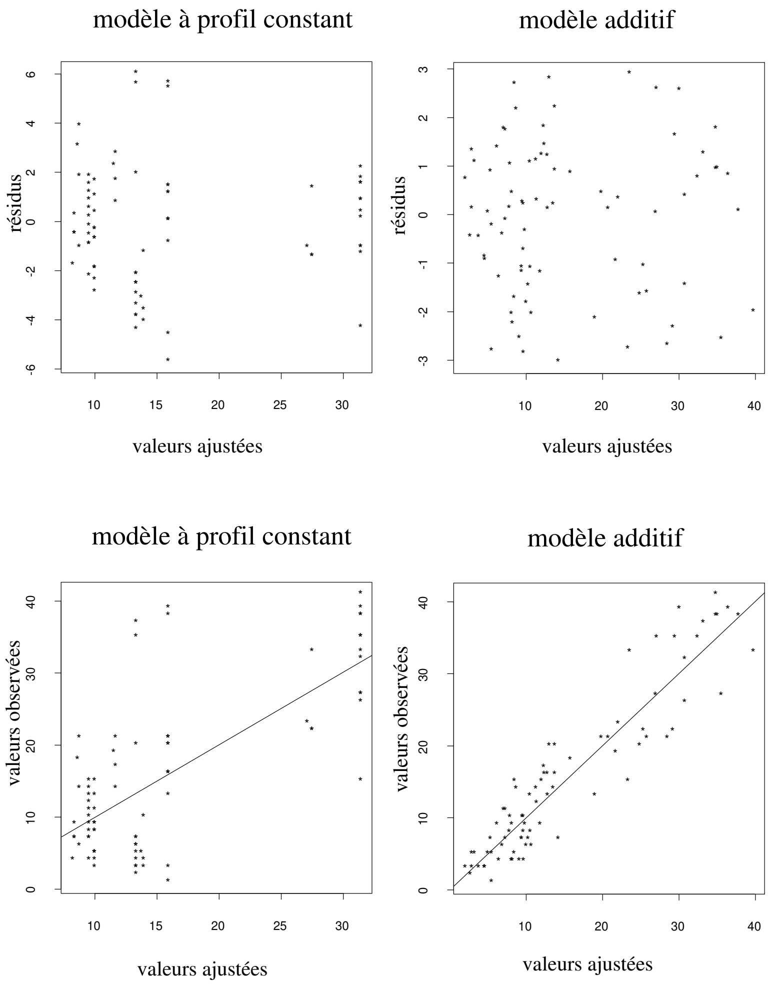
Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	630.1957	0	.	.
ligne	626.0194	15	4.1763	0.9971
cat	351.0251	4	274.9942	0.0001
Mbase(cat)	338.2405	4	12.7846	0.0124
Kin(cat)	266.9794	4	71.2611	0.0001
Géno(cat)	90.9367	12	176.0427	0.0001
Kin*Géno(cat)	37.6440	12	53.2927	0.0001

Version SAS :
La déviance résiduelle demeure non significative. 37.64 est à comparer à qchisq(0.95,28)=41.33. On accepte l'hypothèse d'adéquation du modèle.

On peut aussi réaliser un test de sous-modèle, pour valider l'hypothèse d'adéquation de ce modèle (cela peut faire l'objet d'un exercice). $(37.64 - 18.17 = 19.47 \sim \chi^2(0.95, 28 - 12) = 26.30)$

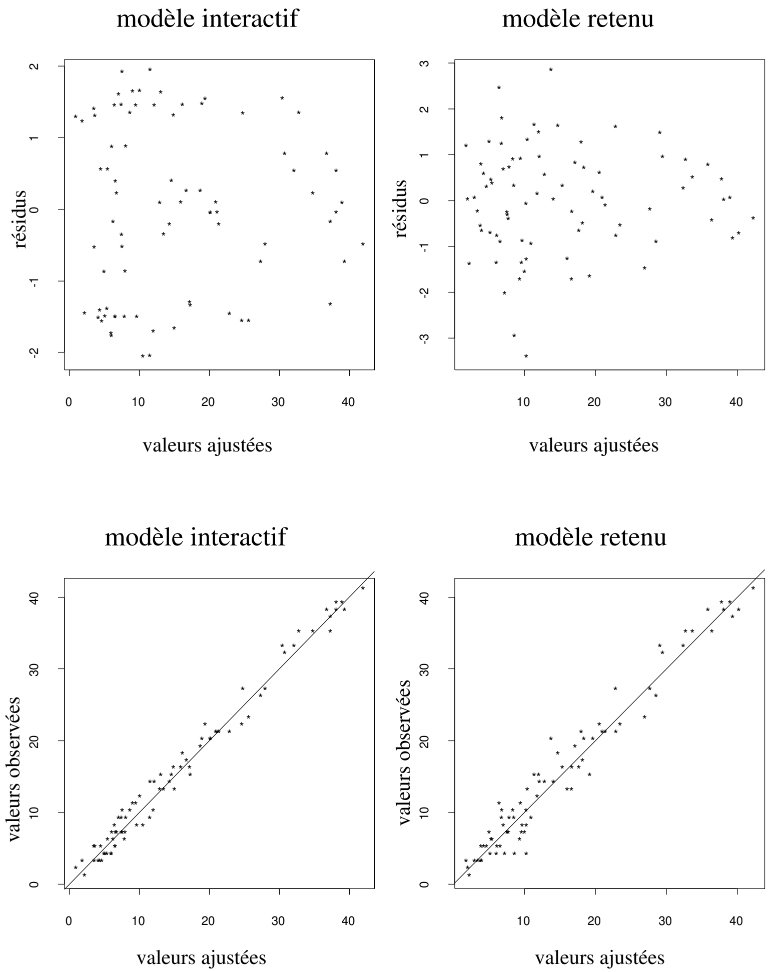
Le modèle obtenu est plus parcimonieux que le précédent, il est adopté, on l'appellera modèle retenu.

Modèle à profil constant, modèle additif



On vérifie graphiquement que le modèle à profil constant présente une structure des résidus et des résidus grands. Il n'est pas adapté. Le modèle additif apporte une amélioration très nette. On ne détecte pas de structure systématique des résidus, si ce n'est, peut-être, une liaison entre la variance des résidus et la valeur estimée.

Modèle interactif d'ordre 2, modèle retenu

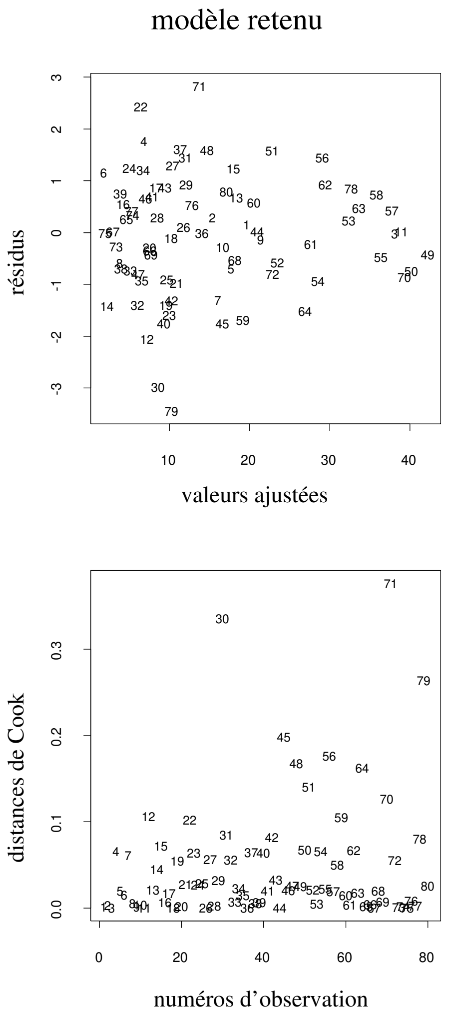


Les graphiques obtenus sont tout à fait satisfaisants pour le modèle interactif d'ordre 2, et pour le modèle retenu.

C'est bien le modèle retenu qui sera utilisé, ce n'est pas le modèle interactif. Ces graphiques sont présentés pour montrer que le fait de passer du modèle retenu au modèle interactif diminue encore les résidus. Mais il faut garder à l'esprit le fait que trop de termes non significatifs sont conservés dans le modèle interactif, ce n'est pas le modèle qu'il faut utiliser pour interpréter.

Dans une pratique réelle, on aurait utilisé une stratégie descendante de choix de modèle : on aurait d'abord ajusté un modèle interactif d'ordre 2 ou 3, et on aurait réalisé des tests de sous-modèle, pour choisir le modèle final.

Modèle retenu - validation



Il ne semble pas exister de structure dans les résidus. Il n'existe pas de distance de Cook anormalement élevée. Ce modèle est retenu.

Remarque : dans le cas multinomial, il ne faut pas estimer un modèle après avoir supprimé une observation du tableau déroulé (c'est à dire une observation correspondant à une modalité). Si une distance de Cook est très élevée et nécessite d'enlever l'observation correspondante, il faut enlever la ligne complète (du tableau non déroulé) avant de réajuster le modèle.

Analyse génotype par génotype

génotype 1 :

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
ligne	3	0.00	16	143.58	1.000
cat	4	127.81	12	15.76	0.000
cat.Mbase	4	1.65	8	14.11	0.799
cat.Kin	4	10.63	4	3.47	0.031

génotype 2 :

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
ligne	3	0.00	16	200.762	1.000
cat	4	18.01	12	82.75	0.000
cat.Mbase	4	1.73	8	81.01	0.734
cat.Kin	4	74.96	4	6.05	0.000

génotype 3 :

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
ligne	3	0.00	16	183.01	1.000
cat	4	19.46	12	63.55	0.000
cat.Mbase	4	18.80	8	44.75	0.001
cat.Kin	4	40.46	4	4.28	0.000

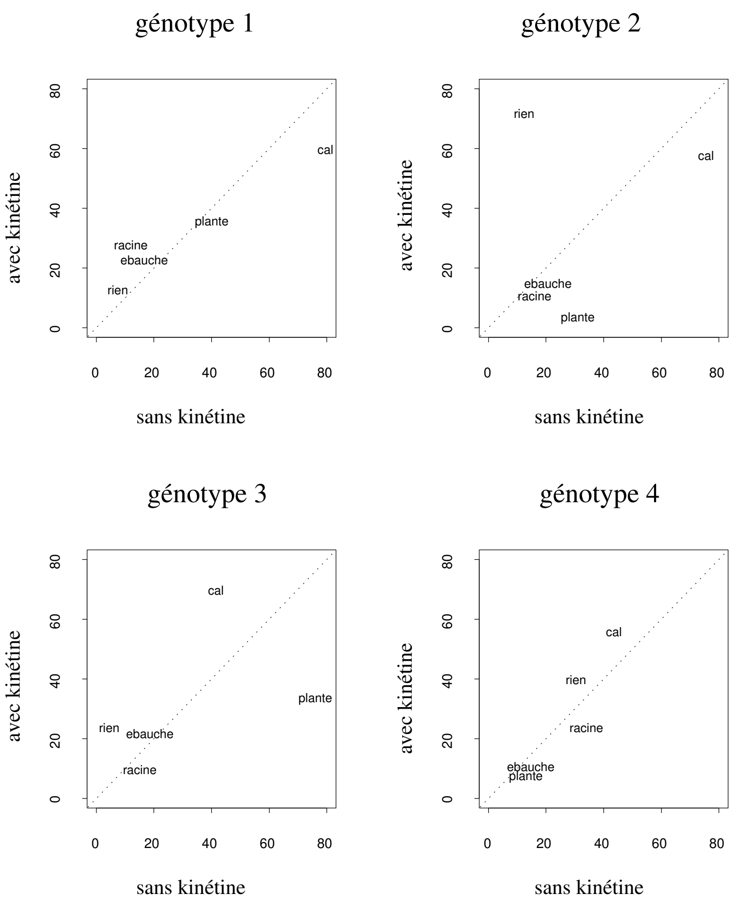
génotype 4 :

	Df	Dev	Res.Df	Res.Dev	Pr.Chi
ligne	3	0.01	16	98.67	1.000
cat	4	77.31	12	21.24	0.000
cat.Mbase	4	8.35	8	12.99	0.079
cat.Kin	4	7.27	4	5.73	0.120

Compte tenu de l'importance de l'interaction entre les facteurs kinétine et génotype, et pour faciliter l'interprétation de cette interaction, on se propose de refaire une analyse pour chaque niveau du facteur génotype.

Seuls les génotypes 2 et 3 présentent une réaction hautement significative à la kinétine. On pourrait déterminer la nature de cette réaction, en examinant les paramètres du modèle. On peut aussi faire des représentations graphiques.

Réponse des 4 génotypes à la kinétine



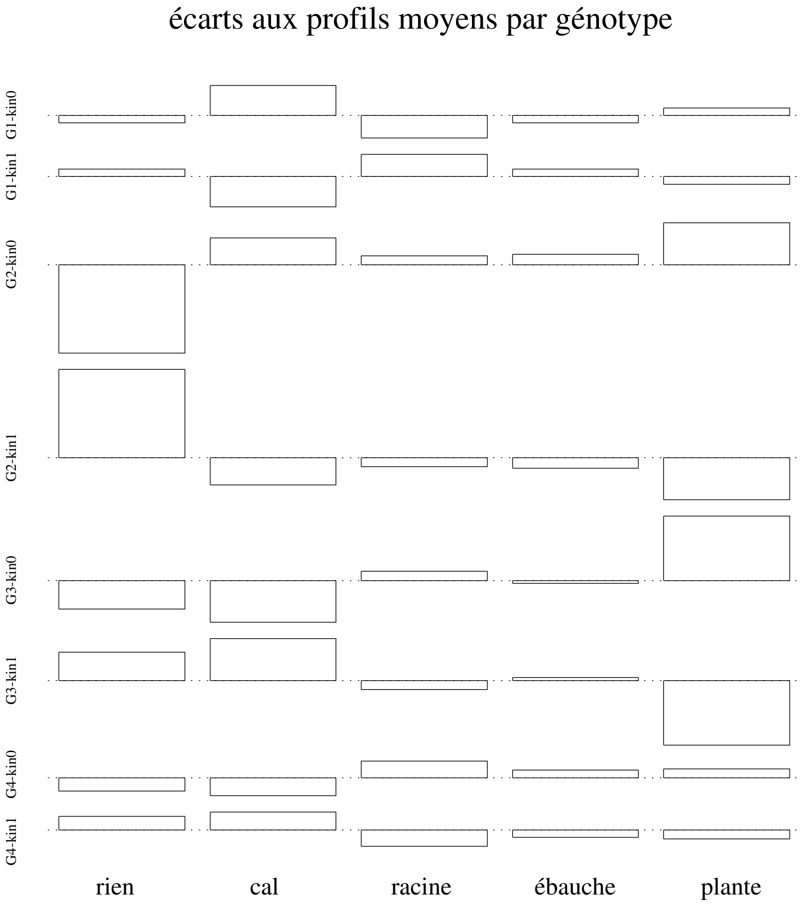
On peut utiliser différents types de graphiques pour visualiser la réponse à la kinétine, pour chaque génotype.

On représente ici, pour chaque génotype, les effectifs observés de chaque catégorie, obtenus sans kinétine en abscisse, et avec kinétine en ordonnée (sommés sur Mbase). L'absence d'effet kinétine se traduit par une position proche de la diagonale pour l'ensemble des catégories. En cas d'effet kinétine, on visualise les catégories favorisées ou au contraire inhibées par la kinétine, ce sont celles qui s'écartent de la diagonale.

Pour les génotypes 2 et 3 qui présentent un effet kinétine significatif, la catégorie plante diminue au profit de la catégorie rien et/ou de la catégorie cal.

L'apport de kinétine entraîne donc une dégradation de la réponse morphogénétique pour ces deux génotypes, alors qu'il n'a pas d'effet significatif pour les deux autres génotypes : s'agit-il d'une insensibilité des génotypes 1 et 4, ou d'un effet de seuil ? il aurait fallu tester un plus grand nombre de niveaux de kinétine, pour conclure.

Réponse des 4 génotypes à la kinétine



Cette représentation est obtenue à l'aide de la commande Splus mulbar.

On représente, pour chaque génotype, l'écart aux effectifs moyens obtenus avec les traitements avec et sans kinétine. Les diagrammes d'un même génotype sont symétriques, et font bien apparaître les catégories favorisées ou non.

Comme l'échelle est commune pour les 4 génotypes, la hauteur des boîtes traduit l'importance des effets, qui est plus grande pour les génotypes 2 et 3.

Conclusion sur l'exemple triticale

AVANTAGE :

Par rapport à l'analyse catégorie par catégorie :

interprétation globale en terme de profil.

LIMITES :

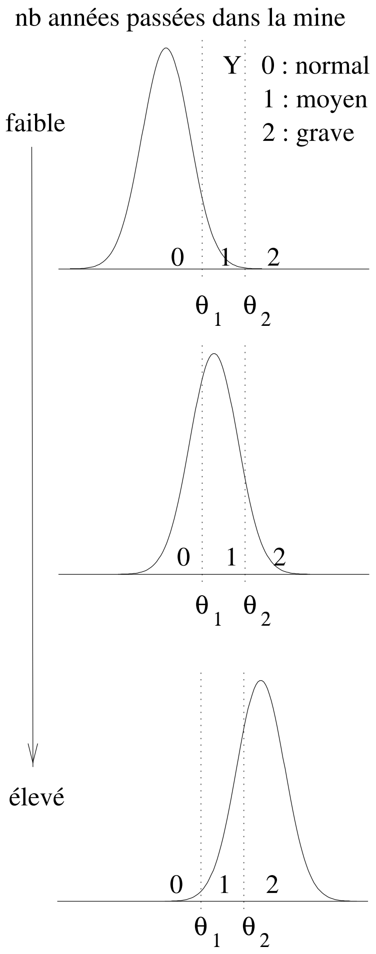
1. Nécessité de regrouper les catégories à effectif faible,
2. Ordre des catégories non pris en compte.

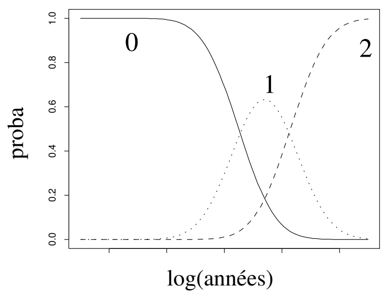
$\longrightarrow$ Modèles pour catégories ordonnées ?

Le modèle utilisé jusqu'à présent (logistique multinomial) ne tient pas compte du fait que les catégories peuvent être ordonnées. Existe-t-il des modèles spécifiques pour cas multinomiaux à catégories ordonnées ?

catégories ordonnées

Approche seuil : seuils fixés





facultatif

Si les catégories sont ordonnées, il est préférable de le prendre en compte dans le modèle, on obtient un modèle plus parcimonieux.

L'approche seuil est une des approches possibles : elle est basée sur l'hypothèse selon laquelle la variable observée Y est une version catégorisée d'une variable latente continue U. U est une variable aléatoire sous-jacente, non observée. Dans l'exemple mineurs, U pourrait être une quantité de cellules détruites, la distribution de U dépendant du nombre d'années passées dans la mine, elle se décale progressivement au fur et à mesure que les années passent, mais conserve une variance constante. Les seuils sont constants aussi. C'est la distribution de U qui est représentée sur le transparent. Y est une version catégorisée de U, déterminée par les seuils $\theta_1, \ldots, \theta_{q-1}$ (q=3 dans l'exemple mineurs).

Si $U < \theta_1$, Y prend la modalité 0,
Si $\theta_1 < U < \theta_2$, Y prend la modalité 1,
Si $\theta_2 < U$, Y prend la modalité 2.

Dans les modèles simples, les variables explicatives agissent sur Y, au moyen d'un décalage de la variable latente, les seuils restant fixés.

Il existe des modèles généraux, où les variables explicatives agissent à la fois sur les seuils, et sur la variable latente.

Catégories ordonnées

Approche seuil :

Modèle cumulatif simple (seuils fixés)

$$p_r = P(Y \leq r) = P(U \leq \theta_r) = F(\alpha_r + \beta x)$$

F: logistique $\Rightarrow$ modèle cumulatif logistique simple.

$$P(Y \leq r) = \frac{exp(\alpha_r + \beta x)}{1 + exp(\alpha_r + \beta x)}$$

$$\log\left(\frac{P(Y \leq r)}{P(Y > r)}\right) = \underbrace{\alpha_r}_{=\theta_r} + \beta\ x$$

contrainte $\theta_1 < \ldots < \theta_{q-1}$

proportional odds model

$\rightarrow$ ordre strict des populations

facultatif

- β est constant, il ne dépend pas de la catégorie r.
- le modèle logit multinomial, avec les mêmes notations, s'écrit :
$$\log\left(\frac{P(Y=r)}{P(Y=0)}|x\right) = \alpha_r + \beta_r\ x$$
ou
$$P(Y = r|x) = F(\alpha_r + \beta_r\ x)$$
(on introduit ici la notation "conditionnellement à x" parce qu'elle est nécessaire dans l'expression mathématique ci-dessous. Cette notation n'était pas utilisée jusqu'à présent, mais le conditionnement existait, il était implicite).
- le modèle cumulatif simple, par rapport au modèle logit multinomial, ajoute une contrainte d'ordre sur les seuils.
- le terme "simple" signifie que les seuils sont constants : aucune des variables explicatives n'a d'influence sur les seuils.
- une caractéristique de ce modèle : les odd-ratios sont indépendants de la catégorie r.
$$\frac{P(Y \leq r|x_1)/P(Y > r|x_1)}{P(Y \leq r|x_2)/P(Y > r|x_2)} = exp(\beta(x_1 - x_2))$$
On appelle "odd" la quantité $P(Y \leq r)/P(Y > r)$.
Il est possible d'ordonner strictement les populations correspondant aux différents niveaux des facteurs étudiés.
- Si F=logistique, le modèle est appelé modèle cumulatif logistique, ou "proportional odds model".
- Si F est la distribution "extreme-minimal-value" associée au lien cloglog, le modèle est appelé "grouped cox model", ou "proportional hazard model" (Fahrmeir and Tutz p.78).

Catégories ordonnées

Approche seuil :

- modèle cumulatif : $P(Y \leq r) = F(\theta_r + \beta\, x)$
  séquentiel : $P(Y = r | Y \geq r) = F(\theta_r + \beta\, x)$
- modèle simple : seuils fixés
  généralisé : seuils variables

choix de F $\rightarrow$ logit
$\rightarrow$ extreme minimal value distribution
$\rightarrow$ extreme maximal value distribution
$\rightarrow$ ...

$\rightarrow\ F^{-1}(\eta) = \log(\frac{p}{1-p}) \qquad F(p) = \frac{\exp \eta}{1+\exp \eta}$
$\rightarrow\ F^{-1}(\eta) = \log(-\log(1-p)) \qquad F(p) = 1 - \exp(-\exp \eta)$
$\rightarrow\ F^{-1}(\eta) = -\log(-\log p) \qquad F(p) = \exp(-\exp(-\eta))$
$\rightarrow$ ...

facultatif

La variété des modèles possibles, basés sur l'approche seuil, est très grande.

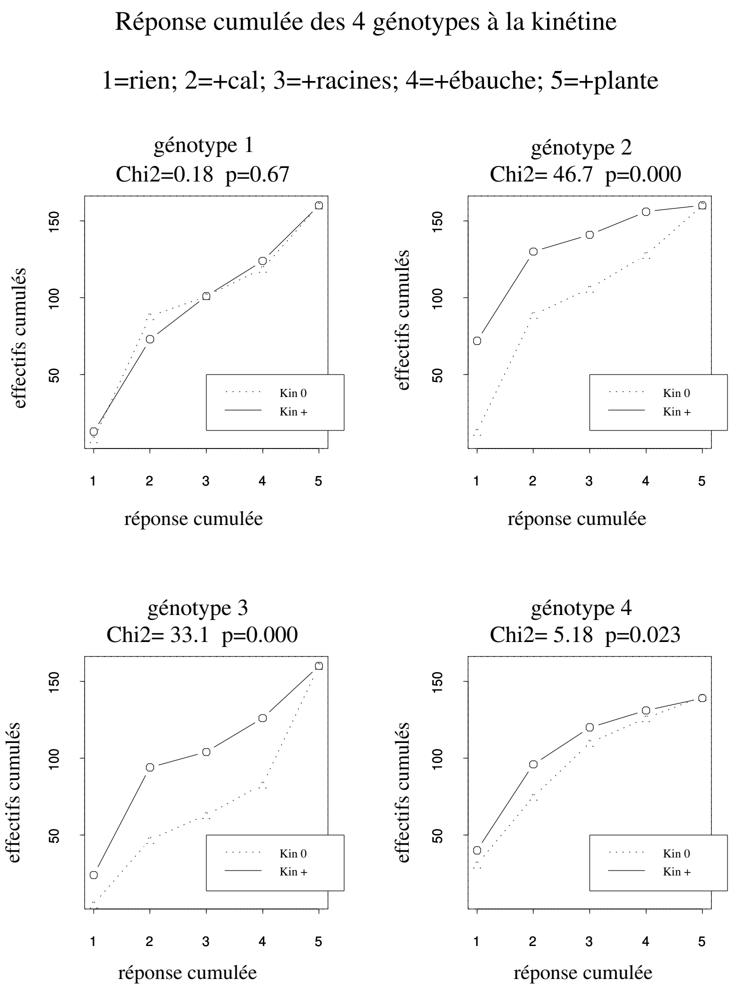
- dans les modèles séquentiels, on suppose que, pour atteindre une catégorie, il est nécessaire d'être passé par les catégories précédentes (ex. des différents stades dans l'évolution d'une maladie).
- dans les modèles généralisés, les seuils dépendent de certaines des variables explicatives. Par exemple, on peut imaginer que le fait d'être fumeur abaisse le seuil de sensibilité à certaines maladies, alors que l'âge agirait sur une variable latente, laissant les seuils constants.
- Les deux types de modèles cumulatifs et séquentiels peuvent être combinés avec les deux types de modèles simple ou généralisé. On peut aussi choisir la distribution F de son choix.

Remarque formateur : des modèles à deux étapes peuvent être envisagés, appelés composés cumulatif-cumulatif. Les catégories sont divisées en grands groupes, eux-mêmes subdivisés en sous-groupes. L'appartenance à un grand groupe est modélisée par un modèle cumulatif. L'appartenance à un sous-groupe est modélisée par un modèle cumulatif, conditionné par l'appartenance à un grand groupe.

Remarque formateur : d'autres approches alternatives existent, comme le modèle logit à catégories adjacentes :

$$\log\left(\frac{P(Y = r)}{P(Y = r-1)}\right) = \beta_r\, x$$

Ex. triticale, modèle cumulatif



<u>facultatif</u>

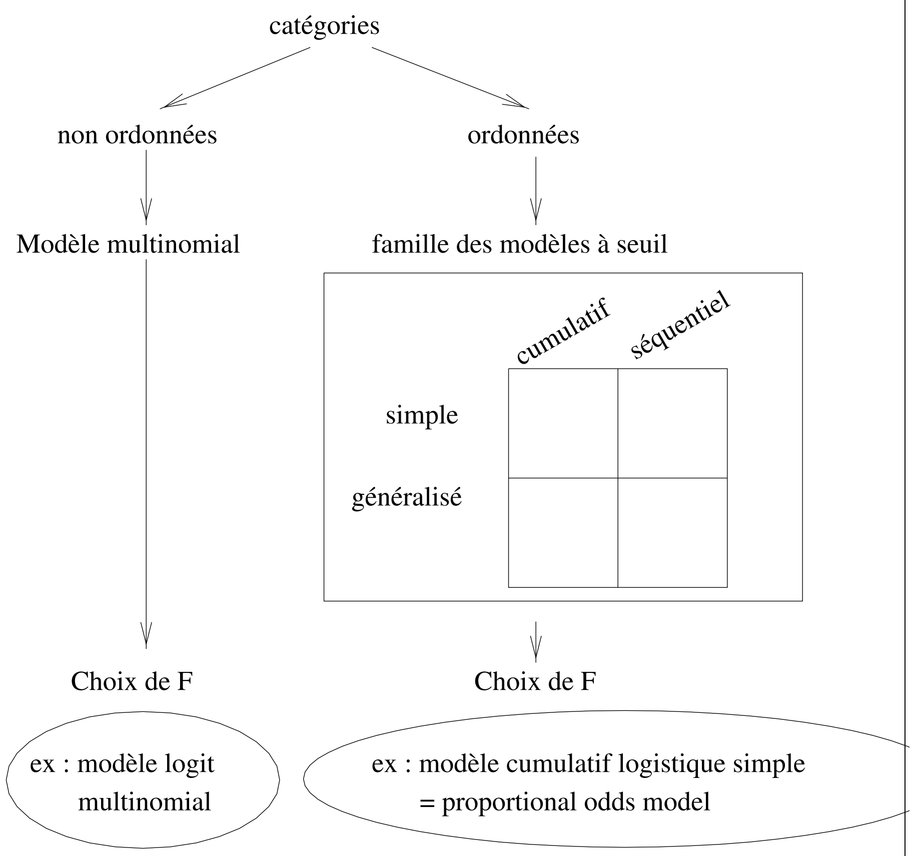
On a ici considéré les 5 catégories comme ordonnées, c'est à dire que l'on suppose que les 5 niveaux de réponse (de cal à plante entière) sont ordonnés.

On peut étudier la variable "effectifs cumulés" avec un modèle de régression logistique, par exemple avec la procédure catmod de SAS.

Cette procédure donne un test global de la différence entre les courbes de réponse cumulées. On a présenté sur cette figure les réponses des 4 génotypes, et la valeur du Chi2 de l'effet kinétine (et la probabilité associée) .

Pas de surprise, l'effet est toujours hautement significatif pour les génotypes 2 et 3. Par contre, il le devient au seuil 5% pour le génotype 4 alors qu'il ne l'était pas quand on considérait les catégories non ordonnées. Cette hypothèse permet donc un gain de puissance, mais elle doit être biologiquement justifiée.

RÉSUMÉ



<u>facultatif</u> La famille des modèles à catégories ordonnées est grande. On peut choisir un modèle simple ou généralisé selon que les seuils sont supposés fixés ou variables, cumulatif ou séquentiel, avec un lien logit ou cloglog ($\log(-\log(1-p))$) ou $-\log(-\log p)$ ... (le lien cloglog est associée à une distribution “extreme minimal value”, alors que le lien $-\log(-\log p)$ est associé à une distribution “extreme maximal value”).

Remarque : le modèle cumulatif logistique simple est appelé “proportional odds model”. Le modèle cumulatif simple à lien cloglog est appelé “proportional hazard model” ou encore “Grouped Cox Model”.