Comprendre les modèles linéaires généralisés

Emmanuel Paradis
paradis@isem.univ-montp2.fr

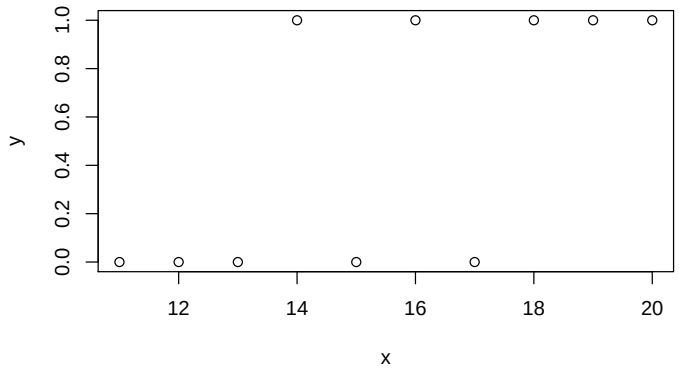
15 novembre 2004

1 Le cas d'une variable de Bernoulli (0/1)

Supposons que l'on observe une variable binaire y (habituellement appelée la réponse) qui prend la valeur 0 ou 1. Nous avons, associée à chaque observation de y, une variable continue x (habituellement appelée le prédicteur). Le graphe ci-dessous montre un échantillon avec $n = 10$. Il semble y avoir une relation entre x et y mais on voudrait tester statistiquement si elle est significative. Intuitivement, on peut utiliser les tests statistiques courants, par exemple avec un test de t pour comparer les moyennes de x pour le sous-échantillon où $y = 0$ et celui où $y = 1$. Cette approche peut être satisfaisante dans des cas particuliers, mais elle s'avère limitée notamment si l'on a non pas un mais plusieurs prédicteurs $(x_1, x_2, ...)$, certaines pouvant être continues, d'autres catégoriques, et que l'on doit envisager des possibles interactions entre ces variables.

Une approche générale est fournie par les modèles linéaires généralisés (*generalized linear models* ou GLM). Faisons l'hypothèse que y suit une loi de Bernoulli de paramètre p (c'est-à-dire $y = 1$ avec la probabilité p, et $y = 0$ avec la probabilité $1-p$), le graphique suggère alors que p est influencé par x : p semble plus petit pour les faibles valeurs de x. On va chercher à formaliser cette apparente relation avec un modèle du genre $p = \beta x$. Une difficulté réside dans le fait que p étant entre 0 et 1, la droite résultante de ce modéle peut être très plate si x varie grandement et β risque d'être proche de 0. Ce problème est résolu en transformant p tel que son domaine de variation soit entre $-\infty$ et $+\infty$, par exemple en utilisant la fonction logit définie par :

$$\text{logit}p = \ln\frac{p}{1-p}$$



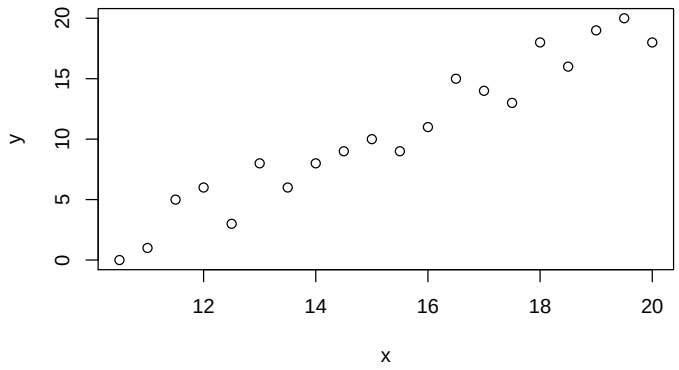
Quand p tend vers 0, logit(p) tend vers $-\infty$, et quand p tend vers 1, logit(p) tend vers $+\infty$. On a donc maintenant le modèle logit$p = \beta x$ qui définit pour chaque valeur de x (x_i) une valeur de p (p_i). Ainsi pour chaque x_i on peut en déduire la probabilité que $y = 1$ et celle que $y = 0$. Le calcul de la vraisemblance suit très logiquement.

Avant de généraliser du cas d'une variable binaire aux modèles linéaires généralisés, voyons un autre exemple, celui d'une régression de Poisson.

2 La régression de Poisson

Supposons maintenant que notre variable y est un effectif : ce genre de variable est très courant, par exemple en écologie avec des données d'abondance (effectifs de populations). La loi de Poisson est appropriée pour analyser ce genre de données. Rappelons qu'une variable de Poisson ne prend que des valeurs entières entre 0 et $+\infty$, que sa distribution est déterminée par un seul paramètre (noté λ), et que la moyenne est égale à la variance et au paramètre λ ($\mu = \sigma^2 = \lambda$).

Le graphique suivant montre un échantillon avec $n = 20$, x est une variable continue, et y est une variable entière. La relation entre x et y est ici évidente, et l'on peut être tenté d'ajuster une simple droite des moindres carrés entre ces deux variables. Le test statistique sur la pente de cette droite sera sûrement significatif, mais le résultat ainsi acquis manquera forcément



de généralité : il est clair que la droite des moindres carrés ne pourra être extrapolée aux faibles valeurs de x puisqu'elle prédirait alors des valeurs de y négatives.

Prenons la même approche que précédemment dans le cas d'une variable binaire : si l'on considère que y suit une loi de Poisson, alors le graphe suggère une relation entre x et λ et on obtient le modèle $\lambda = \beta x$. On se trouve de nouveau devant une difficulté qui est cette fois que λ est contraint d'être entre 0 et $+\infty$: la difficulté est levée en prenant le logarithme de λ, ce qui donne le modèle $\ln\lambda = \beta x$. Pour chaque valeur x_i on a donc une valeur λ_i qui permet de calculer la probabilité que $y_i = 0, 1, 2, ... + \infty$; le calcul de la vraisemblance suit sans problème.

3 Généralisation

Quel est le point commun entre les deux exemples précédents ? Dans le cas de la variable binaire, on est arrivé à formuler une relation linéaire entre une transformation de p et la variable x. Dans le cas de la variable de Poisson, on a fait la même opération avec λ. Notons maintenant que p et λ correspondent aux moyennes attendues de la variable y dans les deux cas respectifs. Pour comprendre ce que cela signifie, imaginons que l'on fixe la valeur de x et que l'on fasse un grand nombre de mesures indépendantes de y. Dans le cas où y est binaire, la moyenne des ces y_i tendra vers p ; dans le cas où y est Poisson,

cette moyenne tendra vers λ (voir ci-dessus).

On a donc dans les deux cas réalisé une modélisation linéaire de la valeur attendue de la moyenne des y, ce qui peut s'écrire $g(\mu_i) = \beta x_i$, avec $\mu_i = E(y_i)$. C'est le premier pas dans la formulation d'un GLM. Pour que la théorie des GLM soit applicable à d'autres distributions (en particulier normale et gamma), un second pas est nécessaire : c'est l'introduction du paramètre de dispersion ϕ qui contrôle la variance de la distribution des y_i autour de μ_i. Dans le cas où y suit une loi binomiale ou de Poisson, on a en fait $\phi = 1$ (on verra plus loin que ce paramètre ϕ peut être ajouté dans le modèle et estimé avec les autres paramètres, il détermine alors la sur-dispersion $[\phi > 1]$ ou la sous-dispersion $[\phi < 1]$ des observations).

3.1 Cas d'une variable normale

Lorsque les observations y_i sont normalement distribuées, l'analyse avec un GLM est exactement similaire, si x est continue, à une régression linéaire des moindres carrés, c'est-à-dire que les observations sont également distribuées de part et d'autre de la valeur prédite par μ_i. La variance de cette distribution est donnée par σ^2, on a donc $\phi = \sigma^2$.

3.2 Cas d'une variable gamma

Une variable gamma est distribuée de façon continue entre 0 et $+\infty$, la forme de la distribution est contrôlée par deux paramètres : le taux λ, et l'échelle ν. C'est une généralisation de la loi exponentielle : cette dernière est obtenue en fixant $\nu = 1$. La moyenne d'une variable de gamma est égale à $1/\lambda$ (tout comme une variable exponentielle). Le paramètre ν contrôle la dispersion des observations relativement à une loi exponentielle, on a en fait $\phi = 1/\nu$.

3.3 Variables de Bernoulli et binomiale (la régression logistique)

On a vu dans le permier exemple le cas où y est une variable de Bernoulli prenant les valeurs 0 ou 1, les GLM peuvent s'appliquer aussi aux cas où y est un ensemble d'observations 0 ou 1, c'est-à-dire une variable binomiale. Admettons que pour chaque valeur de x_i il y a associé n_i observations. La moyenne attendue pour chaque observation est $p_i n_i$. Ce type de régression

est appellé régression logistique. L'exemple du § 1 est donc simplement un cas particulier de régression logistique avec tous les $n_i = 1$.

3.4 Formulation des modèles linéaires généralisés

Revenons maintenant à la formulation générale des GLM. On a besoin de :

- un type de distribution pour la réponse ; ce choix va déterminer la fonction de dispersion $Var(y_i)$ des observations autour de la valeur attendue μ_i en fonction du paramètre de dispersion ϕ ;
- un choix pour la fonction de lien g (qui est une "transformation" de la moyenne attendue, et non pas des valeurs observées) ;
- un modèle linéaire qui spécifie la relation entre la réponse moyenne μ_i et des prédicteurs x_i en fonction de paramères β.

Le tableau ci-dessous récapitule les propriétés des différentes distributions couramment utilisées pour modéliser des réponses avec des GLM. La fonction de lien g indiquée est celle qui est la plus fréquemment utilisée, mais d'autres peuvent être utilisées (*cf.* ci-dessous).

Réponse	ϕ	**lien** g	$Var(y_i)$
Normale	σ^2	identité	σ^2
Gamma	$1/\nu$	log	μ^2/ν
Poisson	1	log	μ
Binomiale	1	logit	$\mu(1-\mu)/n_i$

4 Sur- et sous-dispersion avec les variables binomiales et de Poisson

On vient juste de voir que le paramètre de dispersion ϕ est "libre" pour une réponse normale ou gamma, c'est-à-dire qu'il varie en fonction des données et doit donc être estimé. Ce n'est pas le cas des réponses de type Poisson ou binomiale où ϕ est contraint d'être égal à 1 : la dispersion des observations est alors fonction de la moyenne attendue μ_i. Mais il arrive très fréquemment que cette relation ne soit pas vérifiée, et la variance observée est supérieure (sur-dispersion) ou inférieure (sous-dispersion) à celle prédite, le phénomène de sur-dispersion étant le plus courant.

Il y a deux causes (non exclusives) généralement invoquées pour expliquer la présence de sur-dispersion avec ce types de données : la non-indépendance des observations, et la présence d'hétérogénéité non prise en compte par le modèle. Dans le deuxième cas, la sur-dispersion peut souvent être diminuée en choisissant des prédicteurs plus appropriés (c'est-à-dire en trouvant un meilleur modèle, ce qui bien sûr dépend uniquement de l'utilisateur). Quelque soit la cause, il faut bien noter que la sur-dispersion ne biaise pas l'estimation des paramètres β mais sous-estime les erreurs-standards associées : on risque donc de conclure injustement à un effet significatif (risque de première espèce accru). Le biais inverse se produira en présence de sous-dispersion. Il convient donc de corriger éventuellement ces biais en introduisant dans le modèle le paramètre ϕ (le plus souvent via les options du logiciel) qui sera estimé avec les autres paramètres. Si $\phi > 1$, il y a sur-dispersion ; si $\phi < 1$, il y a sous-dispersion. Notez qu'il n'y a pas de tests formels pour tester l'hypothèse $\phi = 1$, mais on considère en pratique que si $\phi > 2$ ($\phi < 0.5$), il convient de prendre soin de corriger les effets de la sur-dispersion (sous-dispersion).

5 Extensions des modèles linéaires généralisés

Les GLM peuvent être utilisés avec d'autres distributions. Mise à part les quatre distributions discutées ici, les distributions gaussienne inverse et binomiale négative sont disponibles sur la plupart des logiciels.

Comme mentionné plus haut, le choix d'une fonction lien est à la discrétion de l'utilisateur. Le choix d'un lien donné peut s'avérer plus pertinent qu'un autre. Par exemple, dans le cas d'une réponse normale le lien logarithmique permet de spécifier un modèle non-linéaire $\mu_i = \exp(\beta x_i)$.

D'une façon plus générale, le modèle linéaire peut être remplacé par une fonction non-linéaire définissant ainsi un modèles non-linéaires généralisés (GNLM), par opposition aux modèles non-linéaires traditionnels qui utilisent une approche des moindres carrés. Les GNLM présentent de vastes perspectives, les modèles non-linéaires étant généralement considérés comme plus réalistes. Divers modèles non-linéaires peuvent être comparés avec l'AIC. Il faut noter cependant que les GNLM sont encore peu répandus dans les logiciels de statistiques.

Finalement, il convient de signaler que l'approche utilisée dans la régression logistique peut être facilement appliquée dans tout problème où le paramètre d'intérêt est formulé sous forme d'une probabilité (spéciation, extinction, sur-

vie, ...), et a donc un large potentiel d'application en biologie évolutive.

6 Bibliographie

Altham P.M.E. 1998. Introduction to generalized linear modelling in R. Mimeographed document, University of Cambridge, Statistical Laboratory. (http ://www.statslab.cam.ac.uk/˜pat)

Crawley M.J. 1993. GLIM for ecologists. Blackwell Scientific Publications, Oxford.

Lindsey J.K. 1996. Parametric statistical inference. Clarendon Press, Oxford.

McCullagh P. & Nelder J.A. 1989. Generalized linear models (second edition). Chapman & Hall, London.

Nelder J.A. & Wedderburn R.W.M. 1972. Generalized linear models. *Journal of the Royal Statistical Society A* 135 :370-384.

SAS Institute Inc. 1996. SAS/STAT® Software : changes and enhancements through release 6.11. SAS Institute Inc., Cary, NC.