Introduction au Modèle Linéaire Généralisé (Generalized Linear Model ; GLM)

Eric Wajnberg (wajnberg@sophia.inra.fr)
UE7 – Université de Nice-Sophia-Antipolis
Octobre 2011

Rappel 1 – la régression linéaire simple ou multiple

On cherche à ajuster le modèle suivant sur un jeu de données :

$$y = a_1 x_1 + a_2 x_2 + \cdots + b$$

L'équation, linéaire, est celle de la droite (ou du plan, etc.) qui passe « au mieux » dans les points. Les paramètres a_1, a_2, …, et b peuvent être estimés par la méthode des moindres carrées (« least-square method »), leurs estimations sont celles qui minimisent la quantité suivante, qui correspond à la somme des carrés des écarts au modèle :

$$\sum_{i=1}^{n} \left(y_i - (a_1 x_{1i} + a_2 x_{2i} + \cdots + b)\right)^2$$

Sous R, en supposant l'exemple d'une régression linéaire simple avec une variable explicative *x* et une variable à expliquer *y*, la syntaxe suivante peut être adoptée :

Pour calculer et tracer la régression, avec la fonction `lm()`, par exemple :

```
res=lm(y~x)
summary(res)
plot(res)
plot(y~x)
abline(res)
```

Pour rajouter, par exemple, un intervalle de confiance à la régression sur le graphe :

```
new=data.frame(x=seq(min(x), max(x), length=20))
p=predict(res, new, interval= "conf", level=0.99)
points(p[,2]~new$x, type= "l")
points(p[,3]~new$x, type= "l")
```

Pour calculer, par exemple, un intervalle de confiance sur les paramètres de la régression :

```
confint(res, level=0.95)
```

etc., tout ceci pouvant être généralisé à plusieurs dimensions.

L'ensemble des résultats statistiques peut être fourni sous la forme d'un tableau d'ANOVA, avec la fonction `anova()` :

```
anova(res)
```

qui décompose et teste la variance de *y* due à la régression par rapport aux variances résiduelle et totale.

Rappel 2 – l'Analyse de la variance à un facteur ou plus

On cherche ici à tester l'effet d'un (ou de plusieurs) facteur(s) qualitatif(s) sur une variable quantitative. Plus précisément, l'objectif est de comparer statistiquement les moyennes de la variable quantitative mesurée dans chacune des modalités de la – ou des – facteur(s) contrôlé(s). Pour ce faire, la variance totale de la variable mesurée est décomposée en variance due au(x) facteur(s) contrôlé(s) et variance résiduelle, et ces deux variances sont comparées par un test F.

Sous R, en supposant l'exemple d'une Analyse de la Variance (ANOVA) sur une variable *x* avec un facteur ou plusieurs facteurs « factor1 », « factor2 », etc., la syntaxe suivante peut être adoptée :

Pour calculer l'ANOVA, avec la fonction `aov()`, par exemple :

```
res=aov(x~factor1+factor2+etc.)
summary(res)
plot(res)
plot(x~factor1+factor2+etc.)
```

Pour calculer des statistiques pour chaque modalité d'un facteur, par exemple :

```
tapply(x, factor1, mean)
tapply(x, factor1, var)
tapply(x, factor1, summary)
…
```

Le modèle linéaire général

Il se trouve que la régression linéaire présentée ci-dessous (cf. rappel 1) aurait aussi pu être calculée avec la fonction `aov()`! En effet, la sortie de :

```
anova(lm(y~x))
```

est la même que la sortie de :

```
summary(aov(y~x))
```

La raison est que ces deux modèles, celui de la régression linéaire simple (ou multiple) et celui de l'ANOVA à un (ou plusieurs) facteur(s), sont juste des cas particuliers d'un cadre général que l'on nomme le « modèle linéaire général », et qui a pour caractéristiques les éléments suivants :

1) Le modèle à ajuster est de la forme :

$$y = a_1 x_1 + a_2 x_2 + \cdots + b$$

Dans le cadre de la régression linéaire simple ou multiple, c'est l'équation que l'on a vu plus haut. Dans le cadre de l'ANOVA, les facteurs à tester sont préalablement codés sous forme matricielle, et on retombe sur ce schéma linéaire de base.

(Pour ceux que cela intéresse, il est possible de récupérer la forme matricielle qui sert à l'ajustement dans le cas d'une ANOVA. La fonction `model.matrix()` fournie l'information cherchée, *e.g.*, `model.matrix(x~factor1)`).

2) Plus précisément, et c'est là la remarque la plus importante, le modèle à ajuster est en fait dans tous les cas de la forme :

$$E(y) = a_1 x_1 + a_2 x_2 + \cdots + b + \varepsilon$$

Où $E(y)$ représente l'espérance de la variable y (qui est équivalente à sa moyenne, ou plus précisément sa moyenne attendue) et ε représente le terme d'erreur (*i.e.*, de bruit) non contrôlé qui doit impérativement suivre une distribution normale et de même variance. C'est-à-dire que la variance de ce terme d'erreur doit être indépendante de la valeur des différentes variables x_1, x_2, etc. (qui, rappelons-le, peuvent correspondre au codage de facteurs quantitatifs).

Accessoirement, cette hypothèse forte (normalité et égalité des variances) peut être vérifiée graphiquement en étudiant le graphe des résidus, c'est-à-dire le graphe des valeurs de la variable *y* auxquelles on retranche leurs prédictions issues du modèle. Les sorties de R fournissent ce genre de graphes diagnostiques.

Il existe, sous R, une fonction qui permet d'ajuster un modèle linéaire général, et qui rend donc caduques les fonctions `lm()` et `aov()`. Il s'agit de la fonction `glm()`. Attention cependant, le nom de cette fonction ne veut pourtant pas dire « general linear model » contrairement à ce qu'on pourrait penser, mais signifie « general**ized** linear model », qui est un cadre encore plus général que nous verrons plus bas.

Ainsi, une régression linéaire simple (ou multiple) peut se calculer, par exemple, comme suit :

```
res=glm(y~x)
summary(res)
plot(res)
anova(res, test="F")
```

Et une ANOVA avec exactement la même syntaxe :

```
res=glm(x~factor1+factor2+etc.)
summary(res)
plot(res)
anova(res, test="F")
```

Ces syntaxes retournent évidement les mêmes résultats que ceux issus des fonction `lm()` et `aov()`, respectivement.

Le bestiaire du modèle linéaire général

Nous avons vu qu'une régression linéaire est un cas particulier du modèle linéaire général. Elle consiste à chercher à expliquer une variable quantitative par une autre variable quantitative. Nous avons vu qu'une régression linéaire multiple l'est également. Elle consiste à chercher à expliquer une variable quantitative par plusieurs variables quantitatives. Enfin,

nous avons vu qu'une ANOVA l'est également. Elle consiste à chercher à expliquer une variable quantitative par une ou plusieurs autre(s) variable(s) qualitative(s). Dans tous les cas, on cherche à expliquer une variable quantitative par une ou plusieurs variables quantitatives ou qualitatives. On tombe donc sur le tableau suivant, qui donne les différents types de modèle linéaire général que l'on peut rencontrer.

Si	**Alors on a**
Il y a une variable explicative qui est quantitative	Une régression linéaire simple
Il y a plusieurs variables explicatives qui sont toutes quantitatives	Une régression linéaire multiple
Il y a une variable explicative qui est qualitative	Une ANOVA à un facteur
Il y a plusieurs variables explicatives qui sont toutes qualitatives	Une ANOVA à plusieurs facteurs
Il y a une combinaison de variables explicatives quantitatives et qualitatives	Une analyse de covariance (ANACOV ou ANCOVA)

Et dans tous les cas, la syntaxe dans R est la même et l'interprétation des résultats également (cf. exemples ci-dessus).

Le modèle linéaire généralisé

Nous avons vu que le modèle linéaire général repose sur une hypothèse forte : le terme d'erreur suit une loi normale et de même variance. Nous avons pourtant parfois (souvent..) le besoin d'expliquer des variables (et donc leurs erreurs) qui ne suivent pas ce pré-requis.

Prenons deux exemples, celui où la variable *y* mesurée est un pourcentage (*e.g.*, le pourcentage de mâles ou de femelles dans une population où il n'y évidement que des mâles et des femelles), et celui où la variable *y* mesurée est un comptage (*e.g.*, nombre d'œufs pondus par une poule).

Dans ces deux cas – et d'autres encore – le simple modèle linéaire présenté ci-dessus ne peux pas convenir et ce pour au moins deux raisons importantes :

1) la principale et la plus importante est que la distribution de la variable à expliquer n'est pas compatible avec le modèle linéaire présenté ci-dessus. Par exemple, dans le cas du comptage, seules des valeurs entières et positives ou nulles peuvent être mesurées, alors que le modèle linéaire simple ci-dessus pourra malgré tout prédire des valeurs décimales et/ou négatives. De même, un pourcentage est par définition compris dans l'intervalle [0, 1] (ou [0%, 100%] ; on ne peut avoir moins de 0% ou plus de 100% de mâles ou de femelles) alors que le modèle linéaire ci-dessus pourra prédire des valeurs qui pourront sortir de cet intervalle. Par ailleurs considérer la variable à expliquer (et son erreur) comme suivant une loi normale suppose une distribution symétrique autour de la moyenne, alors que ce n'est très probablement pas le cas, par exemple, du comptage où la majorité des valeurs mesurées seront par exemple plus fréquemment vers zéro ou un que vers 100 ou 200.

2) L'autre raison est que, dans le modèle linéaire simple (général), les variables prédictives ont un effet linéaire sur la variable mesurée (effet qui se traduit par les coefficients de régression), or ces effets ne sont peut-être pas linéaires en réalité. Par exemple, le nombre d'œufs pondus par une poule ne change peut-être pas linéairement avec son âge.

Pour tenir compte de ces points plusieurs solutions s'offrent à nous. La plus répandue consiste à trouver une transformation mathématique de la variable à expliquer pour la rendre normale (et son erreur avec) et pour en stabiliser les variances. On parle de transformations « normalisantes ». Plusieurs sont connues. Ces transformations ne sont pas toutes efficaces, et leur effet normalisant est parfois difficile à quantifier. Par ailleurs, il reste évidement préférable d'utiliser les données d'origine plutôt que leurs valeurs transformées, ne serait-ce que pour rendre l'interprétation des résultats plus aisée. C'est dans ce cadre que se développe le modèle linéaire généralis**é** (General**ized** Linear Model ; GLM).

L'idée reste d'utiliser une transformation mathématique sur la variable à expliquer *y* mais en tenant compte cette fois-ci de la véritable distribution des erreurs (par exemple, une loi de Poisson dans le cas de comptages ; une loi Binomiale dans le cas de pourcentages, etc.). Ceci implique entre autre que les paramètres ne sont alors plus estimés par la simple méthode des moindres carrés – comme dans le modèle linéaire général – mais par une autre méthode d'estimation : la méthode dite du « maximum de vraisemblance ». La fonction mathématique utilisée pour transformer la variable à expliquer est appelée « fonction de lien » (« link function »), et plusieurs peuvent être utilisées selon la distribution réelle de la variable d'intérêt (et de son erreur). Du coup, le modèle à ajuster devient :

$$f\big(E(y)\big) = a_1x_1 + a_2x_2 + \cdots + b$$

Où *f*(…) est la fonction de lien.

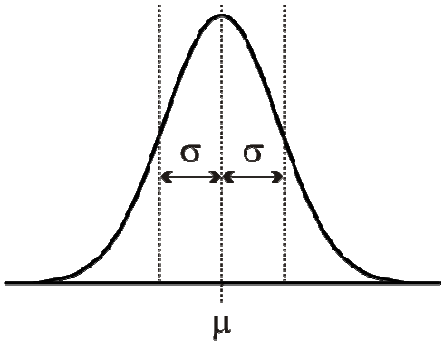
Cas Gaussien

Un GLM peut être virtuellement utilisé quelle que soit la distribution de la variable à expliquer *y*, y compris si cette variable suit une loi normale de même variance. Ainsi donc, le modèle linéaire général vu ci-dessus (test-t, régression simple ou multiple, ANOVA, etc.) est lui-même un cas particulier du modèle linéaire généralisé.

Rappelons qu'une loi normale a une forme « en cloche », symétrique de part et d'autre de sa valeur maximale qui correspond à la moyenne de la variable étudiée. Elle décrit la distribution d'une variable continue. Son équation est :

$$\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Où μ est la moyenne, et σ est l'écart-type.



Des variables telles que le poids et la taille des individus, la surface de cellules, le taux de glucide, entre autres exemples, sont supposées suivre une loi normale.

En toute logique, un GLM utilisé pour analyser une variable suivant une loi normale aura pour fonction de lien la fonction identité $f(y)=y$. Et l'on retombe donc sur le modèle linéaire général pris ici comme un cas particulier :

$$E(y) = a_1x_1 + a_2x_2 + \cdots + b$$

Sous R, nous l'avons vu, l'ajustement pourra se faire de la manière suivante :

```
res=glm(x~factor1+factor2+etc., family=gaussian)
summary(res)
plot(res)
anova(res, test="F")
```

l'argument « `family=gaussian` » peut être omis, car le cas gaussien est le cas par défaut de la fonction `glm()`.

Cas Poissonien

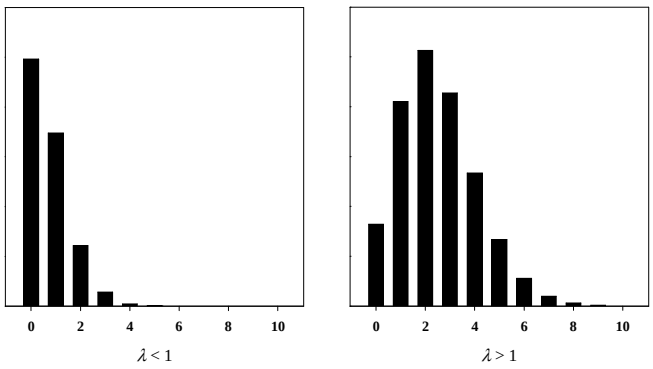
Nous voulons à présent analyser une variable de comptage (par exemple, comme nous l'avons vu ci-dessus, le nombre d'œufs pondus par une poule).

La variable à analyser suit cette fois-ci une loi de Poisson qui décrit des variables discrètes positives ou nulles. Selon cette loi, la probabilité d'observer une valeur k vaut :

$$\frac{e^{-\lambda}\lambda^k}{k!}$$

où λ est la moyenne de la distribution et est également sa variance (l'égalité de la moyenne et de la variance est d'ailleurs une caractéristique de cette loi).

Si la moyenne est inférieure à 1, cette distribution aura une forme en « i », sinon elle aura une forme dissymétrique, avec une « queue » sur la droite.



La fonction de lien utilisée pour analyser une variable suivant une loi de Poisson (*i.e.*, un comptage) est généralement la fonction log : $f(y)=\log(y)$, et le modèle s'appelle dans ce cas un modèle « log-linéaire » :

$$\log\big(E(y)\big) = a_1x_1 + a_2x_2 + \cdots + b$$

Qui peut se réécrire sous la forme suivante :

$$E(y) = e^{a_1x_1+a_2x_2+\cdots+b}$$

Dans le cas d'un comptage ne pouvant prendre que des valeurs positives, cette fonction de lien est logique. Elle permet d'avoir des valeurs prédites par le modèle qui s'étendent de $-\infty$ à $+\infty$, ce qui reste interprétable alors que ca ne l'aurait pas été avec un simple modèle linéaire général, comme nous l'avons vu ci-dessus.

Sous R, l'ajustement pourra se faire de la manière suivante :

```
res=glm(x~factor1+factor2+etc., family=poisson)
summary(res)
plot(res)
anova(res, test="Chisq")
```

Il est effectivement préconisé dans ce cas d'utiliser un test de χ^2 (« `Chisq` ») plutôt qu'un test F pour vérifier la significativité des effets.

Il se peut parfois que la distribution observée de la variable à expliquer ne suive pas exactement une loi de Poisson telle qu'attendue, mais que sa variance (*i.e.*, sa dispersion) soit plus forte que celle issue d'une simple loi de Poisson. Il est possible de prendre en compte cette possible « sur-dispersion », en remplaçant « `family=poisson` » par « `family=quasipoisson` » dans la syntaxe ci-dessus. Dans ce cas, un terme supplémentaire est estimé à partir des données, Ce terme, qualifié de paramètre de dispersion, indique l'augmentation de la variance observée par rapport à celle attendue d'une loi de Poisson. Une valeur de 1.0 signifie que les données suivent bien une loi de Poisson, une valeur supérieure à 1.0 indique que la variance observée est supérieure à celle attendue d'une loi de Poisson.

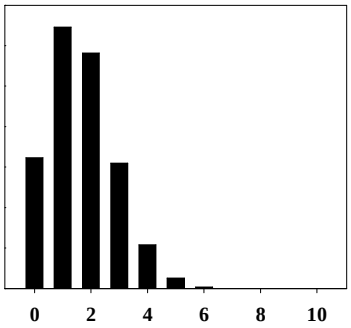
Cas Binomial

La variable mesurée est à présent une proportion (entre 0.0 et 1.0) ou, de manière équivalente, un pourcentage (entre 0.0 % et 100.0 %), par exemple, comme nous l'avons vu ci-dessus, la proportion de mâles ou de femelles dans une population.

La variable à analyser suit cette fois-ci une loi Binomiale qui décrit le nombre de fois où un événement – parmi deux possibles – se produit lorsque l'on répète l'observation (par exemple : nombre de femelles parmi 100 individus ; nombre d'individus vivants sur 30 ; nombre de faces sur 55 jets d'une pièce de monnaie). Selon cette loi, sur n répétitions, si pour chacune d'elle la probabilité d'observer l'événement étudié est p, la probabilité que l'événement se produise k fois est :

$$\frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

Par exemple, le graphe suivant donne la distribution du nombre d'as attendus (p=1/6) au cours de 10 jets d'un dès.



Généralement, on ne s'intéresse pas tant au nombre de fois où l'on observe l'événement, mais à sa fréquence parmi toutes les répétitions : k/n (exemples : proportion de femelles dans l'échantillon ; proportion d'individus vivants ; proportion de faces sur plusieurs jets d'une pièce de monnaie). Dans ce cas, la moyenne de cette proportion vaut p, et sa variance vaut $p(1\text{-}p)/n$.

La fonction de lien utilisée pour analyser une variable suivant une loi Binomiale (*i.e.*, un pourcentage) est généralement la fonction logit : $f(y)=\log(y/(1\text{-}y))$, et le modèle s'appelle dans ce cas une régression « logistique » :

$$\log\left(\frac{E(y)}{1-E(y)}\right) = a_1x_1 + a_2x_2 + \cdots + b$$

Qui peut se réécrire sous la forme suivante :

$$E(y) = \frac{1}{1+e^{-(a_1x_1+a_2x_2+\cdots+b)}}$$

Sous R, un tracé de la fonction logit peut être obtenu de la manière suivante :

```
curve(log(x/(1-x)), 0, 1)
```

Dans le cas d'un pourcentage restant dans l'intervalle [0, 1], cette fonction de lien est logique également. Elle permet d'avoir des valeurs prédites par le modèle qui s'étendent de $-\infty$ à $+\infty$, ce qui reste interprétable alors que ca ne l'aurait pas été avec un simple modèle linéaire général, comme nous l'avons vu ci-dessus.

Sous R, l'ajustement pourra se faire à partir de deux manières différentes de présenter les données.

Dans la première, la variable à expliquer est codée sur deux colonnes, disons y_1 et y_2, l'une contenant le nombre de fois où un évènement (*e.g.*, mâles) a été observé, l'autre contient le nombre de fois où l'autre évènement (*e.g.*, femelles) est observé, dans chaque situation. Dans

la seconde, la variable à expliquer (*e.g.*, l'observation d'un mâle) est codée sur une colonne binaire qui indique si l'événement est observé (1) ou non (0) pour chaque mesure.

Dans le premier cas, l'ajustement pourra se faire de la manière suivante :

```
res=glm(cbind(y1,y2)~factor1+factor2+etc., family=binomial)
summary(res)
plot(res)
anova(res, test="Chisq")
```

Dans le second cas :

```
res=glm(y~factor1+factor2+etc., family=binomial)
summary(res)
plot(res)
anova(res, test="Chisq")
```

Il est effectivement préconisé dans le cas d'une régression logistique également d'utiliser un test de χ^2 (« `Chisq` ») plutôt qu'un test F pour vérifier la significativité des effets.

Comme dans le cas d'un modèle log-linéaire (pour une variable de comptage suivant une loi de Poisson ; voir ci-dessus) Il se peut parfois que la distribution observée de la variable à expliquer ne suive pas exactement une loi Binomiale telle qu'attendue, mais que sa variance (*i.e.*, sa dispersion) soit plus forte que celle issue d'une simple loi Binomiale. Ici aussi, il est possible de prendre en compte cette possible « sur-dispersion », en remplaçant « `family=binomial` » par « `family=quasibinomial` » dans les syntaxes ci-dessus. Dans ce cas, comme dans le cas précédent, un terme supplémentaire est estimé à partir des données, Ce terme, qualifié de paramètre de dispersion, indique l'augmentation de la variance observée par rapport à celle attendue d'une loi Binomiale. Son interprétation est la même quand dans le cadre du cas poissonien.

D'autres cas encore : une brève synthèse

Comme son nom l'indique, le modèle linéaire généralisé est un outil qui peut être utilisé dans de nombreuses situations, afin d'analyser des variables qui présentent différents types de distribution statistique. Nous avons vu, rapidement, les cas où la variable suit une loi Normale, une loi de Poisson ou une loi Binomiale, cas qui restent les plus fréquents. D'autres lois de distribution peuvent être considérées, conduisant à chaque fois à l'utilisation de fonctions de lien différentes. Le tableau suivant donne une petite synthèse des principales situations rencontrées.

Distribution	Type de données	Type de GLM	Fonction de lien
Normale	Variable suivant une loi normale	Modèle linéaire général	Identité : $f(y)=y$
Poisson	Comptage	Modèle log-linéaire	Log : $f(y)=\log(y)$
Binomiale	Pourcentage	Régression logistique	Logit : $f(y)= \log(y/(1-y))$
Gamma	Durée	Modèle Gamma avec fonction de lien inverse	Inverse : $f(y)=1/y$