Partie I : Le Modèle Linéaire *Général*

UE7Statistiques pour la biologie
P. Coquillard, 2015

On ne confondra pas, malgré la ressemblance, le modèle linéaire général et le modèle linéaire généralisé. Ils reposent tout deux sur des hypothèses et des méthodes de résolutions différentes.

Quelques réflexions sur la régression et l'ANOVA.

Nous avons dit à plusieurs reprises que la régression consistait en l'établissement d'un modèle du type :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$$

Où les X_k sont les variables descriptives (= explicatives) et Y la variable dépendante. Reprenons le fichier tomates aov.txt. Sa structure est simplement en 3 colonnes: un numéro d'observation, une mesure (Long) et une dose d'extrait dans laquelle les tomates ont germé (Dose).
En principe, face à ce genre de données incluant un facteur catégoriel comme variable explicative, on choisi l'ANOVA comme test statistique. Mais ce n'est pas obligatoire, on peut aussi opter pour une régression linéaire.
Ainsi, après avoir « monté » les données en mémoire dans R (**E1**), on effectuera une régression linéaire au moyen de lm().
On remarquera que le summary() de la régression (**E3**) fournit un intercept + 4 valeurs dosages : D2 à T. Pour chacune d'entre elle un test *t* de significativité nous dit si ces coefficients ont une action réelle (non due au hasard) sur la valeur de *Y*. On constate que c'est le cas. Par contre, D1 n'est pas mentionnée alors qu'elle existe bel et bien dans le jeu de données. En réalité, D1 a été prise comme valeur de référence pour calculer l'intercept de la régression, toutes les autres valeurs étant mises à 0. L'intercept est donc simplement la moyenne de D1[1]. Mais R fourni de surcroit (ce n'est pas le cas de tous les logiciel) la statistique F : 111 et une p-value: < 2.2e-16.

Effectuons maintenant le test ANOVA classique (**E1 et E4**). Surprise : les coefficients sont exactement les mêmes que précédemment, ainsi que la statistique F… !!!

Il n'y a donc aucune différence entre les deux analyses !??

- Dans la régression, la variable catégorielle est recodée au moyen de 0 et de 1, ce qui implique que l'ordonnée à l'origine de chaque catégorie est comparée à l'intercept du groupe de référence (dans notre cas D1), car l'intercept est défini comme la valeur moyenne lorsque tous les autres prédicteurs = 0.

- Dans l'analyse de variance, la moyenne de chaque catégorie est comparée à la moyenne générale (totale).

```
E1
mydata <- read.table("tomates aov.txt", sep="\t", header =T)
names(mydata)=c("Obs","Long","Dose")
attach(mydata)
Dosage = factor(Dose)
boxplot(Long~Dose)

A <- lm(Long~Dose)
summary(A)
coefficients(A)

B <- aov(Long~Dose)
coefficients(B)
```

[1] On peut changer cela en utilisant la fonction relevel(Dose, x = ″T″) qui fait de T la référence de l'analyse.

```
E2
> mydata
    Obs Long Dose
1     1 27.6    T
2     2 26.6    T
3     3 19.9    T
4     4 22.9    T
5     5 16.2    T
6     6 24.6    T
7     7 26.4    T
…/…
> Dose
  [1] T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T  T
 [26] T  T  T  T  T  D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D1
 [51] D1 D1 D1 D1 D1 D1 D1 D1 D1 D1 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2
 [76] D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D2 D3 D3 D3 D3 D3 D3 D3 D3 D3
[101] D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D3 D4 D4 D4 D4 D4
[126] D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4 D4
Levels: D1 D2 D3 D4 T
```

Pour conclure, l'ANOVA vous indique seulement si au moins deux moyennes sont significativement différentes entre elles (sans vous dire lesquelles) alors que la régression vous signifie quels sont les coefficients de régression qui influent significativement sur la variable dépendante *Y*).

La régression linéaire simple, multiple, l'ANOVA (MANOVA incluse) et l'ANCOVA (MANCOVA incluse) constituent le ***modèle linéaire général*** dont ils sont tous des applications particulières.
Le modèle linéaire général repose sur les hypothèses fondamentales de la régression par les moindres carrés ordinaires :

- Normalités des erreurs (ce qui implique aussi l'absence de lien entre la variance des réponses et leur moyenne…)
- Homogénéité des variances.

Il n'y a aucune différence fondamentale entre les méthodes. C'est le même code qui est appelé dans R ! Ainsi vous pouvez employer indifféremment lm() ou aov().

Considérations :
Si on souhaite réaliser une prédiction à partir de nouvelles valeurs, ou bien si la question est « *lesquelles parmi les catégories ont un effet sur le résultat et si oui, combien* ? » on utilisera plutôt la régression.
Si la question est « *dans quelle mesure les différences entre catégories influent sur le résultat* ? », on utilisera plutôt l'ANOVA

Mais attention ! Il s'agit bien de la même méthode. L'une ne donne pas plus d'informations que l'autre. On retiendra :

Si les variables explicatives sont toutes quantitatives : régression multiple (lm())
Si les variables explicatives sont toutes qualitatives : ANOVA (aov())
Si les variables explicatives sont un mix des deux types : ANCOVA (lm())
Si on examine les réponses de 2 ou plus variables dépendantes : MANOVA ou MANCOVA

Mais il ya encore mieux : le *modèle linéaire généralisé* pouvant traiter un encore plus grand nombre de cas, vous pourrez employer glm() pour tous vos traitements statistiques paramétriques… mais il faudra apporter quelques précisions.

E3 : Régression sur tomates aov.txt

```
➢   summary(A)
Call:
lm(formula = Long ~ Dose)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5500 -2.6492  0.2067  2.9433  7.9033

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.3900     0.6873  26.756  < 2e-16 ***
DoseD2       -6.0933     0.9720  -6.269 3.94e-09 ***
DoseD3      -10.4633     0.9720 -10.764  < 2e-16 ***
DoseD4      -10.3433     0.9720 -10.641  < 2e-16 ***
DoseT         6.3600     0.9720   6.543 9.70e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
                '.' 0.1 ' ' 1

Residual standard error: 3.765 on 145 degrees of freedom
Multiple R-squared:  0.7537,    Adjusted R-squared:  0.7469
F-statistic:   111 on 4 and 145 DF,  p-value: < 2.2e-16

➢   coefficients(A)
(Intercept)      DoseD2      DoseD3      DoseD4       DoseT
  18.390000   -6.093333  -10.463333  -10.343333    6.360000
```

E4: ANOVA sur tomates aov.txt

```
➢   summary(B)
Call:
aov(formula = Long ~ Dose)

Terms:
                    Dose Residuals
Sum of Squares  6289.976  2055.065
Deg. of Freedom        4       145

Residual standard error: 3.764686
Estimated effects may be unbalanced

             Df Sum Sq Mean Sq F value Pr(>F)
Dose          4   6290  1572.5     111 <2e-16 ***
Residuals   145   2055    14.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
                '.' 0.1 ' ' 1

➢   coefficients(B)

(Intercept)      DoseD2      DoseD3      DoseD4       DoseT
  18.390000   -6.093333  -10.463333  -10.343333    6.360000
```

Partie II : préambule au Modèle Linéaire *Généralisé*

Il s'agit en fait d'une famille de modèles dus à Nelder et Wedderburn (1972).

Les modèles linéaires généralisés permettent d'étudier la liaison entre une variable dépendante (= réponse) Y et un ensemble de variables explicatives (= prédicteurs) : $X_1, \ldots, X_K$.

Ils englobent entre autres :

- le modèle linéaire général (Loi normale: régression multiple, analyse(s) de la variance et analyse(s) de la covariance)
- le modèle log-linéaire (Loi de Poisson)
- la régression logistique (Loi binomiale)
- les cas de distribution Gamma, Gamma inverse, Bernoulli…

Les modèles linéaires généralisés ont 3 composantes

1. La variable de réponse Y à laquelle est associée une loi de probabilité. C'est la **composante aléatoire,** appartenant à la famille des exponentielles de paramètres $(y, \theta\ \boldsymbol{\phi}, \boldsymbol{\omega})$ dont la loi de densité est de la forme :

$$\boldsymbol{f(y_i, \theta_i, \phi, \omega_i) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)}\omega_i + c(y_i, \phi, \omega_i)\right\}}$$

Où ϕ est le paramètre de dispersion (connu) et θ le paramètre canonique (inconnu). a, *b* et *c* sont des fonctions connues et spécifiées selon la loi exponentielle en question (Poisson, binomiale, etc). On a par ailleurs :

$$E(Y) = b'(\theta) = \mu$$
$$\mathrm{Var}(Y) = \phi b''(\theta) = \phi V(\mu)$$

2. Les variables explicatives $X_1, \ldots, X_K$ utilisées comme prédicteurs définissent, sous forme d'une combinaison linéaire $\eta = \beta X$, la **composante déterministe.** V est la *fonction de variance* qui lie la moyenne à la variance de la réponse.

3. **Le lien** (g) décrit la relation fonctionnelle entre la combinaison linéaire des variables ($X_1, \ldots, X_K$) et l'espérance mathématique de la variable de réponse : $E(Y) = \mu = g(\eta)$.

La Composante aléatoire

Notons ($Y_1, \dots, Y_n$) un échantillon aléatoire de taille *n* de la variable de réponse Y, les variables aléatoires $Y_{1, \dots,} Y_n$ étant supposées indépendantes :

Yi peut être binaire (succès-échecs, présence-absence). **Loi de Bernoulli, loi binomiale**,
Yi peut être distribuée selon une **loi de Poisson** (comptages),
Yi peut être distribuée selon une **loi normale.**

Dans ce dernier cas, on a **:** $\epsilon \sim \mathcal{N}(0, \sigma^2)$, le préditeur linéaire habituel :

$$\eta_i = \beta_0 + \beta_1 X_{11} + \cdots + \beta_k X_{ki},$$
la fonction de lien est $g(\mu_i) = \mu_i,$
et la fonction de variance $V(\mu_i) = 1.$

On retombe ainsi dans le cas du modèle linéaire général, qui n'est donc qu'un cas particulier (celui de la loi normale caractérisée par l'absence de dépendance entre moyenne et variance et par sa symétrie) du modèle linéaire généralisé !

La suite dans le cours IV….

Annexe. Quelques composantes de la famille exponentielle

Distribution	θ(μ)	b(θ)	a(ϕ)
Normale N(μ,σ²)	μ	θ²/2	σ²
Bernouilli	$\log\frac{\mu}{1-\mu}$	$\log(1+e^{\theta})$	1
Poisson	log(μ)	e^{θ}	1
Gamma Γ(μ,ν)	1/μ	-log(-θ)	1/ ν²

Distribution	E(Y)=b'(θ)	Var(Y)=b''(θ)a(ϕ)
Normale N(μ,σ²)	μ = θ	σ²
Bernouilli	$\mu = \frac{e^{\theta}}{1+e^{\theta}}$	μ(1-μ)
Poisson	$\mu = e^{\theta}$	μ
Gamma Γ(μ,ν)	$\mu = -\frac{1}{\theta}$	$\frac{\mu^2}{\nu}$