MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

Le modèle linéaire généralisé avec R : fonction glm()

Sébastien BALLESTEROS

UMR 7625 Ecologie Evolution
Ecole Normale Supérieure
46 rue d'Ulm
F-75230 Paris Cedex 05

sebastien.ballesteros@biologie.ens.fr

1) Approche

Le modèle linéaire avec R

régression linéaire, *ANOVA* sont des cas particuliers d'un **même** modèle statistique, le **modèle linéaire** que l'on peut écrire :

$$y_{tj} = m_t + e_{tj}$$

partie fixe, linéaire

partie aléatoire, normale

erreurs indépendantes entres elles, suivant chacune une loi normale d'espérance nulle et de même variance.

- quantitative → régression linéaire
- qualitative → ANOVA
- quantitative et qualitative → ANCOVA

- t est l'indice d'un traitement. Les differents facteurs pouvant intervenir dans sa définition sont contôlés, ils sont donc fixes, non aléatoires.
- j est un indice de répétition (pouvant ne pas exister explicitement)
- m_t est l'espérance de y_{tj}

$$y_{tj} \sim N(m_t, \sigma^2),\ \{y_{tj}\} \text{ indépendants}$$

Sous R : lm(variable à expliquer ~ variable(s) explicative(s), ...)

Non application du modèle linéaire

Influence de la dose d'un poison (disulfide de carbone) sur la mortalité de cafards.

Données

```
>cafards<-read.table("cafards.dat", header=TRUE)
> cafards
  ldose total morts
1 1.691    59     6
2 1.724    60    13
3 1.755    62    18
4 1.784    56    28
5 1.811    63    52
6 1.837    59    53
7 1.861    62    61
8 1.884    60    60
```

On note :

i = 1...8 groupes

n_i = taille du $i^{ème}$ groupe

N_i = nombre de morts dans le groupe i

x_i = dose de poison

Avec modèle linéaire, on peut étudier :

$$Y_i = \frac{N_i}{n_i}$$

$$Y_i = a + b\,x_i + E_i$$

Problèmes :

- Les valeurs prédites peuvent sortir de la zone [0,1]
- homoscédasticité

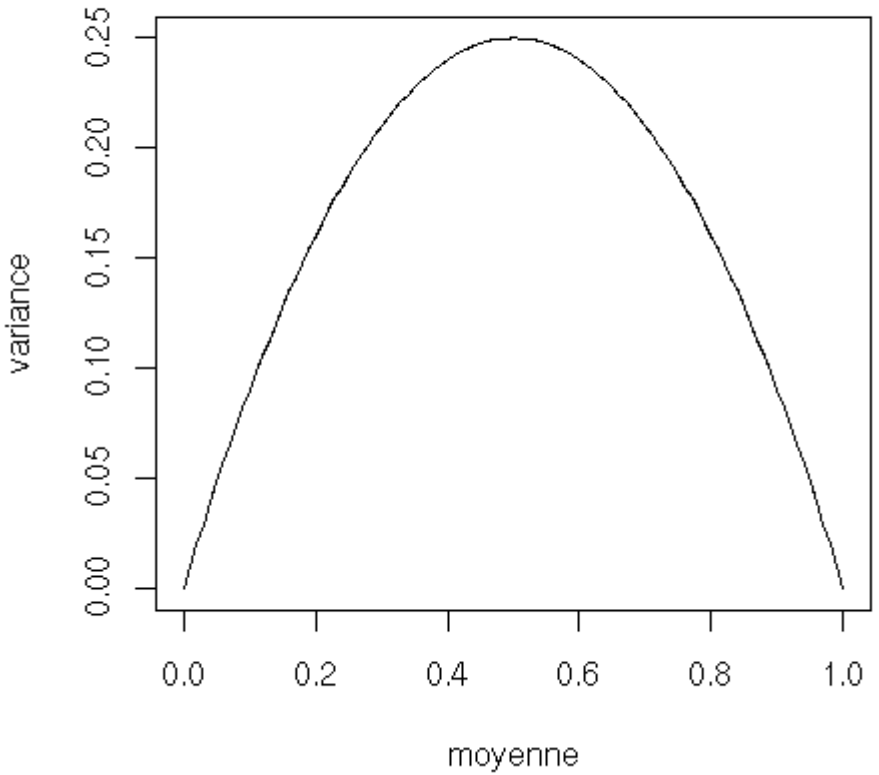
Homoscédascticité

Rappel, on note :
i = 1...8 groupes
n_i = taille du $i^{ème}$ groupe
N_i = nombre de morts dans le groupe i
x_i = dose de poison

$$N_i \sim B(n_i, \pi_i)$$

$$Y_i = \frac{N_i}{n_i} = \hat{\pi}_i$$

$$E(Y_i) = \pi_i \quad ; \quad V(Y_i) = \frac{\pi_i(1-\pi_i)}{n_i}$$



On est dans une situation hétéroscedastique par construction

Longtemps, on a utilisé une transformation pour stabiliser la variance

$$Z_i = \arcsin(\sqrt{Y_i})$$

Marche bien quand $n_i \approx cst$

Modèle linéaire généralisé

Modèle linéaire généralisé

Modèle

i = 1...8 groupes
n_i = taille du $i^{ème}$ groupe
Y_i = nombre de morts dans le groupe i
x_i = dose de poison
π_i = proba de mourir dans le groupe i

$$Y_i \sim B(n_i, \pi_i)$$

On veut garder la simplicité d'interprétation du modèle linéaire

$$\pi_i = a + b\,x_i$$

Problème, π_i doit rester entre 0 et 1
On ne modélise pas directement π_i mais $g(\pi_i)$

g : fonction de lien

$$g(\pi_i) = a + b\,x_i$$

$g : [0,1] \to \mathbb{R}$ π est astreint entre 0 et 1 mais on laisse a et b faire ce qu'ils veulent
monotone croissante

Fonction de lien logit (logistique)

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Modèle linéaire généralisé sous R

Modèle

$$Y_i \sim B(n_i, \pi_i)$$

$$g(\pi_i) = a + b\, x_i$$

i = 1…8 groupes

n_i = taille du $i^{ème}$ groupe

Y_i = nombre de morts dans le groupe i

x_i = dose de poison

π_i = proba de mourir dans le groupe i

Sous R : glm(variable à expliquer ~ variable(s) explicative(s), type de loi (fonction de liens), ...)

```
>cafards<-read.table("cafards.dat", header=TRUE)
>attach(cafards)
> cafards
  ldose total morts
1 1.691    59     6
2 1.724    60    13
[...]
```

```
> y<-cbind(morts,total-morts)
> model<-glm(y~ldose, family=binomial(link="logit"))
```

```
> y.prop<-morts/total
> model.prop<-glm(y.prop~ldose, weights=total, family=binomial(link="logit"))
```

Summary

```
> summary(model)
```

```
Call:
glm(formula = y ~ ldose, family = binomial(link = "logit"))
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5878  -0.4085   0.8442   1.2455   1.5860
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.740      5.182  -11.72   <2e-16 ***
ldose         34.286      2.913   11.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
AIC: 41.314
```

```
Number of Fisher Scoring iterations: 4
```

Estimation des paramètres et test sur les paramètres

$$Y_i \sim B(n_i, \pi_i)$$

Estimation des paramètres par maximum de vraisemblance

$$\nu(y_1, y_2, \dots, y_I) = \prod_{i=1}^{I} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

$$\log(\nu) = \sum_i \binom{n_i}{y_i} + \sum_i y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i)$$

Si $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ est linéaire, pas très dure a maximiser

$$g(\pi_i) = a + b\, x_i$$

logit: fonction de lien canonique, ca va bien se passer avec elle

Les estimateurs du max de vraisemblance sont **asymptotiquement gaussien**

➡ On a la loi des estimateur, on peut faire des tests

Estimation des paramètres et test sur les paramètres

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.740      5.182  -11.72   <2e-16 ***
ldose         34.286      2.913   11.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Interprétation des paramètres

$$g(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)=a+b\,x_i$$

$$\pi_i=\frac{\exp(a+b\,x_i)}{1+\exp(a+b\,x_i)}$$

$\pi(\cdot)=\frac{\exp(a)}{1+\exp(a)}$ Proba de décès quand on ne met pas de poison

$\pi(?)=0.5 \qquad x_i=\frac{-a}{b}$ Dose létale à 50%

$\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}=\exp(b(x_2-x_1))$ Si x augmente de 1 unité, log(b)=log(odd ratio)

Deviance

Modèle saturé : la moyenne de la variable à expliquer est défini par l'observation elle même. $E(Y_i)=y_i$

$$Y_i \sim B(n_i, \pi_i)$$

$$E(Y_i)=n_i\pi_i \rightarrow \pi_i = y_i/n_i$$

La probabilité d'observer l'observation vaut $\binom{n_i}{y_i}\frac{y_i}{n_i}^{y_i}(1-\frac{y_i}{n_i})^{n_i-y_i}$. On a donc la vraisemblance du modèle saturé ν_{sat}

```
> LVsat <- sum(log(dbinom(morts,total,morts/total)))        [1] -13.09902
```

Modèle nul : $E(Y_i)$=cst estimé comme la moyenne p0 par max de vraisemblance

```
> p0<-sum(morts)/sum(total)
> LV0 <- sum(log(dbinom(morts,total,p0)))          [1] -155.2002
```

Deviance nul = $-۲*\log(\nu_{restr}/\nu_{sat})$

```
> dev0 = 2*(LVsat-LV0)          [1] 284.2024
```

Modèle x : estimé par max de vraisemblance

```
> LVx <- sum(log(dbinom(morts,total,predict(model,type="response"))))
[1] -18.65681
```

Deviance residuelle = $-2*\log(\nu_{restr}/\nu_{sat})$

```
> devx = 2*(LVsat-LVx)      [1] 11.11558
```

```
Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
AIC: 41.314
```

calcul de l'AIC = 2LVx + 2p

```
> aicx = -2*LVx + 2*2        [1] 41.31361
```

Deviance : test de modèles emboîtés

Une stratégie intuitive consiste à comparer deux modèles emboîtés sur la base d'une mesure de la qualité de leur ajustement aux données.

$$2(\log(\nu_{complet}) - \log(\nu_{restr})) \approx \chi^2_{(r-r_0)}$$

Test du rapport de vraisemblance

Test du *modèle restreint* contre le *modèle complet*

Le test de rapport de vraisemblance est correcte seulement asymptotiquement pour de grands jeux de données.

Deviance = $-2 * \log(\nu_{restr} / \nu_{sat})$

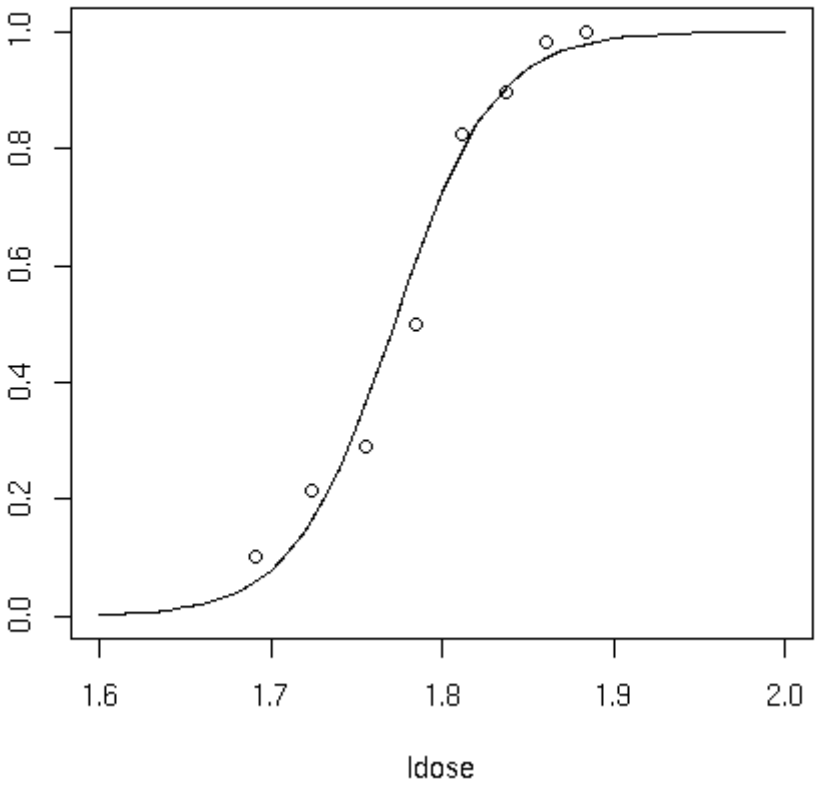
LV0 = -155.2002 2*(LVsat-LV0) = 284.2024
LVx = -18.65681 2*(LVsat-LVx) = 11.11558
LVsat = -13.09902 2*(LVx-LV0) = 273.0869

```
> anova(model,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: y
Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     7    284.202
ldose  1  273.087         6     11.116 2.411e-61
```

Test du *modèle constant* contre le *modèle complet*

Bilan cafards

```
> y<-cbind(morts,total-morts)
> model<-glm(y~ldose, family=binomial(link="logit"))
> summary(model)
[...]
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.740      5.182  -11.72   <2e-16 ***
ldose         34.286      2.913   11.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
AIC: 41.314

Number of Fisher Scoring iterations: 4
```




```
> anova(model,test="Chisq")
[...]
Analysis of Deviance Table
Model: binomial, link: logit
Response: y
Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      7    284.202
ldose  1  273.087         6     11.116 2.411e-61
```

predict(model,type="response")

2) généralisation du modèle linéaire généralisé

généralisation

Type d'erreur
- Binomial (Bernoulli)
- Poisson

Fonction de lien
- logit, probit,...
- log,...

Prédicteur linéaire

$$g(\pi_i) = x_i \theta$$

x_i : vecteur des covariables

θ : vecteur des paramètres

Fonction de lien probit

Ce modèle est particulièrement naturel lorsque la variable binaire Y dont on peut observer les réalisations n'est que l'expression simplifiée d'une autre variable continue Y* impossible à observer, parfois seulement conceptuelle. Par exemple, dans un contexte médical, une problématique classique est le classement d'un patient dans le système de catégories (malade et sain).

Modèle de seuil, en fonction du seuil succès ou échec.

La fonction de lien est la réciproque de la fonction de répartition d'une loi normale.

$$g(\pi)=\Phi^{-1}(\pi)$$

$$\Phi^{-1}(\pi)=a+b\,x_i$$
$$\pi_i=PHI(a+b\,x_i)$$
$$\pi_i=P(Z\leqslant a+b\,x_i)$$

Y est une variable de comptage

On peut faire des transformation de variable pour tenter d'appliquer le modèle linéaire mais dans le cadre du glm, on prend une loi de poisson

$$Y_i \sim P(\lambda_i)$$

$$\lambda_i > 0 \quad ; \quad E(Y_i) = V(Y_i) = \lambda_i$$

$$g : \mathbb{R}+ \to \mathbb{R}$$

monotone croissante

Vraisemblance

$$P(Y_i = y) = \exp(-\lambda) \frac{\lambda_i^{\,y}}{y!}$$

$$\nu(y_1, \ldots, y_n) = \prod_i \exp(-\lambda_i) \frac{\lambda_i^{\,y_i}}{y_i!}$$

$$\log(\nu) = -\sum_i \lambda_i + \sum_i y_i \log(\lambda_i) - \sum_i \log(y_i!)$$

On maximise la vraisemblance, ca serait bien si log(λ) linéaire et donc on prend :

$$g(\lambda) = \log(\lambda)$$

$$g(\lambda_i) = x_i \theta$$

Modèle saturé

$$\hat{\lambda}_i = y_i$$

Modèle linéaire généralisé sous R

Sous R : glm(variable à expliquer ~ variable(s) explicative(s), type de loi (fonction de liens), ...)

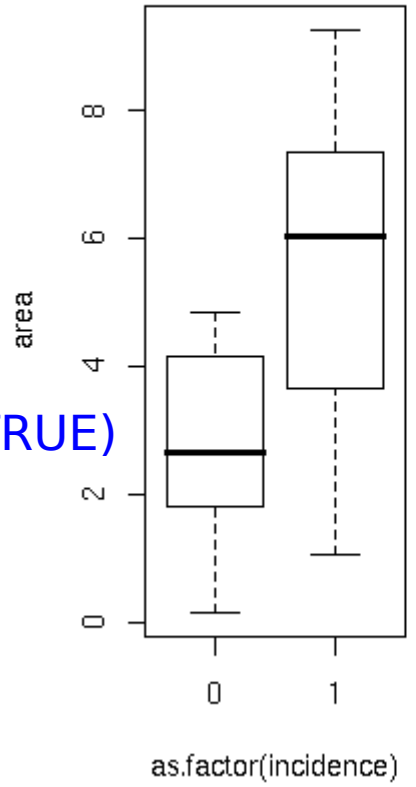
```
>?family
[...]
    binomial(link = "logit")
    gaussian(link = "identity")
    Gamma(link = "inverse")
    inverse.gaussian(link = "1/mu^2")
    poisson(link = "log")
    quasi(link = "identity", variance = "constant")
    quasibinomial(link = "logit")
    quasipoisson(link = "log")
[...]
```

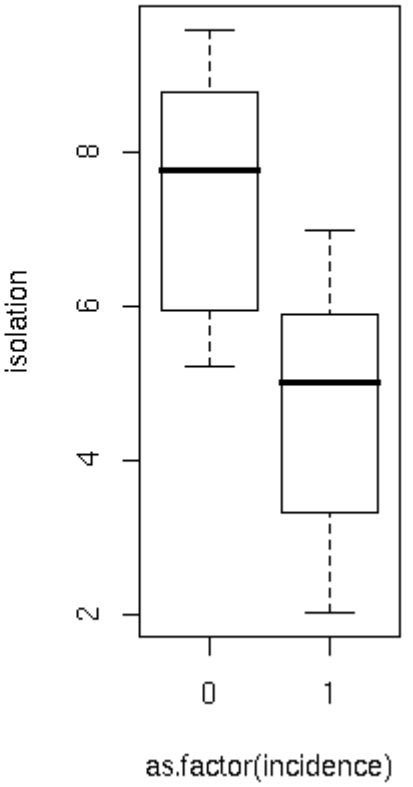
```
> model<-glm(y~ldose, family=binomial(link="logit"))
```

3) Exemples

Données binaires : nidification d'une espèce d'oiseau

Îles de taille et de distance au continent variable (area et isolation)
On regarde sur chacune des îles si l'espèce d'oiseau est présente (1) ou absente (0)






```
> island<-read.table("isolation.dat", header=TRUE)
> attach(island)
> names(island)
[1] "incidence" "area"      "isolation"
```

$Y_i \sim B(\pi_i)$ Loi de Bernoulli

$g(\pi_i) = a + b\,x_{1i} + c\,x_{2i} + d\,x_{1i}\,x_{2i}$

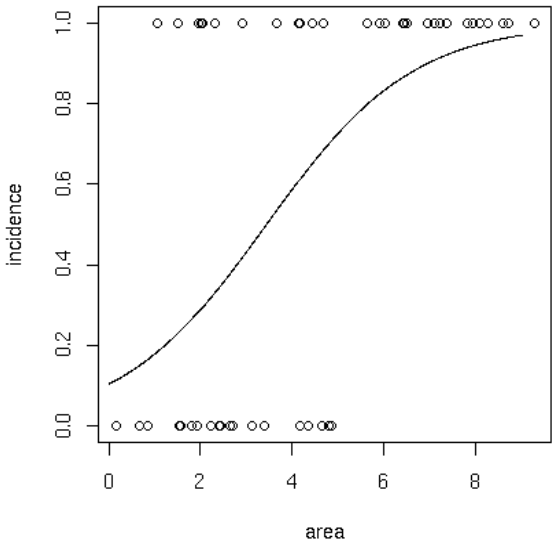
```
> model1<-glm(incidence~area*isolation,family=binomial(link="logit"))
```

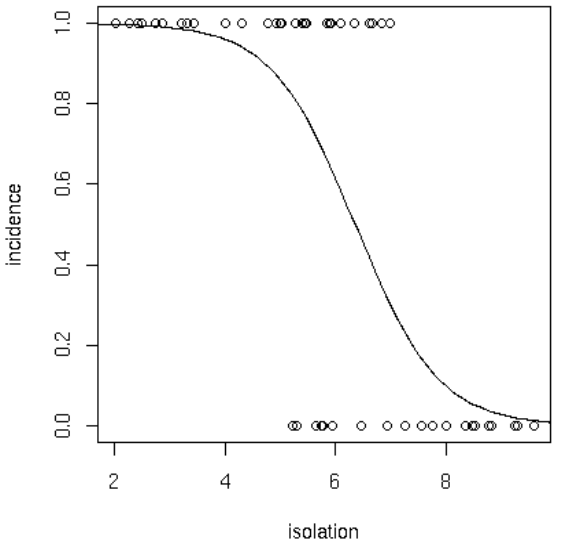
nidification d'une espèce d'oiseau

```
> model1<-glm(incidence~area*isolation,family=binomial(link="logit"))
> model2<-glm(incidence~area+isolation,family=binomial(link="logit"))
> anova(model2,model1,test="Chisq")

Analysis of Deviance Table

Model 1: incidence ~ area + isolation
Model 2: incidence ~ area * isolation
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        47    28.4022
2        46    28.2517  1   0.1504    0.6981


> anova(model1,test="Chisq")

Analysis of Deviance Table
Model: binomial, link: logit
Response: incidence
Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         49     68.029
area          1   17.857       48     50.172 2.382e-05
isolation     1   21.770       47     28.402 3.073e-06
area:isolation  1    0.150       46     28.252     0.698
```

nidification d'une espèce d'oiseau

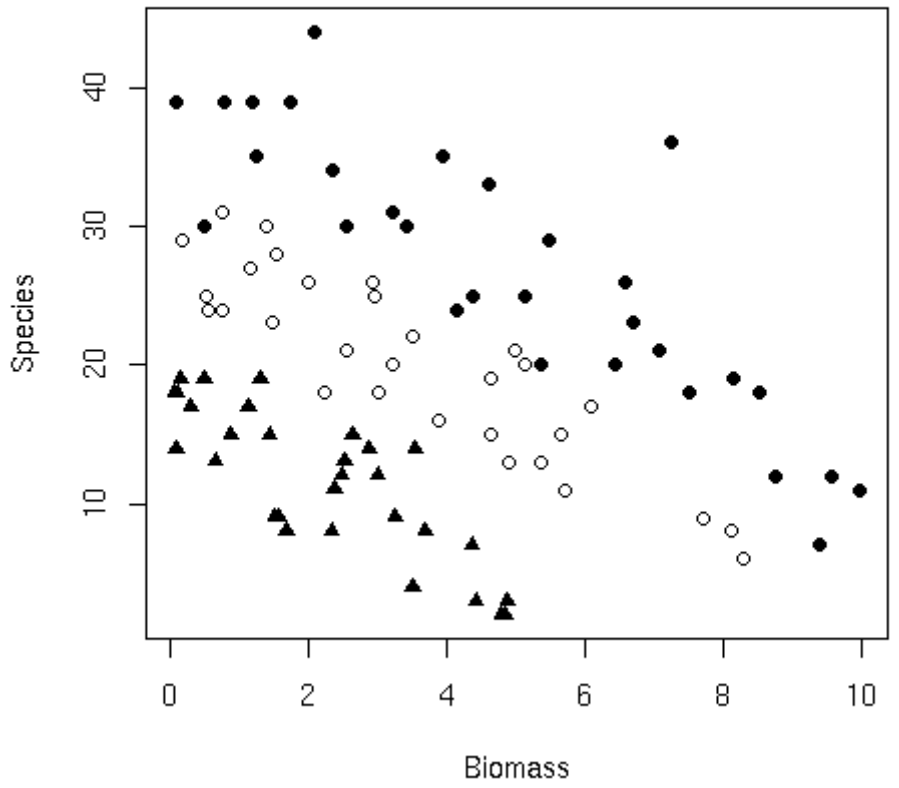
```
> summary(model2)

Call:
glm(formula = incidence ~ area + isolation, family = binomial(link = "logit"))
[...]
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.6417     2.9218   2.273  0.02302 *
area          0.5807     0.2478   2.344  0.01909 *
isolation    -1.3719     0.4769  -2.877  0.00401 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
[...]
Number of Fisher Scoring iterations: 6
```





comptage de diversité spécifique

```
> species<-read.table("species.txt", header=TRUE)
> attach(species)
> names(species)
[1] "pH"      "Biomass" "Species"
```



La pente de la relation entre le nombre d'espèces et la biomasse dépend elle du pH ?

comptage de diversité spécifique

On pose

$$Y_i \sim P(\lambda_i)$$

$$\log(\lambda_i) = \mu + \alpha_i + \beta x_{ij} + \gamma_i x_{ij}$$

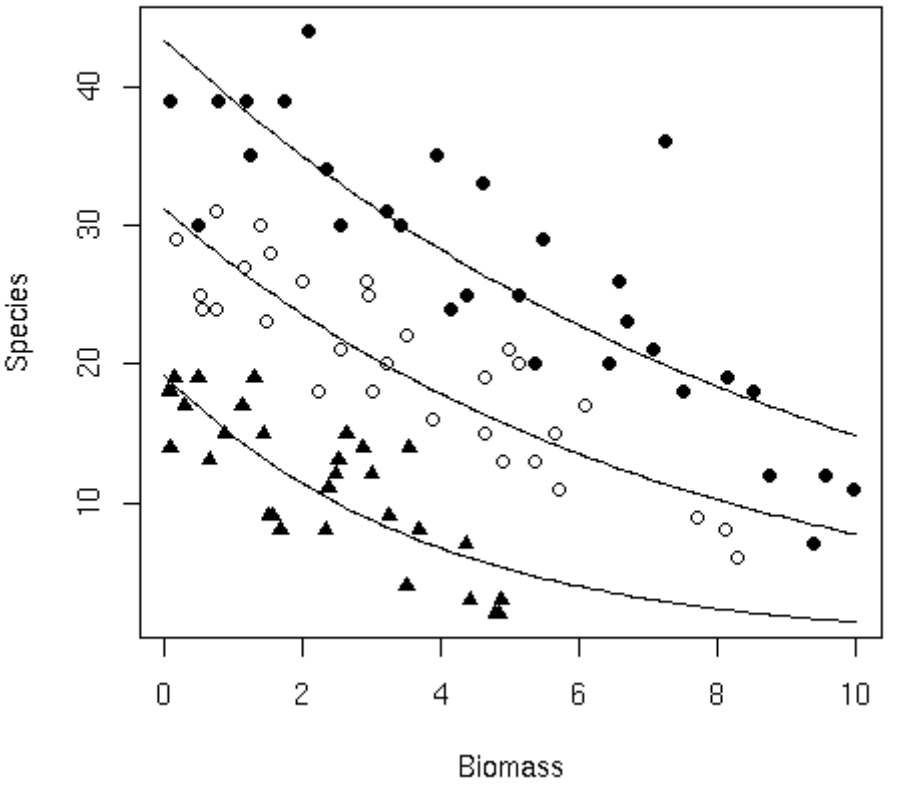
$$E(Y_i) = \lambda_i$$

```
> model1<-glm(Species~Biomass*pH,family=poisson(link="log"))
> anova(model1,test="Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: Species

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     89     452.35
Biomass    1    44.67        88     407.67 2.328e-11
pH         2   308.43        86      99.24 1.059e-67
Biomass:pH  2    16.04        84      83.20 3.288e-04
```

On retient le modèle le plus complexe

comptage de diversité spécifique

```
> summary(model1)

Call:
glm(formula = Species ~ Biomass * pH, family = poisson(link = "log"))
[...]
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
Biomass       -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
pHmid         -0.33146    0.09217  -3.596 0.000323 ***
Biomass:pHlow -0.15503    0.04003  -3.873 0.000108 ***
Biomass:pHmid -0.03189    0.02308  -1.382 0.166954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: 514.39

Number of Fisher Scoring iterations: 4
```



References :

• Modern Applied Statistics with S Fourth edition ;
W. N. Venables and B. D. Ripley

•The R book ; Michael J. Crawley

•Introductory Statistics With R ; Peter Dalgaard

•Le modèle linéaire ; Camille Duby

•Statistique inférentielle ; JJ. Daudin, S.Robin

•http://www.bio.ic.ac.uk/research/mjcraw/therbook/index.htm