Sélection de modèles avec l'AIC et critères d'information dérivés

Renaud LANCELOT et Matthieu LESNOFF

Version 3, Novembre 2005

Ceci n'est pas une revue exhaustive mais une courte introduction sur l'utilisation du critère d'information d'Akaike (AIC) pour la sélection de modèles.

1. Cadre d'application

La démarche présentée ici s'applique aux modèles estimés par une méthode du maximum de vraisemblance : les analyses de variance, les régressions linéaires multiples, les régressions logistiques et de Poisson peuvent rentrer dans ce cadre (Burnham et Anderson, 2002). En toute rigueur, les modèles estimés par des méthodes basées sur la quasi-vraisemblance (par exemple, les modèles GEE) n'en font pas partie mais des adaptations ont été proposées (Lebreton et al., 1992).

2. Problématique

Il est fréquent d'être confronté à des données provenant d'enquêtes dans lesquelles les variables explicatives n'ont pas fait l'objet d'un plan d'observation (i.e., échantillonnage en fonction de ces variables explicatives). Même si c'est le cas, des variables doivent souvent être relevées pour leur rôle connu ou supposé dans le phénomène étudié, sans qu'il soit possible de les contrôler (ex. : température externe ou hygrométrie dans une enquête sur la pathologie respiratoire). L'analyste est alors confronté à de nombreux modèles possibles, correspondant aux différentes combinaisons de variables explicatives (effets principaux et interactions).

Le test du rapport des vraisemblances (*likelihood ratio test*) est souvent utilisé pour comparer des modèles deux à deux. Il ne s'applique qu'à des modèles emboîtés (dérivant l'un de l'autre par ajout ou suppression de termes) et il est supposé que les deux modèles comparés ajustent correctement les données (tests globaux, examen des résidus, comparaison des valeurs observées et prédites,…) (McCullagh and Nelder, 1989). Quand de nombreux modèles doivent être comparés entre eux, le risque de rejeter l'hypothèse nulle alors qu'elle est vraie augmente substantiellement (fig. 1).

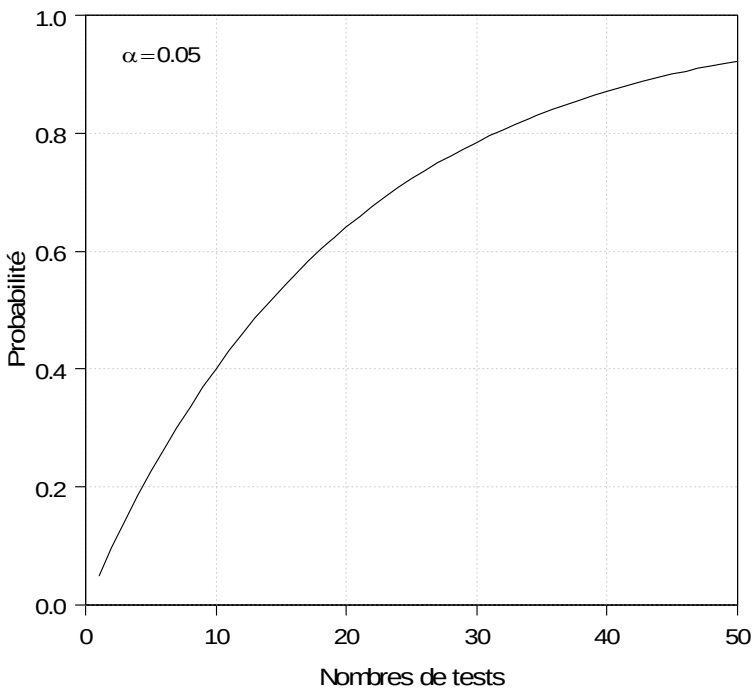


Figure 1. Risque de rejeter au moins une fois l'hypothèse nulle alors qu'elle est vraie quand le nombre de comparaisons augmente

Pour résoudre ce problème, une solution possible (il y en a d'autres) consiste à comparer les modèles en utilisant le critère d'information d'Akaike (Akaike, 1974) :

$$AIC = - 2 * \log(L) + 2 * k$$

où L est la vraisemblance maximisée et k le nombre de paramètres dans la modèle. Avec ce critère, la déviance du modèle (-2 * log(L)) est pénalisée par 2 fois le nombre de paramètres. L'AIC représente donc un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètres possible).

3. Application

3.1. Principe

- En toute rigueur, il est nécessaire que les modèles comparés dérivent tous d'un même modèle « complet » (Burnham et Anderson, 2002). Ce modèle doit être intégré dans la liste des modèles comparés.
- Il est nécessaire de vérifier que ce modèle « complet » ajuste correctement les données : test de la qualité de l'ajustement de Pearson, examen des résidus, comparaison des valeurs observées et prédites,…
- Il est souhaitable que la liste des modèles comparés soit établie avant l'analyse selon des critères de plausibilité biologique.
- Le meilleur modèle est celui possédant l'AIC le plus faible.
- Quand le nombre de paramètres k est grand par rapport au nombre d'observations n, i.e., si n / k < 40, il est recommandé d'utiliser l'AIC corrigé : $AICc = AIC + \frac{2 * k * (k+1)}{n-k-1}$

  (Hurvich et Tsai, 1995).

3.2. Mise en œuvre

Sous R, la fonction générique `AIC()` permet de calculer l'AIC pour plusieurs types de modèles, notamment les modèles linéaires (estimés avec la fonction `lm()`) et linéaires généralisés (estimés avec la fonction `glm()`).

Cependant, l'$AICc$ n'est pas directement disponible et il n'est pas possible de spécifier la taille de l'échantillon, ce qui peut poser des problèmes avec les modèles linéaires généralisés sur données groupées. Une fonction `sic()` (acronyme pour *some information criteria*) a ainsi été écrite et mise à disposition des utilisateurs dans le package **`metomet`** (à partir de la version 0.5-0), disponible sur le site GuR.

Cette fonction consiste en fait en une fonction générique `sic()` (programmation de style S3) avec deux méthodes pour les objets de classe « lm » et « glm », et une méthode par défaut gérant les autres cas par un message d'erreur.

Certains packages proposent des fonctions permettant de calculer l'AIC et d'autres critères d'information. C'est le cas par exemple du package `aod`, pour les modèles de régression béta-binomiale et binomiale négative.

4. Exemples

4.1. Modèles linéaires généralisés

On utilise un jeu de données proposé dans l'aide de la fonction `glm()`.

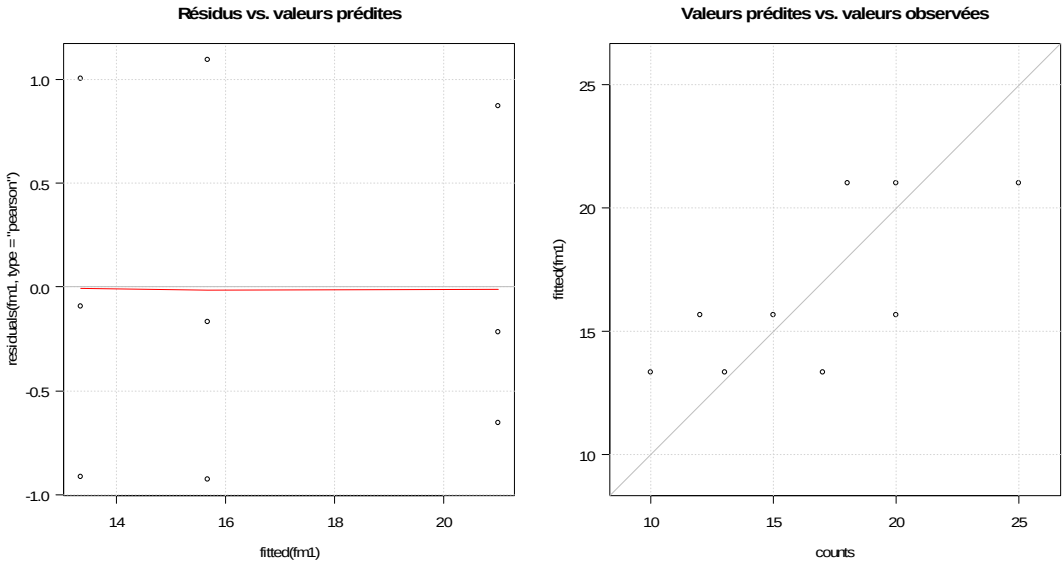
```
> counts <- c(18,17,15,20,10,20,25,13,12)
> outcome <- gl(3,1,9)
> treatment <- gl(3,3)
> Data <- data.frame(treatment, outcome, counts)
```

```
> # Modèle le plus complet que l'on est prêt à examiner
> fm1 <- glm(counts ~ outcome + treatment, family = poisson(), data = Data)
> summary(fm1)

Call:
glm(formula = counts ~ outcome + treatment, family = poisson(), data = Data)

Deviance Residuals:
       1         2         3         4         5         6         7         8
-0.67125   0.96272  -0.16965  -0.21999  -0.95552   1.04939   0.84715  -0.09167
       9
-0.96656

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  3.045e+00  1.709e-01   17.815   <2e-16 ***
outcome2    -4.543e-01  2.022e-01   -2.247   0.0246 *
outcome3    -2.930e-01  1.927e-01   -1.520   0.1285
treatment2   8.717e-16  2.000e-01 4.36e-15   1.0000
treatment3   4.557e-16  2.000e-01 2.28e-15   1.0000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10.5814  on 8  degrees of freedom
Residual deviance:  5.1291  on 4  degrees of freedom
AIC: 56.761

Number of Fisher Scoring iterations: 4
```

Ce modèle ajuste-t-il correctement les données ? On calcule tout d'abord le test de la qualité de l'ajustement basé sur les résidus de Pearson.

```
> # test global
> X2 <- sum(residuals(fm1, type = "pearson")^2)
> ddl <- df.residual(fm1)
> 1 - pchisq(X2, ddl)
[1] 0.2699831
```

On ne rejette pas l'hypothèse nulle que le modèle ajuste correctement les données. Les graphes des résidus et des valeurs prédites ne permettent pas non plus d'affirmer que le modèle ajuste mal les données.



```
> # modèles
> fm2 <- glm(counts ~ outcome, family = poisson(), data = Data)
> fm3 <- glm(counts ~ treatment, family = poisson(), data = Data)
> fm4 <- glm(counts ~ 1, family = poisson(), data = Data)

> library(metomet)
Package metomet, version 0.5-0
> sic(fm1)
    n k        LL      AIC     AICc      BIC
fm1 9 5 -23.38066 56.76132 76.76132 57.74744
> sic(fm1, fm2, fm3, fm4)
    n k        LL      AIC     AICc      BIC
fm1 9 5 -23.38066 56.76132 76.76132 57.74744
fm2 9 3 -23.38066 52.76132 57.56132 53.35299
fm3 9 3 -26.10681 58.21362 63.01362 58.80530
fm4 9 1 -26.10681 54.21362 54.78505 54.41085
```

Avec l'*AIC*, le meilleur modèle est celui impliquant la variable *treatment* alors que l'*AICc* indique le modèle de la moyenne générale (sans variable explicative).

4.2. Modèles pour données surdispersées du package aod

On utilise le jeu de données `orob2` (Fig. 2) du package `aod`, disponible sur le site web du CRAN (http://cran.r-project.org/).

```
> help(orob2)
```

orob2 {aod} R Documentation

Germination Data

Description

"A 2 x 2 factorial experiment comparing 2 types of seed and 2 root extracts. There are 5 or 6 replicates in each of the 4 treatment groups, and each replicate comprises a number of seeds varying between 4 and 81. The response variable is the proportion of seeds germinating in each replicate." (Crowder, 1978, Table 3).

Usage

```
data(orob2)
```

Format

A data frame with 21 observations on the following 4 variables.

seed
: a factor with 2 levels: `O73` and `O75`.

root
: a factor with 2 levels `BEAN` and `CUCUMBER`.

n
: a numeric vector: the number of seeds exposed to germination.

y
: a numeric vector: the number of seeds which actually germinated.

References

Crowder, M.J., 1978. *Beta-binomial anova for proportions*. Appl. Statist. 27, 34-37.

Figure 2. Description du jeu de données orob2

Nous comparons différents modèles de régression logistique, en supposant *a priori* l'existence d'une surdispersion dans les données. En conséquence, nous allons utiliser une régression logistique béta-binomiale, programmée dans la fonction `betabin()`.

```
> library(aod)
Package aod, version 1.1-8
> data(orob2)
> nrow(orob2)
[1] 21
```

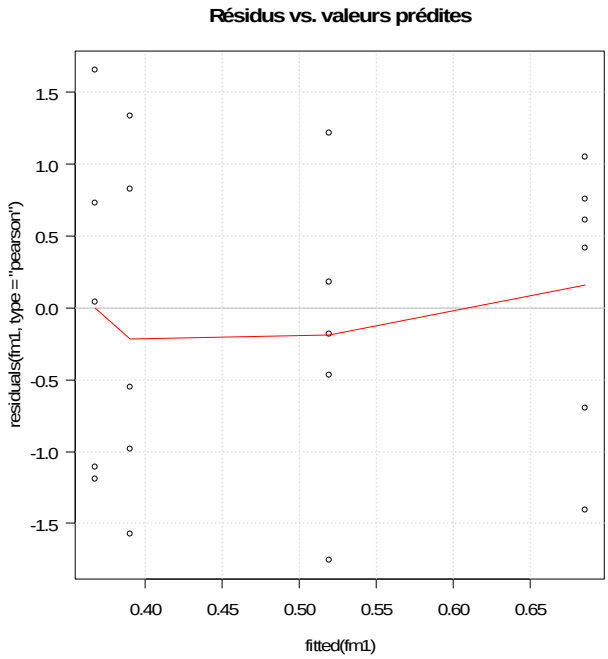
Le jeu de données comporte 21 observations. Nous utiliserons l'*AICc*. Le modèle le plus complet que nous puissions spécifier pour les effets fixes est celui comportant les variables *seed* et *root* comme effets principaux, ainsi que leur interaction :

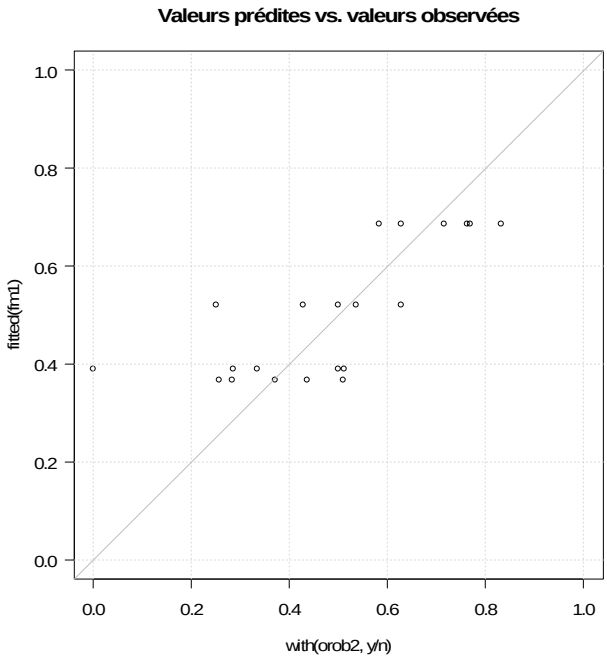
```
> fm1 <- betabin(cbind(y, n - y) ~ seed * root, ~ 1, data = orob2)
> fm1
Beta-binomial model
-------------------
betabin(formula = cbind(y, n - y) ~ seed * root, random = ~1,
    data = orob2)

Convergence was obtained after 196 iterations.
Fixed-effect coefficients:
                    Estimate Std. Error z value Pr(> |z|)
(Intercept)          -0.4456     0.2183 -2.0411    0.0412
seedO75              -0.0961     0.2737 -0.3512    0.7255
rootCUCUMBER          0.5235     0.2968  1.7636    0.0778
seedO75:rootCUCUMBER  0.7962     0.3779  2.1068    0.0351

Overdispersion coefficients:
                Estimate Std. Error z value Pr(> z)
phi.(Intercept)   0.0124     0.0113  1.0927  0.1373

Log-likelihood statistics
 Log-lik    nbpar  df res. Deviance      AIC     AICc
 -53.767        5       16   30.937  117.534  121.534
```

Ce modèle ajuste-t-il correctement les données ? On calcule tout d'abord le test de la qualité de l'ajustement basé sur les résidus de Pearson.

```
> X2 <- sum(residuals(fm1, type = "pearson")^2)
> ddl <- df.residual(fm1)
> 1 - pchisq(X2, ddl)
[1] 0.1556433
```

On ne rejette pas l'hypothèse nulle que le modèle ajuste correctement les données. Les graphes des résidus et des valeurs prédites ne permettent pas non plus d'affirmer que le modèle ajuste mal les données.





On estime les modèles qui vont être comparés à `fm1`.

```
> fm2 <- betabin(cbind(y, n - y) ~ seed + root, ~ 1, data = orob2)
> fm3 <- betabin(cbind(y, n - y) ~ seed, ~ 1, data = orob2)
> fm4 <- betabin(cbind(y, n - y) ~ root, ~ 1, data = orob2)
> fm5 <- betabin(cbind(y, n - y) ~ 1, ~ 1, data = orob2)
```

On utilise ensuite la fonction `AIC()` du package `aod`, qui permet de calculer les critères *AIC* et *AICc*.

```
> res <- AIC(fm1, fm2, fm3, fm4, fm5)
> res
    df      AIC     AICc
fm1  5 117.5336 121.5336
fm2  4 119.6637 122.1637
fm3  3 133.1054 134.5172
fm4  3 120.3920 121.8037
fm5  2 133.0326 133.6992
```

Un utilitaire `summary()` facilite la présentation et l'interprétation des données.

```
> summary(res, which = "AICc")
    df     AICc      Delta           W     Cum.W
fm1  5 121.5336  0.0000000 0.3835560300 0.3835560
fm4  3 121.8037  0.2701845 0.3350880969 0.7186441
fm2  4 122.1637  0.6301115 0.2798993204 0.9985434
fm5  2 133.6992 12.1656787 0.0008751555 0.9994186
fm3  3 134.5172 12.9836137 0.0005813972 1.0000000
```

4.3. Applications plus avancées

Dans cet exemple, 3 modèles sont très proches l'un de l'autre, et il est délicat de décider lequel d'entre eux est réellement le meilleur. Une approche possible est d'utiliser l'ensemble de ces modèles pour réaliser les inférences (Burnham et Anderson, 2002, Posada et Buckley, 2004).

A cet effet, la tendance actuelle est plutôt de se baser sur le BIC (*Bayesian information criterion*):

$$BIC = -2 * LL + k * \log(n)$$

et le package R **BMA** met cette approche en œuvre (Raftery et al., 2005).

Le BIC a été initialement proposé (Schwartz, 1978) pour sélectionner les modèles dans le cas de grands échantillons (plusieurs milliers d'observations) pour lesquels l'AIC et l'$AICc$ ont tendance à sélectionner des modèles comportant de nombreuses variables explicatives : le BIC aboutit à des modèles plus parcimonieux. Cependant, les bases théoriques sous-tendant les deux approches (AIC vs. BIC) sont différentes, l'utilisation de l'AIC étant en premier lieu dans un objectif de prédiction, et non de décision vis-à-vis de la signification statistique des paramètres retenus dans le modèle (Ripley, 2003).

5. Références

Akaike, H., 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control AU-19: 716-722.

Burnham, K. P. and Anderson, D. R., 2002. Model selection and multimodel inference: A practical information-theoretic approach., 2nd ed. Springer-Verlag, New-York.

Hurvich, C. M. and Tsai, C.-L., 1995. Model selection for extended quasi-likelihood models in small samples. Biometrics 51: 1077-1084.

Lebreton, J.-D., Burnham, K. P., Clobert, J. and Anderson, D. R., 1992. Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. Ecological Monographs 62: 67-118.

McCullagh, P. and Nelder, J. A., 1989. Generalized linear models, 2nd ed. Chapman & Hall, London.

Posada, D. and Buckley, T. R., 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53: 793-808.

Raftery, A. E., Painter, I. S. and Volinsky, C. T., 2005. BMA: An R package for Bayesian Model Averaging. R News 5: 2-9.

Ripley, B. D., 2003. Model selection in complex classes of models. http://web.maths.unsw.edu.au/~inge/statlearn/ripley1.pdf

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6: 461-464.