Biostatistiques avancées avec R

Modèle linéaire généralisé : estimation et inférence

Christophe Lalanne

www.aliquote.org

Synopsis

Rappels sur la régression logistique

Estimation des paramètres du modèle logistique

Inférence à partir du GLM

Modèles prédictifs

Extensions du GLM

Parallèle avec le modèle linéaire

Les différences principales avec la régression linéaire sont les suivantes :

- On ne parle plus de sommes de carrés (OLS, résidus, variance) mais de déviance (dans le cas gaussien, elle est équivalente à la somme de carrés de la résiduelle), mais cette dernière reflète toujours l'écart entre les données et le modèle.
- En raison de la nature binaire de la variable réponse, l'analyse classique des résidus en fonction des valeurs prédites ou la notion d'hétéroskédasticité ne font plus sens ; en revanche, on s'intéresse toujours à la qualité d'ajustement du modèle, et à la comparaison de modèles emboîtés à l'aide de tests du rapport de vraisemblance.

Aspects computationnels

Comme il n'existe pas de solution analytique, il est nécessaire d'utiliser des techniques d'optimisation non-linéaires :

- Méthode de Newton (Fisher scoring) : convergence quadratique, plus rapide que la méthode du gradient, mais nécessite la résolution d'un système linéaire à chaque étape (algorithme de Choleski) ;
- Algorithme du gradient Gradient descent : application au modèle linéaire et au GLM, et plus généralement au ML. *Variantes* : coordinate descent, stochastic gradient descent.

Interprétation des coefficients

Le terme d'intercept s'interprète comme un odds, et les coefficients de régression comme des odds-ratio : lorsque X_j augmente de d unité, l'odds de $y = 1$ augmente de $\exp(\beta_j d)$ (de manière équivalente, le log-odds augmente de $\beta_j d$).

Dans le cas où l'on a un seul prédicteur, binaire, on peut vérifier à partir de la relation $\frac{P(x)}{1-P(x)} = \exp(\beta_0 + \beta_1 x)$ que $\frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]} = \exp(\beta_1)$.

Si l'odds-ratio vaut 1, on a bien $\beta_1 = 0$, qui représente l'absence d'association entre le critère binaire et la variable explicative. Le risque relatif se définit quant à lui comme $P(1)/P(0)$.

Illustration

```
> wcgs <- read.table("data/wcgs.txt",
+                    header = TRUE, sep = "\t")
> names(wcgs)[1:14]
 [1] "id"      "age"     "height"  "weight"  "sbp"     "dbp"
 [7] "chol"    "behpat"  "ncigs"   "dibpat"  "chd69"   "typchd
[13] "time169" "arcus"

> xtabs(~ chd69 + cut(age, breaks = seq(39, 59, by = 5)),
+       data = wcgs)
     cut(age, breaks = seq(39, 59, by = 5))
chd69 (39,44] (44,49] (49,54] (54,59]
    0    1127     730     517     276
    1      55      71      66      46
```

```
> m <- glm(chd69 ~ age, data = wcgs, family = binomial)
> coef(m)
```

```
(Intercept)         age
    -5.9395      0.0744
```

$$\begin{aligned}
&\log\left[\frac{P(56)}{1-P(56)}\right] - \log\left[\frac{P(55)}{1-P(55)}\right] \\
&\quad = (-5.940 + 0.074 \times 56) - (-5.940 + 0.074 \times 55) \\
&\quad = 0.074.
\end{aligned}$$

D'où un odds-ratio de $\exp(0.074) = 1.077$ associé à une augmentation d'âge d'un an sur le risque d'infarctus, c'est-à-dire une augmentation du risque de 8 %.

Pour une variation de 10 ans, l'OR est de $\exp(0.074 \times 10) = 2.105$.

Test du rapport de vraisemblance

La déviance D représente l'écart entre les valeurs prédites par le modèle et les valeurs observées. La vraisemblance du modèle est la densité de probabilité calculée à partir des données à l'aide des paramètres du modèle estimés par maximum de vraisemblance. On peut comparer cette vraisemblance à celle correspondant à d'autres modèles possibles :

- modèle saturé : le nombre de paramètres correspond au nombre d'observations, et la déviance vaut $D = 2[\log \mathcal{L}(\text{modèle saturé}) - \log \mathcal{L}(\text{modèle actuel})] \sim \chi^2(n - p - 1)$,
- modèle emboîté : on considère une restriction de l'espace des paramètres du modèle (voire un modèle nul), $\text{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} \cdots + \beta_p x_p$, avec $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$ (test à $p - q$ degrés de liberté).

```
> m0 <- update(m, . ~ - age)
> anova(m0, m)  ## test = "Chisq"
Analysis of Deviance Table

Model 1: chd69 ~ 1
Model 2: chd69 ~ age
  Resid. Df Resid. Dev Df Deviance
1      3153       1781
2      3152       1738  1     42.9

> drop1(m, "age", test = "LRT")
Single term deletions

Model:
chd69 ~ age
       Df Deviance  AIC  LRT Pr(>Chi)
<none>        1738 1742
age     1     1781 1783 42.9  5.8e-11
```

Approche par permutation

Les tests de signification pour les coefficients de régression ($H_0 : \beta_j = 0$) reposent sur des tests de Wald, construits de manière similaire aux tests de Student ($\hat{\beta}_j / \hat{\text{se}}(\hat{\beta}_j)$). Sous H_0, la statistique de Wald suit une loi gaussienne (et son carré une loi $\chi^2(1)$).

Il est également possible d'utiliser une approche par permutation pour évaluer le degré de signification d'un prédicteur. Dans le cas d'un modèle à un seul prédicteur, l'idée revient à permuter les étiquettes 0/1 de la variable réponse et à construire la distribution des pentes de régression pour K permutations.

Dans le cas multivarié, il est préférable de travailler sur les résidus du modèle pour contrôler correctement le conditionnement sur l'ensemble des prédicteurs[7], voir le package `glmperm`[10].

Illustration

```
> set.seed(101)
> K = 100000
> r = numeric(K)
>
> for (i in 1:K) {
+   idx = sample(1:nrow(wcgs), nrow(wcgs), replace = FALSE)
+   y = wcgs$chd69[idx]
+   x = wcgs$bmi
+   m = glm(y ~ x, family = binomial)
+   r[i] = coef(m)[2]
+ }
>
> m <- glm(chd69 ~ bmi, data = wcgs, family = binomial)
> sum(r >= coef(m)[2]) / K
```

Capacité discriminante d'un modèle

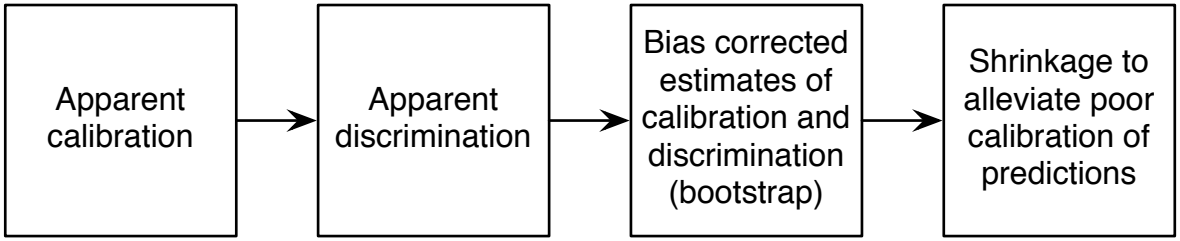
La fonction `lrm` du package `rms`[4] est beaucoup plus informative que `glm`, en particulier en ce qui concerne la qualité d'ajustement du modèle et son pouvoir discriminant.

```
> library(rms)
> ddist <- datadist(wcgs)
> options(datadist="ddist")
> mbis <- lrm(chd69 ~ bmi, data = wcgs)
```

```
> mbis

Logistic Regression Model

lrm(formula = chd69 ~ bmi, data = wcgs)
                       Model Likelihood     Discrimination    Rank Di
                          Ratio Test           Indexes           Inde
Obs           3154    LR chi2     11.82    R2        0.009    C
 0            2897    d.f.            1    g         0.236    Dxy
 1             257    Pr(> chi2) 0.0006    gr        1.266    gamma
max |deriv| 2e-11                          gp        0.018    tau-a
                                           Brier     0.075

          Coef    S.E.   Wald Z Pr(>|Z|)
Intercept -4.5087 0.6063 -7.44  <0.0001
bmi        0.0843 0.0241  3.49  0.0005

> exp(coef(mbis))
Intercept       bmi
    0.011     1.088
```

Sélection de variables, modèles prédictifs

La construction et la validation de modèles prédictifs dans le domaine clinique ont fait l'objet de nombreuses publications, en particulier Harrell[4] et Steyerberg et al.[9]. Le site RMS fournit de nombreuses ressources sur ce sujet : http://biostat.mc.vanderbilt.edu/wiki/Main/RmS.



Internal validation of predictive models

Pour un tour d'horizon rapide des enjeux de la modélisation à partir d'enquêtes épidémiologiques, voir Greenberg and Kleinbaum[3]. De nombreux autres articles fournissent des recommendations pour le reporting des résultats[1,6,2].

Régression *versus* classification

Le modèle de régression logistique est idéal pour fournir des probabilités (au niveau individuel, ou en moyenne conditionnellement à certains co-facteurs). Dans un contexte de classification, on réduit ces prédictions à deux classes, 0/1, à partir d'un cut-off optimisé sur la base d'un compromis entre sensibilité et spécificité, sans prendre en considération un coût différentiel pour les faux-positifs et les faux-négatifs. D'autre part, transformer une probabilité (continue dans [0,1]) en une variable binaire, sans autoriser de zone d'incertitude, est risqué.

> *There is a reason that the speedometer in your car doesn't just read « slow » and « fast ». Frank Harrell on R-help, February 2011*

Concernant la présentation des résultats d'une régression logistique dans les études diagnostiques, voir Steyerberg[8].

Comparaison GLM *versus* ML

Le tableau ci-dessous résume le taux de classification apparent de différents modèles de classification pour la relation suivante : `low` ∼`age + lwt + race + ftv` (données `MASS::birthwt`).

Ces taux de classification ont été estimés par validation croisée (25 × 10 fold) avec le package `caret`[(5)]. Sur les données d'origine, le taux de cas positifs est de 68,8 %.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LR	0.4737	0.6316	0.6842	0.6688	0.6842	0.7895
RF	0.3684	0.5789	0.6316	0.6401	0.6842	0.8421
SVM	0.6842	0.6842	0.6842	0.6880	0.6842	0.7222
KNN	0.4737	0.6316	0.6842	0.6605	0.7332	0.8889

Régression de Poisson

La régression de Poisson fait également partie de la famille des GLM et elle est utilisée pour modéliser des données de comptage, par exemple le nombre de cas de gastroentérites (mesurés par le nombre de consultation à l'hôpital ou en ville).

On considère trois conditions d'applications : les événements surviennent de manière indépendante sur des intervalles de temps disjoints, la probabilité d'observer plus de 2 événements sur une courte durée est faible, et la probabilité qu'un événement apparaisse sur une courte période est proportionnelle à la longueur de l'intervalle de temps.

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

(λ, paramètre d'intensité ou de fréquence, désignant la moyenne et la variance de la v.a. de Poisson.)

Loi binomiale négative

Considérons un processus binomial (succession d'événements binaires indépendants) où la probabilité de succès vaut p. On observe le processus jusqu'à obtenir r succès.

Le nombre total d'essai vaut alors :

$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}.$$

Le modèle correspondant s'écrit (notation de la fonction `glm.nb` de R) :

$$\frac{\mu}{\mu + \theta} = \exp(\beta_0 + \beta_1 x_1 + \dots).$$

(θ, paramètre de la loi binomiale négative.)

Références I

1. SC Bagley, H White, and BA Golomb. Logistic regression in the medical literature : Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54 :979–985, 2001.
2. W Bouwmeester, NPA Zuithoff, S Mallett, MI Geerlings, Y Vergouwe, EW Steyerberg, DG Altman, and KGM Moons. Reporting and methods in clinical prediction research : A systematic review. *PLoS Medicine*, 9(5) :e1001221, 2012.
3. RS Greenberg and DG Kleinbaum. Mathematical modeling strategies for the analysis of epidemiological research. *Annual Review of Public Health*, 6 :223–245, 1985.
4. FE Harrell. *Regression modeling strategies with applications to linear models, logistic regression and survival analysis*. Springer, 2001.
5. M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5), 2008.
6. KJ Ottenbacher, HR Ottenbacher, L Tooth, and GV Ostir. A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. *Journal of Clinical Epidemiology*, 57 :1147–1152, 2004.
7. DM Potter. A permutation test for inference in logistic regression with small- and moderate-sized datasets. *Statistics in Medicine*, 24 :693–708, 2005.

Références II

8. EW Steyerberg. *Clinical prediction models : a practical approach to development, validation, and updating*. Springer, 2009.
9. EW Steyerberg, FE Harrell, GJJM Borsboom, MJC Eijkemans, Y Vergouwe, and JDF Habbema. Internal validation of predictive models : Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54 :774–781, 2001.
10. W Werft and A Benner. `glmperm` : A permutation of regressor residuals test for inference in generalized linear models. *The R Journal*, 2/1 :39–43, 2010.

Index des commandes

A
anova, 9
C
coef, 7, 11, 13
D
datadist, 12
drop1, 9
E
exp, 13
G
glm, 7, 11, 12
L
library, 12
lrm, 12
N
numeric, 11
O
options, 12
R
read.table, 6
S
sample, 11
set.seed, 11
sum, 11
U
update, 9
X
xtabs, 6