Modèle linéaire Généralisé

C. Ambroise

Laboratoire de Mathématiques et Modélisation d'Évry
UMR CNRS 8071

Plan

Modèles Linéaires Généralisés

Le modèle linéaire généralisé I

Le modèle linéaire

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i,\ \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

peut être reformulé comme

$$\mu_i = \mathbf{X}_i\boldsymbol{\beta},\ \ Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

avec $\mu_i = E[Y_i]$

Le modèle linéaire généralisé II

Le modèle linéaire généralisé étend les possibles

$$g(\mu_i) = \mathbf{X}_i'\boldsymbol{\beta} \;\; Y_i \sim EF(\mu_i, \phi)$$

- g est une fonction monotone dérivable dite fonction de lien,
- $EF(\mu_i, \phi)$ une une distribution de la famille exponentielle
- ϕ est un paramètre dit parameètre d'échelle
- $\mathbf{X}'\boldsymbol{\beta} = \eta$ est un prédicteur linéaire

La fonction de lien g I

- La fonction de lien joue un peu le rôle d'un changement de variable, mais
  - la fonction de lien transforme $\mathbb{E}[Y_i]$ et non pas
  - la variable de sortie Y_i.
- Des exemples de fonction de liens (*log*, *logit*, $\sqrt{\cdot}$ ...)

Plan

Modèles Linéaires Généralisés

La famille exponentielle

$$f_\theta(y) = exp[((y\theta - b(\theta))/a(\phi)) + c(y, \phi)],$$

avec

- θ un paramètre de position,
- ϕ un paramètre d'échelle,
- $\mathbb{E}[Y] = b'(\theta) = \mu$
- $var[Y] = b''(\theta)a(\phi) = V(\mu)\phi$ avec $V(\mu) = b''(\theta)/w$

Plan

Modèles Linéaires Généralisés

Déviance I

La déviance est une quantité similaire aux résidus ($RSS/sigma^2$) dans le contexte du modèle linéaire.

$$D = 2[\mathcal{L}(\hat{\beta}_{max}) - \mathcal{L}(\hat{\beta})]\phi$$

où $\mathcal{L}(\hat{\beta}_{max})$ est la log-vraisemblance maximisé du modèle saturé.
Remarques:

- $\mathcal{L}(\hat{\beta}_{max})$ sert de référence (valeur max de la vaisemblance),
- pour calculer $\mathcal{L}(\hat{\beta}_{max})$ il suffit de remplacer $\hat{\mu}$ par $\mathbf{y}$
- dans le cas du modèle binomial $\phi = 1$

Déviance II

La déviance réduite $D^* = D/\phi$ ne dépend pas du paramètre d'échelle (dans le cas binomial et poisson $D^* = D$) et d'après la propriété asymptotique du rapport de vraisemblance:

$$D^* \sim \chi^2_{n-dim(\beta)}$$

Notons que cette approximation est très mauvaise dans le cas du modèle binomiale (car le nombre de paramètre du modèle saturé croît avec n) et exact dans le cas Gaussien.
En pratique on peut utiliser cette approximation pour estimer

$$\hat{\phi} = \frac{D}{n - dim(\beta)}$$

Déviance III

Et pour tester l'intérêt du modèle on pourra utiliser le test suivant

$$\begin{cases} H_0 : \omega \text{ est le bon modèle} \\ H_1 : \overline{H_0} \end{cases}$$

avec comme statistique de décision D^*.
Si $D^* > \chi_{n-dim(\omega);1-\alpha}$ alors on décide H_1 sinon on garde H_0

Comparaison de modèles I

Supposons que l'on veuille comparer deux modèles $\omega \subset \Omega$

$$\begin{cases} H_0 : \omega \text{ est le bon modèle} \\ H_1 : \Omega \text{ est le bon modèle} \end{cases}$$

Sous H_0, de manière approchée

$$F = \frac{(D_\omega - D_\Omega)/(q_\Omega - q_\omega)}{D_\Omega/(n - q_\Omega)} \sim F_{q_\Omega - q_\omega, n - q_\Omega}$$

Si $F > \mathcal{F}_{q_\Omega - q_\omega, n - q_\Omega; 1-\alpha}$ alors on décide H_1 sinon on garde H_0
Remarque: on peut simplement utiliser AIC pour comparer deux modèles. Dans ce cas le modèle le meilleur pour la prédiction sera préféré (un peu plus complexe que le modèle sélectionné par test d'hypothèse).

Distribution de $\hat{\beta}$ I

Les estimateurs du maximum de vraisemblance sont asymptotiquement Gaussiens

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, (\mathbf{X}^t W \mathbf{X})^{-1} \phi)$$

où W est la matrice de pondération diagonale avec $w_i = 1/(g'(\mu_i)^2 V(\mu_i))$.

Distribution de $\hat{\beta}$ II

Cette distribution est utilisée pour les intervalles de confiance et pour les test d'hypothèses pour un unique prédicteur.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Si ϕ est inconnu, on le remplace par une estimation

$$\hat{\phi} = \sum_i w_i \frac{(z_i - \mathbf{X}_i \hat{\beta})^2}{n - dim(\beta)}$$

et la distribution utilisée est une student à $n - dim(\beta)$ degrés de liberté. Avec

- les pseudo données $z_i = g'(\mu_i)(y_i - \mu_i) + \eta_i$
- $\eta_i = X_i \hat{\beta}$ et $\mu = g^{-1}(\eta_i)$

Résidus I

Les résidus d'un GLM acceptent plusieurs définitions:

Résidus de Pearson

$$\epsilon_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Résidus de déviance

$$\epsilon_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$$

où $D = \sum_i d_i$

Plan

Modèles Linéaires Généralisés

Loi binomiale et famille exponentielle I

Supposons que les observations $Y_i \sim \mathcal{B}(n_i, \pi_i)$.

$$f(y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

En mettant la distribution sous la forme canonique:

$$f(y_i) = exp(y_i \ log \frac{\pi_i}{1 - \pi_i} + n \log(1 - \pi_i) + \log \binom{n_i}{y_i})$$

donc

- $\theta = \ log \frac{\pi_i}{1 - \pi_i}$ et inversement $\pi_i = \frac{exp\theta}{1 + exp\theta}$
- $c(\theta, y_i) = \log \binom{n_i}{y_i}$
- $b(\theta) = n \log(1 - \pi_i)$
- $a(\phi) = \phi = 1$

Loi binomiale et famille exponentielle II

La fonction de lien canonique est donc

$$g(\pi_i) = \ log \frac{\pi_i}{1 - \pi_i} = \mathbf{X}_i^t \beta$$

Plan

Modèles Linéaires Généralisés

Nombre de cas de SIDA en Belgique I

$$\mu_i = \mathbb{E}[Y_i] = N_0 e^{\beta_1 X_i}, \; Y_i \sim P(\mu_i)$$

- Y_i est le nombre de nouveaux cas dans l'année x_i
- N_0 est le nombre de cas en 1980
- le nombre de cas observé suit une loi de Poisson

$$\begin{aligned} \log \mathbb{E}[Y_i] &= \log N_0 + \beta_1 x_i \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

GLM avec une fonction de lien log

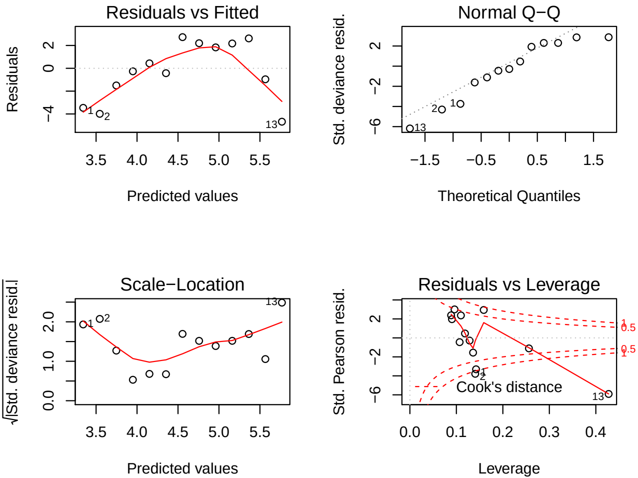
Nombre de cas de SIDA en Belgique II

```
> cases<- c(12,14,33,50,67,74,123,141,165,204,253,246,24
> year <-1:13
> plot(year+1980,cases ,xlab="Année",ylab="Nouveaux cas
```

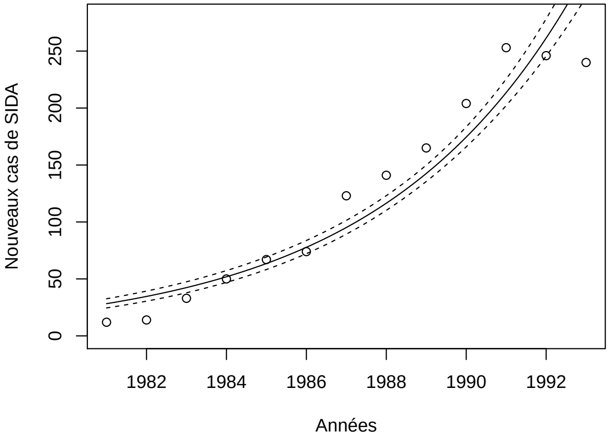
Nombre de cas de SIDA en Belgique III

```
> m0 <- glm(cases~year,poisson)
> summary(m0)

Call:
glm(formula = cases ~ year, family = poisson)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-4.678   -1.501   -0.264   2.176   2.731

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.14059    0.07825    40.1   <2e-16
year         0.20212    0.00777    26.0   <2e-16

(Intercept) ***
year        ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 872.206  on 12  degrees of freedom
Residual deviance:  80.686  on 11  degrees of freedom
AIC: 166.4

Number of Fisher Scoring iterations: 4
```

```
> par(mfrow=c(2,2))
> plot(m0)
```



Residuals vs Fitted
Residuals
Predicted values
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
Scale−Location
√|Std. deviance resid.|
Predicted values
Residuals vs Leverage
Std. Pearson resid.
Cook's distance
Leverage


Nouveaux cas de SIDA
0
50
100
150
200
250
1982
1984
1986
1988
1990
1992
Années