M1 Iséfar

Modèles de Régression Chapitre 3 : Le modèle linéaire généralisé

Cécile Durot
cecile.durot@gmail.com

Université Paris-Ouest-Nanterre-La Défense

2014-2015

Motivation

Dans le cadre de la régression (expliquer une variable réponse par une ou plusieurs variables explicatives) :

1. La loi de la variable réponse peut être continue sans être Gaussienne. Par exemple, les lois exponentielle et Gamma sont souvent utilisées pour modéliser une variable réponse positive dont la loi est dissymétrique (modèles de durées: durée de découvert sur un compte banquaire, temps séparant deux achats successifs d'un certain type de bien).
2. La loi de la variable réponse peut ne pas être continue. Par exemple,
   - données de comptage (modèles de Poisson: nombre de faillites par secteurs industriels, nombre d'accidents du travail dans une entreprise pendant une certaine période)
   - données dichotomiques (loi de Bernoulli: présence ou absence de maladie, réponse ou non-réponse, bon ou mauvais client).

Cadre du modèle linéaire généralisé

- Les réponses $Y_1, \ldots, Y_n$ sont supposées indépendantes
- Les réponses sont modélisées par une unique famille de lois (e.g. Gaussienne, Poisson, Binomiale)
- La relation entre la réponse Y_i et les variables explicatives n'est pas nécessairement linéaire mais prend la forme

$$g(m_i) = x_i\beta.$$

où $m_i = E(Y_i)$, x_i est un vecteur ligne contenant les valeurs des différentes variables (éventuellement fictives) pour le i-ème individu, et g est une fonction donnée strictement monotone sur son domaine.

Définition : g est appelée fonction de lien.
Remarque : $Y_i = f_\beta(x_i) + \varepsilon_i$ où $f_\beta(x_i) = g^{-1}(x_i\beta)$ et $E(\varepsilon_i) = 0$.

Comparaison avec le modèle linéaire Gaussien

	Modèle linéaire Gaussien	Modèle linéaire Généralisé
Réponse	continue	continue ou discrète
Loi	gaussienne	famille exponentielle
Espérance	$E(Y_i) = x_i\beta$	$E(Y_i) = g^{-1}(x_i\beta)$
Variance	$\mathrm{Var}(Y_i) = \sigma^2$	$\mathrm{Var}(Y_i) = v(x_i\beta)$

Le modèle linéaire gaussien entre donc dans le cadre du modèle linéaire généralisé (la fonction de lien est la fonction identité : $g(m) = m$ pour tout $m \in \mathbb{R}$).

Définition : Familles exponentielles

- Soit Y une variable aléatoire réelle telle qu'il existe une mesure dominante par rapport à laquelle sa densité (Radon-Nikodym) s'écrit
$$y \mapsto \exp\left(\frac{ya(m) - b(m)}{\phi} + c(y, \phi)\right)$$
où $m = E(Y)$, $\phi \in \mathbb{R}$ est un paramètre éventuellement inconnu, les fonctions a, b, c sont données, a étant strictement monotone. On dit que la loi de Y appartient à la famille exponentielle caractérisée par a, b, c, ϕ.
- On appelle ϕ le paramètre de nuisance (typiquement, 1 ou la variance de Y).
- On appelle $a(m)$ le paramètre naturel de la famille exponentielle.

Exemples

1. $\{\mathcal{E}xp(\theta), \theta > 0\}$,
2. $\{\mathcal{P}(\theta), \theta > 0\}$,
3. $\{\mathcal{B}(n, \theta), \theta \in]0, 1]\}$ pour un entier $n \geqslant 1$ fixé,
4. $\{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$ pour un $\sigma > 0$ fixé éventuellement inconnu,

sont des familles exponentielles de lois.

loi	m	ϕ	$a(m)$	$b(m)$	$c(y, \phi)$
$\mathcal{E}xp(\theta)$	$1/\theta$	1	$-1/m$	$\log(m)$	0
$\mathcal{P}(\theta)$	θ	1	$\log m$	m	$-\log(y!)$
$\mathcal{B}(n, \theta)$	$n\theta$	1	$\log\left(\frac{m}{n-m}\right)$	$n\log\left(\frac{n}{n-m}\right)$	$\log\binom{n}{y}$
$\mathcal{B}(\theta)$	θ	1	$\log\left(\frac{m}{1-m}\right)$	$-\log(1-m)$	0
$\mathcal{N}(\theta, \sigma^2)$	θ	σ^2	m	$m^2/2$	$-\frac{y^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$

Remarques

Si la loi de Y appartient à une famille exponentielle, alors :

- Les paramètres de la loi sont $m = E(Y)$ et ϕ (nuisance).
- La variance peut dépendre de m.

Exemples :

loi	m	ϕ	$\mathrm{Var}(Y)$
$\mathcal{E}xp(\theta)$	$1/\theta$	1	m^2
$\mathcal{P}(\theta)$	θ	1	m
$\mathcal{B}(n,\theta)$	$n\theta$	1	$\frac{m}{n}(n-m)$
$\mathcal{B}(\theta)$	θ	1	$m(1-m)$
$\mathcal{N}(\theta,\sigma^2)$	θ	σ^2	ϕ

Définition : modèle linéaire généralisé

- On observe des variables réelles indépendantes $Y_1, \dots Y_n$, où pour chaque i, Y_i est la réponse en $x_i = (x_{i1}, \dots, x_{ip})$. Notant $m_i = E(Y_i)$, on dit que le modèle est linéaire généralisé si
  1. les lois des Y_i appartiennent à une même famille exponentielle :
  $$f(y_i, m_i) = \exp\left(\frac{y_i a(m_i) - b(m_i)}{\phi} + c(y_i, \phi)\right)$$
  2. pour une fonction g donnée dérivable et strictement monotone, il existe un vecteur $\beta \in \mathbb{R}^p$ inconnu tel que $g(m_i) = x_i\beta$.
- On appelle alors g la fonction de lien (link en anglais).

Dans la suite, on note X la matrice de i-ème ligne x_i (matrice du plan d'expérience contenant les variables explicatives et fictives) et on suppose que X est injective (si nécessaire, on pose des contraintes d'identifiabilité).

Déclarer un modèle linéaire généralisé

Déclarer un modèle linéaire généralisé consiste donc à spécifier les trois éléments suivants :

- la matrice X (ou l'ensemble de ses colonnes),
- la famille exponentielle de lois,
- la fonction de lien.

Choix de la fonction de lien

Toute fonction dérivable strictement monotone qui envoie l'espace des valeurs de m sur $\mathbb{R}$.

- Dans le cas gaussien, $g(m) = m$ pour tout $m \in \mathbb{R}$.
- Dans le cas d'une loi de Poisson, on peut considérer par exemple $g(m) = \log(m)$ pour tout $m > 0$.
- Dans le cas d'une loi binomiale, g est une fonction quantile de $p = m/n$. Il y en a trois principales : pour tout $m \in]0, n[$,
  - lien logit: $g(m) = \log\{p/(1-p)\} = \log(m/(n-m))$
  - lien probit: $g(m) = \Phi^{-1}(p) = \Phi^{-1}(m/n)$ où Φ est la fonction de répartition de $\mathcal{N}(0,1)$
  - lien complementary log-log: $g(m) = \log\{-\log(1-p)\} = \log\{-\log(1-m/n)\}$

Lien canonique : $g(m) = a(m)$. Dans ce cas, $g(m_i) = x_i\beta$ est le paramètre naturel de la i-ème réponse.

Déclarer un modèle linéaire généralisé sur SAS ou R :

Sur SAS : plusieurs procédures

- `logistic`: réponse binaire ou ordinale, covariables quantitatives ou qualitatives, en option un algorithme de sélection de modèles
- `genmod`: modèles linéaires généralisés
- `catmod`: pour les covariables catégorielles uniquement
- `probit`: pour des covariables qualitatives
- attention, la procédure `glm` ne traite que de modèles linéaires.

Sur R, la fonction générique est la fonction `glm`.

- On déclare la variable réponse et les effets comme dans la fonction `lm`, en précisant quels sont les effets qualitatifs.
- Syntaxe : `glm(y ~ effet1+ effet2 ..., family=...)`

L'option `family=` permet de spécifier la famille de lois et la fonction de lien caractérisant le modèle.

Exemple: Régression logistique

Exemple issu de "Statistiques avec R", de Cornillon.

La traitement du cancer de la prostate change selon que le cancer a atteint ou non les noeuds lymphatiques entourant la prostate.
On étudie la variable binaire `Y` prenant la valeur 1 si le cancer a atteint le réseau lymphatique et 0 sinon. On souhaite expliquer `Y` par

- l'âge du patient (variable age),
- le niveau d'acide phosphatase sérique (variable `acide`)
- le résultat d'une analyse par rayon X (variable `rayonx` : 0 pour négatif, 1 pour positif),
- la taille de la tumeur (variable `taille` : 0 pour petite, 1 pour grande),
- état de la tumeur (variable `grade` : 0 pour moyenne, 1 pour grave),
- log du niveau d'acidité (variable `log.acid`).

Sur R

Les premières lignes du fichier cancer.txt sont :

```
age;acide;rayonx;taille;grade;Y;log.acid
66;0.48;0;0;0;0;-0.7339691750802
68;0.56;0;0;0;0;-0.579818495252942
66;0.5;0;0;0;0;-0.693147180559945
56;0.52;0;0;0;0;-0.653926467406664
58;0.5;0;0;0;0;-0.693147180559945
```

On importe les données, on déclare les variables qualitatives et on réalise une régression logistique grâce aux commandes

```
> data=read.table("cancer.txt", sep=";",header=TRUE)
> for (i in 3:6) {data [,i]<-factor (data [,i])}
> complet=glm(Y~.,data=data,family=binomial(link="logit"))
> summary(complet)
```


Régression logistique : On suppose que les Y_i sont indépendantes, et en notant $p(x) = E(Y|x) = P(Y = 1|x)$, on suppose

$$\log\left(\frac{p(x)}{1-p(x)}\right) = x\beta$$

où x est le vecteur ligne des covariables et β est un paramètre inconnu. Ceci caractérise entièrement la loi des observations.

Le nombre d'observations est $n = 53$.

On obtient la sortie R :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.08672    7.83450   1.287   0.1979
age         -0.04289    0.06166  -0.696   0.4867
acide       -8.48006    7.63305  -1.111   0.2666
rayonx1      2.06673    0.85469   2.418   0.0156 *
taille1      1.38415    0.79546   1.740   0.0819 .
grade1       0.85376    0.81247   1.051   0.2933
log.acid     9.60912    6.21652   1.546   0.1222
---
Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 44.768  on 46  degrees of freedom
AIC: 58.768
Number of Fisher Scoring iterations: 5
```


Cadre

On observe des variables réelles $Y_1, \dots Y_n$, où pour chaque i, Y_i est la réponse en $x_i = (x_{i1}, \dots, x_{ip})$, et on suppose que

1. $Y_1, \dots Y_n$ sont indépendantes,
2. pour tout i, Y_i possède une espérance notée m_i,
3. la densité de Y_i par rapport à une mesure dominante donnée s'écrit

$$f(y_i, m_i) = \exp\left(\frac{y_i a(m_i) - b(m_i)}{\phi} + c(y_i, \phi)\right)$$

où a, b, c sont des fonctions données, a étant strictement monotone, et ϕ éventuellement inconnu,

4. pour une fonction g donnée dérivable et strictement monotone, il existe un vecteur $\beta \in \mathbb{R}^p$ inconnu tel que $g(m_i) = x_i\beta$.

Score et information de Fisher

Dans le cadre du modèle linéaire généralisé précédent, supposant la matrice X (dont les lignes sont les x_i) injective, on souhaite estimer le paramètre β.

Définition. Soit $U_n(\beta) = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}$ où

$$u_j = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \log f(Y_i; g^{-1}(x_i\beta)).$$

On dit que $U_n(\beta)$ est le vecteur des scores. L'information de Fisher est définie par $\mathcal{I}_n(\beta) = E[U_n(\beta)\ {}^t U_n(\beta)]$.

Propriété. $E(U_n(\beta)) = 0$, $\mathcal{I}_n(\beta) = \sum_{i=1}^{n} Var(\frac{\partial}{\partial \beta_j} \log f(Y_i; g^{-1}(x_i\beta)))$

Estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance $\widehat{\beta}_n$ est solution des équations normales

$$U_n(\beta) = 0.$$

Sous des hypothèses de régularité adaptées, ($\beta \in \Theta$ ouvert convexe, g deux fois continûment différentiable, conditions sur le plan d'expérience pour que la matrice d'information soit définie positive), $\widehat{\beta}_n$ existe et est un estimateur consistant.

Remarques :

- Cet estimateur ne dépend pas de ϕ.
- Il se calcule par méthodes itératives (méthode de Newton-Raphson ou de Fisher's scoring).

Intervalle de confiance

Soit $h : \mathbb{R}^p \to \mathbb{R}$ continuement dérivable et de dérivée de rang plein. On utilise la delta-méthode:

$$\frac{h(\widehat{\beta}_n) - h(\beta)}{\widehat{S}} \xrightarrow{\mathcal{L}} \mathcal{N}_1(0,1)$$

où

$$\widehat{S} = \sqrt{\nabla h(\widehat{\beta}_n)\mathcal{I}_n(\widehat{\beta}_n)^{-1}\ {}^t(\nabla h(\widehat{\beta}_n))}.$$

Donc

$$[h(\widehat{\beta}_n) \pm \Phi^{-1}(1-\alpha/2)\widehat{S}],$$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$, est un intervalle de confiance pour $h(\beta)$ de coefficient de sécurité asymptotique $1-\alpha$.

Propriétés asymptotiques

Qu'est-ce que l'asymptotique ?

- Cas de données individuelles. L'asymptotique signifie $n \to \infty$.
- Cas de données groupées. Les variables explicatives définissent K groupes et pour tout $k \in \{1, \ldots, K\}$, on dispose de n_k observations dans le groupe k. Le nombre d'observations est alors $n = \sum_k n_k$. L'asymptotique signifie que $n \to \infty$ de telle sorte que chaque n_k tend vers l'infini et n_k/n tend vers une constante strictement positive (K ne dépend pas de n).

Sous des hypothèses de régularité adaptées, ($\beta \in \Theta$ ouvert convexe, g deux fois continûment différentiable, conditions sur le plan d'expérience pour que la matrice d'information soit définie positive),

$$\mathcal{I}_n(\widehat{\beta}_n)^{1/2}(\widehat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, I_p)$$

quand $n \to \infty$, où I_p désige la matrice identité dans $\mathbb{R}^p$.

Test de Wald

Pour bâtir un test asymptotique, on peut s'appuyer

- sur le résultat de convergence vers une gaussienne lorsque $p = 1$ (paramètre réel):

$$\mathcal{I}_n(\widehat{\beta}_n)^{1/2}(\widehat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

- sur la convergence vers un chi-deux quel que soit p:

$${}^t(\widehat{\beta}_n - \beta)\mathcal{I}_n(\widehat{\beta}_n)(\widehat{\beta}_n - \beta) \xrightarrow{\mathcal{L}} \chi^2(p)$$

quand $n \to \infty$. Si nécessaire, on combine ces résultats avec la delta-méthode.

Exemple : Pour tester la significativité d'une variable explicative ou fictive, on veut tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$ pour un $j \in \{1, \ldots, p\}$. Notons v_{nj} le j-ème terme diagonal de la matrice $\mathcal{I}_n(\widehat{\beta}_n)^{-1}$. On a sous des hypothèses adaptées

$$\frac{\widehat{\beta}_{nj} - \beta_j}{\sqrt{v_{nj}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

d'où l'on déduit la p-valeur

$$2\left(1 - \Phi\left(\frac{|\widehat{\beta}_{nj}|}{\sqrt{v_{nj}}}\right)\right)$$

avec Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$.

Remarque : La variance de $\widehat{\beta}_{nj}$ est estimée par v_{nj}.

Test du rapport des vraisemblances

Soit à tester

$$H_0 : h(\beta) = 0 \qquad \text{contre} \qquad H_1 : h(\beta) \neq 0$$

où $h : \mathbb{R}^p \to \mathbb{R}^q$ $(q \leqslant p)$ possède une matrice de dérivées partielles de rang q.

Sous des hypothèses de régularité adaptées, on a

$$S_{RV} = 2(\log L_n(\widehat{\beta}_n) - \log L_n(\widehat{\beta}_{n,0})) \xrightarrow{\mathcal{L}} \chi^2(q) \text{ sous } H_0$$

où L_n désigne la vraisemblance des observations et $\widehat{\beta}_{n,0}$ est l'EMV de β sous H_0.

Exemple : Si le modèle comporte un intercept, alors le nombre de degrés de liberté du chi-deux limite lorsqu'on teste la significativité globale du modèle est $p - 1$.

Modèle saturé et déviance

- On définit le modèle saturé comme le modèle ayant le plus grand nombre de paramètres possibles (tout en restant estimable, dans la famille exponentielle considérée): n pour les données individuelles et K pour les données groupées en K groupes.
- Soit L_n la vraisemblance des observations où on considère le paramètre de nuisance fixé à 1. La déviance d'un modèle $\mathcal{M}$ est définie par

$$D(\mathcal{M}) = 2(\log L_{n,S}(\widehat{\beta}_{n,S}) - \log L_n(\widehat{\beta}_n))$$

où $\widehat{\beta}_n$ est l'EMV de β dans le modèle $\mathcal{M}$, $L_{n,S}$ et $\widehat{\beta}_{n,S}$ désignent la vraisemblance et l'EMV de β dans le modèle saturé.

- Le modèle correspondant à l'hypothèse que les Y_i sont i.i.d. est appelé modèle nul. Sa déviance est appelée déviance nulle.

Sortie R :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.08672    7.83450   1.287   0.1979
age         -0.04289    0.06166  -0.696   0.4867
acide       -8.48006    7.63305  -1.111   0.2666
rayonx1      2.06673    0.85469   2.418   0.0156 *
taille1      1.38415    0.79546   1.740   0.0819 .
grade1       0.85376    0.81247   1.051   0.2933
log.acid     9.60912    6.21652   1.546   0.1222
---
Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 44.768  on 46  degrees of freedom
AIC: 58.768
Number of Fisher Scoring iterations: 5
```


Reprenons les données du fichier cancer.txt :

```
> data=read.table("cancer.txt", sep=";",header=TRUE)
> for (i in 3:6) {data [,i]<-factor (data [,i])}
> complet=glm(Y~.,data=data,family=binomial(link="logit"))
> sous.modele=glm(Y~acide+rayonx+taille+log.acid,
+ data=data,family=binomial(link="logit"))
> anova(sous.modele,complet,test="Chisq")
```

Sortie R :

```
Analysis of Deviance Table

Model 1: Y ~ acide+rayonx+taille+log.acid
Model 2: Y ~ age+acide+rayonx+taille+grade+log.acid
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        48     46.425
2        46     44.768  2   1.6568    0.4367
```


Tests d'une sous-hypothèse quelconque sur R

Si `W` et `V` sont deux objets de classe `glm` tels que `W` correspond à un sous-modèle de `V`, alors la commande

```
anova(W,V,test="Chisq")
```

réalise le test du rapport des vraisemblances de `W` contre `V`. Si l'option `test=` n'est pas spécifiée, la p-valeur du test n'est pas calculée et seule la différence des déviances entre les deux modèles est calculée.

Des tests de Wald peuvent être effectués avec la fonction `wald.test` du package `aod`.

Validation de modèle

- Sélection de variables
- Lien linéaire pour chacune des variables
- Qualité de l'ajustement
- Graphes de résidus

Cadre de la régression logistique

On suppose que pour chaque individu, la réponse est 0 (échec) ou 1 (succès) et on a un vecteur ligne $x_i = (x_{i1}, \ldots, x_{ip})$ de p variables, éventuellement fictives, associé à chaque individu. On écrit:

$$P(Y_i = 1|x_i) = p(x_i).$$

Exemples:

- Banque : Y_i vaut 1 si l'emprunteur i est bon payeur, 0 sinon. La variable x_i donne par exemple l'âge, la profession, le statut matrimonial, le fait d'être ou non propriétaire.
- Assurance : Y_i vaut 1 si l'assuré i est "bon conducteur" (pas de sinistre dans l'année), 0 sinon. La variable x_i donne par exemple l'âge, le sexe, le degré de bonus-malus de l'année précédente, la vétusté du véhicule, le code postal.
- Biologie : Y_i vaut 1 si l'insecte i est vivant une période de temps donnée après diffusion d'un insecticide en milieu clos, 0 sinon. La variable x_i donne la dose d'insecticide.

Définition de la régression logistique

On suppose que

$$E(Y|x) = F(x\beta)$$

où F est une fonction de répartition inversible, c'est-à-dire que

$$p(x) = F(x\beta),$$

ou encore que

$$F^{-1}(p(x)) = x\beta.$$

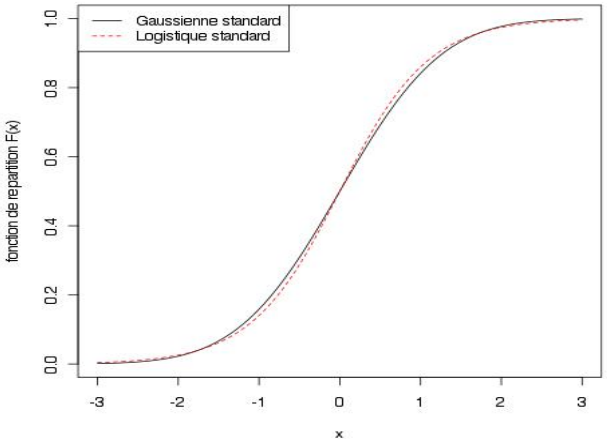
Exemples de fonctions F :

- Modèle logit ou régression logistique : F est la fonction de répartition Λ de la loi logistique.
- Modèle probit : F est la fonction de répartition Φ de la loi Gaussienne standard.

La régression logistique est donc un exemple de Modèle linéaire généralisé.

Comparaison des lois gausienne et logistique standards

de fonctions de répartition respectives $\Phi(x)$ et $\Lambda(\sigma x)$, où $\sigma = \pi/\sqrt{3}$ est l'écart-type de la loi logistique :



Interprêtation en termes de variable latente

Dans l'exemple de l'insecticide, supposons que l'insecte i soit caractérisé par une dose $Y_i^\star$ d'insecticide, reflétant la capacité de résistance de l'insecte, telle que l'insecte est vivant après la période de référence si et seulement si $x_i \leqslant Y_i^\star$. On a donc

$$Y_i = \mathbb{1}_{Y_i^\star \geqslant x_i} = \mathbb{1}_{Z_i^\star \geqslant 0}$$

où $Z_i^\star = Y_i^\star - x_i$ est appelée variable latente (non observée). Supposons que $Z_i^\star$ obéisse à un modèle linéaire

$$Z_i^\star = x_i\beta + \varepsilon_i$$

où les ε_i sont i.i.d. En notant F la fonction de répartition commune des $-\varepsilon_i$, on a

$$E(Y_i|x_i) = F(x_i\beta).$$

Noter qu'il n'est pas restrictif de supposer var(ε_i) connue (sans contrainte, cette variance n'est pas identifiable).

Cotes (odds) et rapports de cotes (odds ratios)

Cote : $C(x) = \dfrac{p(x)}{1-p(x)}$.

Rapport de cotes : $R(x', x) = \dfrac{p(x')(1-p(x))}{(1-p(x'))p(x)} = \dfrac{C(x')}{C(x)}$.

Dans le cas de la régression logistique, on a

$$\exp(x\beta) = C(x),$$

et si toutes les coordonnées de x et x' sont identique hormis la $j$ème, pour laquelle $x'_{[j]} = x_{[j]} + 1$, alors on a

$$\exp(\beta_j) = R(x', x).$$

Cas de données groupées

Supposons que l'on ait K groupes, i.e. seulement K valeurs possibles pour le vecteur de variables explicatives x, et que pour chaque groupe k, $k = 1, \ldots, K$, on dispose de n_k observations Y_{kj}, $j = 1, \ldots, n_k$ i.i.d. à valeurs dans $\{0, 1\}$. Ainsi,

$$P(Y_{kj} = 1|x_k) = p(x_k),$$

et on suppose $p(x) = F(x\beta)$ où F est une fonction de répartition inversible et $\beta \in \mathbb{R}^p$ est inconnu. On peut résumer les données par

$$(x_k, Y_{k.}, n_k), \ k = 1, \ldots, K$$

où on a posé

$$Y_{k.} = \sum_{j=1}^{n_k} Y_{kj} \sim \mathcal{B}(n_k, p(x_k)).$$

Cas de données groupées

- Le modèle saturé comporte K paramètres :

$$p(x_k),\ k = 1, \ldots, K.$$

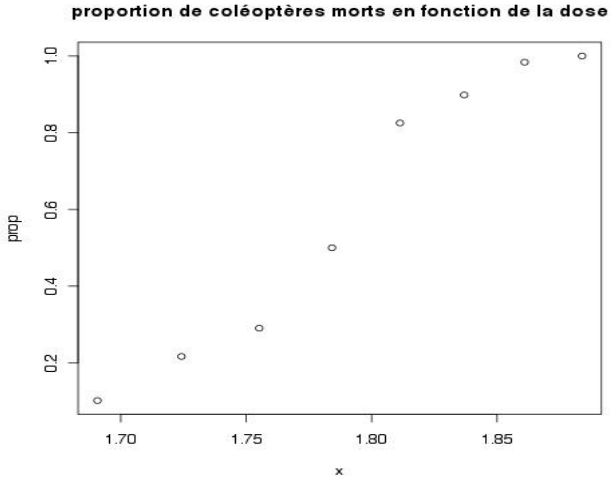
  Le nombre de paramètres du modèle saturé ne dépend donc pas du nombre d'observations.
- Si la matrice X dont les lignes sont les x_k est de rang p, alors la déviance tend en loi vers un $\chi^2(K-p)$ quand n tend vers l'infini de sorte que les n_k/n tendent vers une constante strictement positive (K fixé). On dispose donc d'un test d'adéquation du modèle.
- Pour ramener des données individuelles (Y_i est à valeurs dans $\{0,1\}$) au cas de données groupées, on peut séparer les données en K groupes (segmentation selon les variables explicatives).

Exemple sur R (Dobson p.127)

Le tableau suivant donne le nombre de coléoptères morts après cinq heures d'exposition au disulfure de carbone gazeux à différentes concentrations : x_i est le logarithme de la dose de disulfure de carbone (en mg/l), n_i est le nombre de coléoptères soumis à la dose x_i et y_i est le nombre de coléoptères soumis à la dose x_i et morts après cinq heures.

x_i	n_i	y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Représentation des données



Remarque : Représentation peu informative en cas de données non groupées.

Code R pour une régression logistique

```
> x=c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,
+ 1.8610,1.8839)
> y=c(6,13,18,28,52,53,61,60)
> n=c(59,60,62,56,63,59,62,60)
> n0=n-y
> mat=cbind(y,n0)
> prop=y/n
> plot(prop~x, main="proportion de coleopteres morts
+ en fonction de la dose")
> res.logit=glm(mat~x,family=binomial(link="logit"))
> res.probit=glm(mat~x,family=binomial(link="probit"))
> res.loglog=glm(mat~x,family=binomial(link="cloglog"))
```

Extrait de sorties R pour une régression logistique

```
> summary(res.logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
x             34.270      2.912   11.77   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43
```

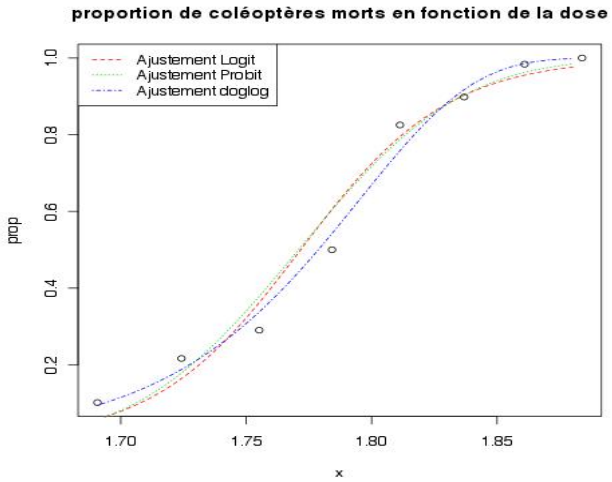

Calcul des proportions ajustées

```
> xgrid=seq(min(x),max(x),by=0.005)
> res.logit["coefficients"]

$coefficients
(Intercept)           x
  -60.71745    34.27033

> ajust.logit=plogis(-60.71745+34.27033*xgrid)
> plot(prop~x, main="proportion de coleopteres morts
+ en fonction de la dose")
> lines(ajust.logit~xgrid,lty=2,col=2)
```


Représentation des données



Comparaison des trois modèles

	Logit	Probit	cloglog
Residual deviance	11.23	10.12	3.45
1-F(Residual deviance)	0.081	0.120	0.751
AIC	41.43	40.318	33.644
Intercept (Std)	-60.72 (5.18)	-34.94 (2.65)	-39.57 (3.24)
x (Std)	34.27 (2.91)	19.73 (1.49)	22.04 (1.80)

On a noté F la fonction de répartition de la loi de Chi-deux à 6 degrés de liberté.