Chapitre 6

Modélisation de données binomiales

On modélise les données binomiales ou bernoulli avec la régression logistique. Ce type de régression est très utile lorsque les données sont dichotomiques, c'est-à-dire lorsqu'il n'y a que deux possibilités :

1. Survie ou décès d'un individu
2. Un traitement est efficace ou ne l'est pas
3. Vote pour Obama versus Romney
4. Réussite ou échec d'un examen
5. etc.

Soit $Y_i^* \sim Binomiale(m_i, \pi_i)$, $m_i > 0$ est un entier et est connu. On transforme Y_i^* en divisant par m_i. Alors, la densité de $Y_i = Y_i^*/m_i$

$$\begin{aligned} f_{Y_i}(y_i; \pi_i) &= \binom{m_i}{m_i y_i} \pi^{m_i y_i} (1-\pi)^{m_i(1-y_i)} \\ &= \exp\left\{ m_i y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \ln(1-\pi_i) + \ln\binom{m_i}{m_i y_i} \right\}. \end{aligned}$$

avec $y_i \in \{0, 1/m_i, 2/m_i, ..., 1\}$. Le paramètre canonique est $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, $\phi = 1$ et $w_i = m_i$ sont des poids.

Le lien canonique est donc le lien logistique, mais on peut également utiliser les liens probit ou log-log complémentaire.

Exemple 6.1. *Cet exemple provient de Bliss (1935). Des coccinelles ont été exposées à différentes concentrations de disulfure de carbone gazeux pendant cinq heures. On note ensuite le nombre de coléoptères qui n'ont pas survécu à cette période. Les données sont présentées ci-dessous :*

Dose	*# Exposées*	*# Mortes*	*Proportion*
49.1	*59*	*6*	*0.102*
53.0	*60*	*13*	*0.217*
56.9	*62*	*18*	*0.290*
60.8	*56*	*28*	*0.500*
64.8	*63*	*52*	*0.825*
68.7	*59*	*53*	*0.898*
72.6	*62*	*61*	*0.984*
76.5	*60*	*60*	*1.000*

On observe une tendance claire dans la proportion de décès lorsque la dose augmente. On modélise la probabilité de décès avec un GLM Binomial avec lien logistique et une variable explicative continue :

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

Code et résultats

```
> beetle <- read.table("Beetledata.txt",header=TRUE,sep=" ")
> fit <- glm(cbind(Killed,Exposed-Killed)~Dose,family=binomial(logit),data=beetle)
> summary(fit)

Call:
glm(formula = cbind(Killed, Exposed - Killed) ~ Dose, family = binomial(logit),
    data = beetle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2746  -0.4668   0.7688   0.9544   1.2990

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.82300    1.28959  -11.49   <2e-16 ***
Dose          0.24942    0.02139   11.66   <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: 284.2024  on 7  degrees of freedom
Residual deviance:   7.3849  on 6  degrees of freedom
AIC: 37.583

Number of Fisher Scoring iterations: 4
```

Age	Sex	Class	Number of survivors	Total number
Child	M	1	5	5
		2	11	11
		3	13	48
		Crew	0	0
	F	1	1	1
		2	13	13
		3	14	31
		Crew	0	0
Adult	M	1	57	175
		2	14	168
		3	75	462
		Crew	192	862
	F	1	140	144
		2	80	93
		3	76	165
		Crew	20	23

TABLE 6.1: Données pour exemple 6.2

Le paramètre β_1 représente l'effet de l'augmentation d'une unité de la dose sur le 'log odds ratio'. Puisque ce paramètre est positif, la probabilité de décès augmente lorsque la dose augmente. □

Exemple 6.2. *Les données dans le tableau 6.1 proviennent du package* `datasets` *en* `R`*. On y trouve le nombre de passagers à bord du Titanic et le nombre qui ont survécu, selon le sexe (1-Homme, 2-Femme), la classe (1,2,3, ou dans l'équipage) et l'âge (1-Enfant, 2-Adulte). On s'intéresse au taux de survie et à l'effet des variables explicatives sur celui-ci. Est-ce que la politique 'les femmes et les enfants d'abord' a été efficace ?*

On utilise un GLM Binomial avec lien logistique pour modéliser π_{jkl}. On considère d'abord le modèle saturé `Age*Sex*Class`*. Après analyse de la déviance on constate que le modèle le plus approprié semble être* `Age*Class+Sex*Class` *:*

$$\ln\left(\frac{\pi_{jkl}}{1-\pi_{jkl}}\right) = \alpha + \beta_j^{age} + \beta_k^{sex} + \beta_l^{class} + \gamma_{j,l}^{age,class} + \gamma_{k,l}^{sex,class},$$

avec $j = \{1,2\}$, $k = \{1,2\}$, $l = \{1,..,4\}$ et les contraintes d'identification

$$\beta_1^{age} = \beta_1^{sex} = \beta_1^{class} = 0 \text{ et } \gamma_{1,l}^{age,class} = \gamma_{j,1}^{age,class} = \gamma_{1,l}^{sex,class} = \gamma_{k,1}^{sex,class} = 0.$$

Code et résultats

```
> fit.satur <- glm(cbind(Survived,Total-Survived)~Age*Sex*Class,family=binomial(logit),data=titanic)
> anova(fit.satur)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Survived, Total - Survived)

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev
NULL                            13     671.96
Age           1    19.56        12     652.40
Sex           1   420.80        11     231.60
Class         3   119.03         8     112.57
Age:Sex       1    18.02         7      94.55
Age:Class     2    29.53         5      65.01
Sex:Class     3    65.01         2       0.00
Age:Sex:Class 2     0.00         0       0.00
> fit2 <- update(fit.satur,~.-Age:Sex:Class)
> drop1(fit2,test="Chisq")
Single term deletions

Model:
cbind(Survived, Total - Survived) ~ Age + Sex + Class + Age:Sex +
    Age:Class + Sex:Class
          Df Deviance     AIC    LRT  Pr(>Chi)
<none>         0.000  70.621
Age:Sex    1    1.685  70.306  1.685    0.1942
Age:Class  2   37.263 103.884 37.263 8.101e-09 ***
Sex:Class  3   65.013 129.634 65.013 4.984e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> fit3 <- update(fit2,~.-Age:Sex)
> drop1(fit3,test="Chisq")
Single term deletions

Model:
cbind(Survived, Total - Survived) ~ Age + Sex + Class + Age:Class +
    Sex:Class
          Df Deviance     AIC    LRT  Pr(>Chi)
<none>         1.685  70.306
Age:Class  2   45.899 110.520 44.214 2.507e-10 ***
Sex:Class  3   76.904 139.525 75.219 3.253e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(fit3)

Call:
glm(formula = cbind(Survived, Total - Survived) ~ Age + Sex +
    Class + Age:Class + Sex:Class, family = binomial(logit),
    data = titanic)

Deviance Residuals:
      17        18        19        20        21        22        23        24
 0.00005   0.00007   0.82651   0.00000   0.00000   0.00001  -0.87452   0.00000
      25        26        27        28        29        30        31        32
 0.00000   0.00000  -0.30431   0.00000   0.00000   0.00000   0.38065   0.00000

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.276e-01  1.613e-01  -4.511 6.45e-06 ***
AgeChild           2.276e+01  1.650e+04   0.001  0.99890
SexF               4.283e+00  5.321e-01   8.049 8.36e-16 ***
Class2            -1.670e+00  3.224e-01  -5.181 2.21e-07 ***
Class3            -8.751e-01  2.018e-01  -4.337 1.44e-05 ***
ClassCrew         -5.222e-01  1.809e-01  -2.887  0.00389 **
AgeChild:Class2    1.998e+00  2.101e+04   0.000  0.99992
AgeChild:Class3   -2.242e+01  1.650e+04  -0.001  0.99892
AgeChild:ClassCrew         NA         NA      NA       NA
SexF:Class2       -6.801e-02  6.712e-01  -0.101  0.91929
SexF:Class3       -2.898e+00  5.636e-01  -5.141 2.73e-07 ***
SexF:ClassCrew    -1.136e+00  8.205e-01  -1.385  0.16616
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 671.9622  on 13  degrees of freedom
Residual deviance:   1.6854  on  3  degrees of freedom
AIC: 70.306

Number of Fisher Scoring iterations: 21

> fit4 <- update(fit3,~.-Age:Class)
> drop1(fit4,test="Chisq")
Single term deletions

Model:
```

```
cbind(Survived, Total - Survived) ~ Age + Sex + Class + Sex:Class
          Df Deviance    AIC    LRT  Pr(>Chi)
<none>         45.899 110.52
Age        1   66.238 128.86 20.339 6.486e-06 ***
Sex:Class  3  112.567 171.19 66.667 2.206e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(fit4)

Call:
glm(formula = cbind(Survived, Total - Survived) ~ Age + Sex +
    Class + Sex:Class, family = binomial(logit), data = titanic)

Deviance Residuals:
     17       18       19       20       21       22       23       24
 2.2841   5.1564  -1.0917   0.0000   0.1406   1.1625  -2.5433   0.0000
     25       26       27       28       29       30       31       32
-0.3259  -1.8814   0.4503   0.0000  -0.0050  -0.1998   1.0832   0.0000

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.6752     0.1576  -4.284 1.84e-05 ***
AgeChild         1.0537     0.2304   4.573 4.81e-06 ***
SexF             4.2331     0.5310   7.972 1.56e-15 ***
Class2          -1.2323     0.2688  -4.584 4.56e-06 ***
Class3          -1.0229     0.1991  -5.139 2.77e-07 ***
ClassCrew       -0.5746     0.1776  -3.235  0.00122 **
SexF:Class2     -0.4483     0.6460  -0.694  0.48772
SexF:Class3     -2.8625     0.5633  -5.082 3.73e-07 ***
SexF:ClassCrew  -1.0862     0.8197  -1.325  0.18516
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 671.962  on 13  degrees of freedom
Residual deviance:  45.899  on  5  degrees of freedom
AIC: 110.52

Number of Fisher Scoring iterations: 5
```

Après analyse de la déviance, on constate que le modèle le plus approprié semble être `Age*Class+Sex*Class` *:*

$$\ln\left(\frac{\pi_{jkl}}{1-\pi_{jkl}}\right) = \alpha + \beta_j^{age} + \beta_k^{sex} + \beta_l^{class} + \gamma_{j,l}^{age,class} + \gamma_{k,l}^{sex,class},$$

avec $j = \{1, 2\}$, $k = \{1, 2\}$, $l = \{1, .., 4\}$ *et les contraintes d'identification*

$$\beta_1^{age} = \beta_1^{sex} = \beta_1^{class} = 0 \text{ et } \gamma_{1,l}^{age,class} = \gamma_{j,1}^{age,class} = \gamma_{1,l}^{sex,class} = \gamma_{k,1}^{sex,class} = 0.$$

Autrement dit,

$$\ln\left(\frac{\hat{\pi}_{jkl}}{1-\hat{\pi}_{jkl}}\right) = \begin{cases} 22.03, & \textit{si Child, M, Class 1} \\ 22.36, & \textit{si Child, M, Class 2} \\ -1.26, & \textit{si Child, M, Class 3} \\ 26.32, & \textit{si Child, F, Class 1} \\ 26.58, & \textit{si Child, F, Class 2} \\ 0.12, & \textit{si Child, F, Class 3} \\ 3.55, & \textit{si Adult, F, Class 1} \\ 1.81, & \textit{si Adult, F, Class 2} \\ -0.22, & \textit{si Adult, F, Class 3} \\ 1.90, & \textit{si Adult, F, Crew} \\ -0.72, & \textit{si Adult, M, Class 1} \\ -2.40, & \textit{si Adult, M, Class 2} \\ -1.60, & \textit{si Adult, M, Class 3} \\ -1.25, & \textit{si Adult, M, Crew} \end{cases}$$

On peut donc calculer les probabilités estimées $\hat{\pi}_{jkl}$ *avec la fonction* `fitted`. *On observe que les femmes et les enfants en troisième classe ont eu moins de chance !*

```
> Age <- factor(rep(c("Child","Adult"),each=8),c("Adult","Child"))
> Sex <- factor(rep(rep(c("M","F"),each=4),2),c("M","F"))
> Class <- factor(rep(c(1,2,3,"Crew"),4),c(1,2,3,"Crew"))
> dat <- data.frame(Age, Sex, Class, count=rep(1,16))
> predictions <- predict(fit3,data=dat,type="response")
> data.frame(dat,predictions)
     Age Sex Class count predictions
17 Child   M     1     1  1.00000000
18 Child   M     2     1  1.00000000
19 Child   M     3     1  0.22014973
20 Child   M  Crew     1  1.00000000
21 Child   F     1     1  1.00000000
22 Child   F     2     1  1.00000000
23 Child   F     3     1  0.53009073
24 Child   F  Crew     1  1.00000000
25 Adult   M     1     1  0.32571429
26 Adult   M     2     1  0.08333333
27 Adult   M     3     1  0.16760349
28 Adult   M  Crew     1  0.22273782
29 Adult   F     1     1  0.97222222
```

```
30 Adult   F     2     1  0.86021505
31 Adult   F     3     1  0.44586174
32 Adult   F  Crew     1  0.86956522

## On peut aussi les calculer "à la main"
## Prob survie Homme Adulte Classe 1
> exp(coef(fit3)[1])/(1+exp(coef(fit3)[1]))
(Intercept)
  0.3257143
```

□

6.1 Cas Bernoulli

Soit $Y_i \sim Bernoulli(\pi_i)$. Alors, on a

$$f_{Y_i}(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

avec $y_i \in \{0,1\}$. La fonction de log-vraisemblance est

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right].$$

Dans le modèle complet (ou saturé) on a que $\hat{\pi}_i = y_i$, alors l'expression de la déviance du modèle est :

$$D(y;\hat{\pi}) = 2\sum_{i=1}^{n} \left[y_i \ln(y_i) + (1-y_i)\ln(1-y_i) - y_i\ln(\hat{\pi}_i) - (1-y_i)\ln(1-\hat{\pi}_i)\right].$$

Or, si $y_i = 0$, alors $y_i \ln(y_i) = (1-y_i)\ln(1-y_i) = 0$ et si $y_i = 1$, $y_i \ln(y_i) = (1-y_i)\ln(1-y_i) = 0$. On trouve donc une forme plus simple pour la déviance :

$$\begin{aligned} D(y;\hat{\pi}) &= 2\sum_{i=1}^{n} \left[-y_i\ln(\hat{\pi}_i) - (1-y_i)\ln(1-\hat{\pi}_i)\right] \\ &= -2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i)}\right) + \ln(1-\hat{\pi}_i)\right]. \end{aligned}$$

Sous le lien canonique, on a que $\eta_i = \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i)}\right)$, alors

$$D(y;\hat{\pi}) = -2\sum_{i=1}^{n} \left[y_i \mathbf{x}_i\hat{\beta} + \ln(1-\hat{\pi}_i)\right].$$

On note que

$$\frac{\partial}{\partial\beta}\ell(\beta) = \sum_{i=1}^{n}(y_i - \pi_i)x_i = 0 \Rightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n}\pi_i x_i.$$

Donc,

$$D(y;\hat{\pi}) = -2\sum_{i=1}^{n}\left[\hat{\pi}_i x_i\hat{\beta} + \ln(1-\hat{\pi}_i)\right].$$

On observe que $D(y;\hat{\pi})$ est une fonction de $\hat{\beta}$ et ne dépend pas de y_i. Quand $n \to \infty$, $\hat{\beta} \to \boldsymbol{\beta}$, donc même de façon asymptotique, il n'est pas vrai que $D(y;\hat{\pi}) \sim \chi^2_{(n-p)}$ puisque la déviance est une fonction de la vraie valeur du paramètre. Alors, pour le modèle binomial avec m petit, soit $m = 1, 2, 3$, on ne peut pas utiliser la statistique de déviance pour déterminer si le modèle est bien ajusté.

D'ailleurs, on ne peut pas non plus utiliser le X^2 de Pearson. Selon McCullagh et Nelder (1989), si $Y \sim Bernoulli(\pi)$,

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{\bar{y}(1-\bar{y})} = n,$$

ce qui est fixé !

6.1.1 Petite étude de simulation

On illustre que la déviance ne suit pas une chi-carrée $n-p'$ lorsque les données sont Bernoulli. Méthode :

1. On simule $n = 200$ données Bernoulli avec paramètre $\pi = 0.1$.
2. On ajuste un GLM binomial avec lien logistique et aucune variable explicative.
3. On calcule la déviance du modèle.
4. On répète les étapes 1, 2 et 3 $N = 200$ fois, et on observe la distribution des déviances calculées.

On illustre que l'analyse de la déviance pour comparer les modèles fonctionne quand même avec les données Bernoulli. Méthode :

1. On simule $n = 200$ données Bernoulli avec paramètre $\pi = 0.1$.
2. On simule $n = 200$ variables explicatives x uniformes.
3. On ajuste un GLM binomial avec lien logistique et aucune variable explicative.
4. On ajuste un GLM binomial avec lien logistique et la variable explicative x.
5. On calcule la déviance des modèles et la différence entre les déviance.
6. On répète les étapes 1, 2 et 3 $N = 200$ fois, et on observe la distribution des différence entre les déviances calculées.

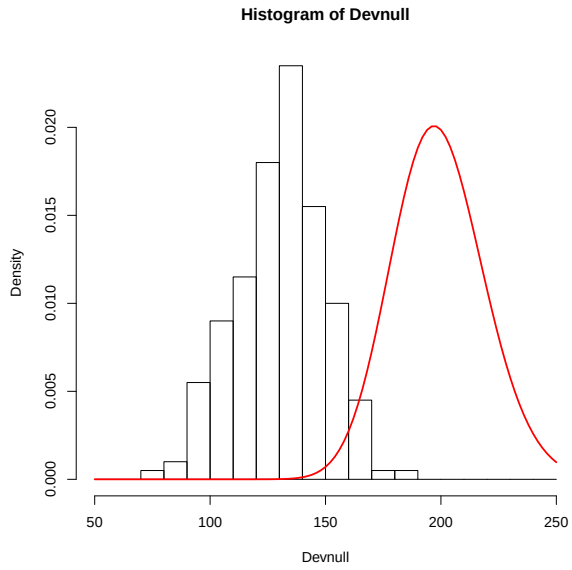

FIGURE 6.1: Histogramme $D(y_i, \hat{\mu})$ et courbe de la distribution $\chi^2_{(199)}$

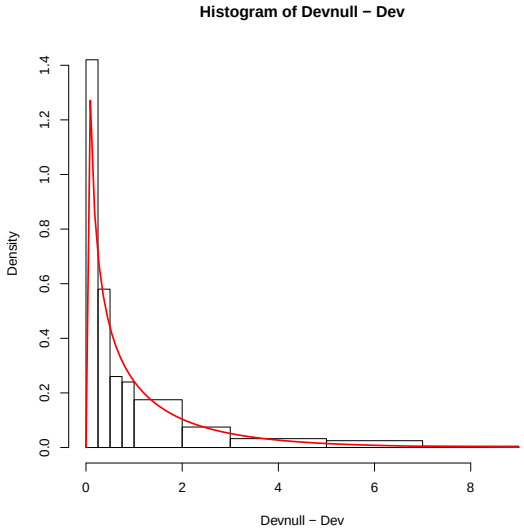

FIGURE 6.2: Histogramme $D(y_i, \hat{\mu}_i) - D(y_i, \hat{\mu})$ et courbe de la distribution $\chi^2_{(1)}$

Il faut être prudent avec la déviance pour un modèle Binomial lorsque m est petit (voir Figure 6.3). Lorsque m est élevé, l'approximation chi-carrée est valide, tel qu'illustré avec $m = 15, \pi = 0.2$ et $m = 100, \pi = 0.6$ dans la Figure 6.4.

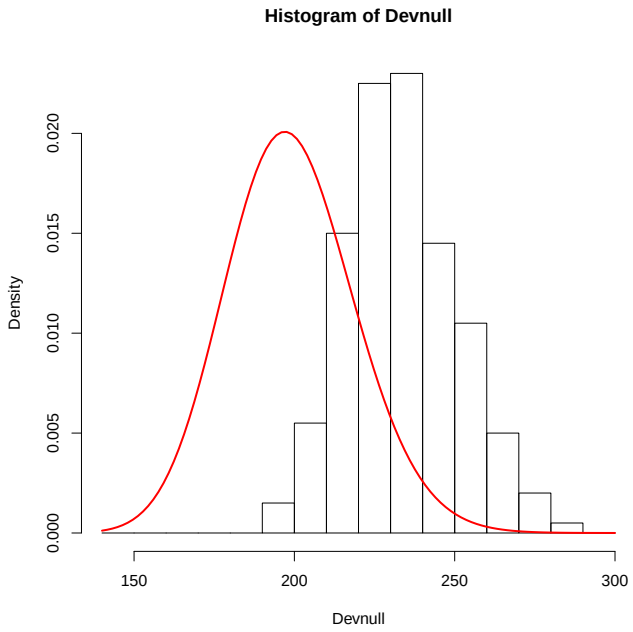

FIGURE 6.3: Histogramme de la déviance avec $m = 3$ et $\pi = 0.2$

6.1.2 Statistique de mauvaise classification

La courbe ROC est une statistique de 'mauvaise classification' (misclassification statistics) qui permet de vérifier l'ajustement du modèle Bernoulli. On forme le tableau

	Prédiction $\hat{Y}_i$	
Vrai Y_i	0	1
0	a	b
1	c	d

où, pour un seuil $\tau \in (0, 1)$,

$$\hat{Y}_i = \begin{cases} 0 & , \hat{\pi}_i < \tau \\ 1 & , \hat{\pi}_i \geq \tau \end{cases} .$$

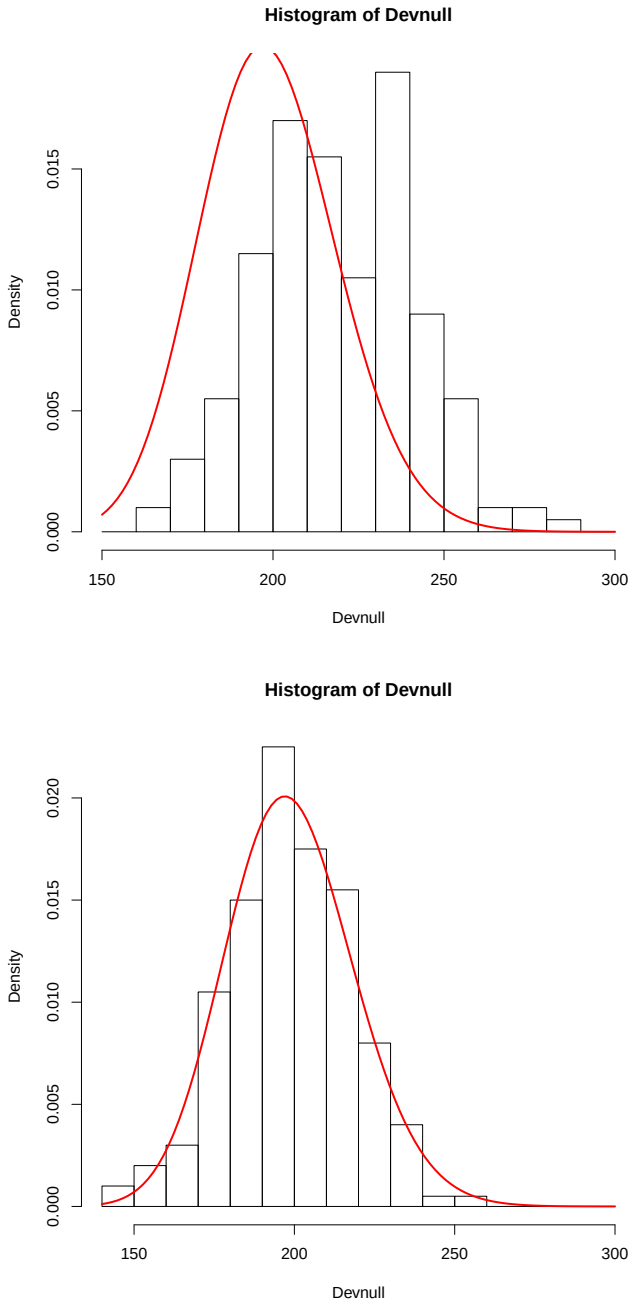

FIGURE 6.4: Histogrammes de la déviance avec $m = 15$ (haut) et $m = 100$ (bas)

Ce tableau indique si le modèle prédit bien les données observées. La courbe ROC est un tracé de la statistique de sensitivité

$$\alpha(\tau) = \frac{d}{c+d}$$

en fonction de $1-\beta(\tau)$, où $\beta(\tau)$ est la statistique de spécificité

$$\beta(\tau) = \frac{a}{a+b}.$$

Un bon modèle va bien classifier les 0 et les 1 lorsque le seuil τ est optimal. Dans ce cas, on verra la courbe ROC s'approcher du coin supérieur gauche $(0,1)$ du graphique. La fonction `lroc` du package `epicalc` en `R` permet de tracer cette courbe. Il est cependant plus simple de la coder directement.

Exemple 6.3. ***Accouplement de la limule***[1]

On utilise des données provenant d'une étude sur les limules femelles sur une île dans le Golfe du Mexique. Pendant la saison de la reproduction, les femelles migrent vers le rivage pour se reproduire, avec un mâle accroché à l'arrière de sa colonne vertébrale. Elle creuse dans le sable et y pond des oeufs. Pendant la même période, d'autres males peuvent se tenir autour du couple et peuvent aussi fertiliser les oeufs. Ces autres mâles sont appelés "satellites".

On s'intéresse à la variable binaire qui indique si une femelle a au moins un satellite. On a $n = 173$ observations, et les variables sont

– couleur du crabe femelle :

$$C = \begin{cases} 2 & \text{pâle} \\ 3 & \text{moyenne} \\ 4 & \text{moyenne foncée} \\ 5 & \text{foncée} \end{cases}$$

– largeur de la carapace (cm)

On modélise la variable

$$y_i = \begin{cases} 0 & \text{aucun satellite} \\ 1 & \text{au moins un satellite} \end{cases}$$

avec un GLM binomial et une variable explicative continue, x (`width`). On essaie trois fonction de liens différentes :

1. Exemple tiré de Agresti (2013), chapitre 4

1. Le lien canonique : $logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$.
2. Le lien probit : $\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$.
3. Le lien log-log complémentaire : $\ln(-\ln(1-\pi_i)) = \beta_0 + \beta_1 x_i$.

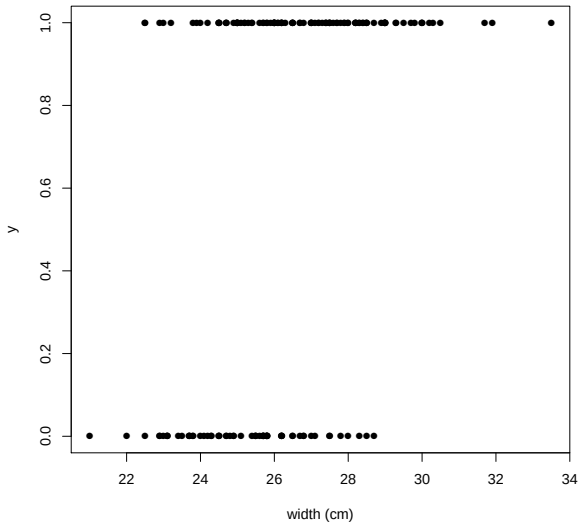

FIGURE 6.5: Graphique en nuage de points de y en fonction de x

Pour tracer un graphique plus facile à interpréter que la Figure 6.5, on peut grouper les données et calculer la moyenne de y sur l'intervalle. On a

Largeur	x moyen	y moyen
< 23.5	22.69286	0.3571429
$23.5 - 24.5$	23.84286	0.2857143
$24.5 - 25.5$	24.77500	0.6071429
$25.5 - 26.5$	25.83846	0.5384615
$26.5 - 27.5$	26.79091	0.6818182
$27.5 - 28.5$	27.73750	0.8333333
$28.5 - 29.5$	28.66667	0.8333333
> 29.5	30.40714	1.0000000

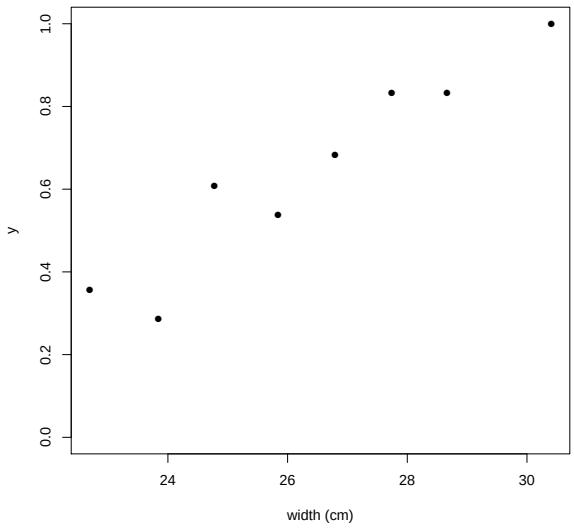

FIGURE 6.6: Graphique en nuage de points de y en fonction de x - données groupées

Modèle logistique

```
> mod1 <- glm(y~width,family=binomial,data=crab)
> summary(mod1)
Call:
glm(formula = y ~ width, family = binomial, data = crab)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.0458   0.5480   0.9066   1.6942
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45
Number of Fisher Scoring iterations: 4
```

Modèle probit

```
> mod2 <- glm(y~width,family=binomial(link=probit),data=crab)
> summary(mod2)
Call:
glm(formula = y ~ width, family = binomial(link = probit), data = crab)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0519  -1.0494   0.5374   0.9126   1.6897
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.50196    1.50712  -4.978 6.44e-07 ***
width        0.30202    0.05804   5.204 1.95e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.04  on 171  degrees of freedom
AIC: 198.04
Number of Fisher Scoring iterations: 5
```

Modèle log-log complémentaire

```
> mod3 <- glm(y~width,family=binomial(link=cloglog),data=crab)
> summary(mod3)
Call:
glm(formula = y ~ width, family = binomial(link = cloglog), data = crab)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1188  -1.0477   0.5089   0.9419   1.6056
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.17442    1.58575  -5.155 2.54e-07 ***
width        0.31299    0.05978   5.236 1.64e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 193.28  on 171  degrees of freedom
AIC: 197.28
Number of Fisher Scoring iterations: 5
```

Comparaison

```
> ilogit <- function(x,coef) exp(coef[1]+coef[2]*x)/(1+exp(coef[1]+coef[2]*x))
> iprobit <- function(x,coef) pnorm(coef[1]+coef[2]*x)
```

```
> icloglog <- function(x,coef) 1-exp(-exp(coef[1]+coef[2]*x))
> plot(grwidth,gry,pch=16,ylim=c(0,1),xlab="width (cm)",ylab="y")
> curve(ilogit(x,mod1$coef),add=TRUE,lwd=2)
> curve(iprobit(x,mod2$coef),add=TRUE,lwd=2,col=2,lty=2)
> curve(icloglog(x,mod3$coef),add=TRUE,lwd=2,col=3,lty=3)
```

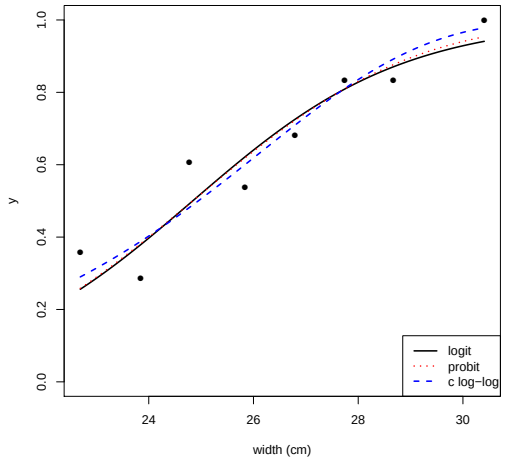

FIGURE 6.7: Graphique en nuage de points de y en fonction de x et courbe ajustées

On continue l'analyse avec le modèle 3, pour lequel le lien est

$$\eta_i = \ln(-\ln(1-\pi_i)).$$

Quelle est la moyenne ajustée $\hat{\pi}_i$ lorsque $x_i = 26.5$? On estime le prédicteur linéaire :

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = -8.174 + 0.313 \times 26.5.$$

Quelle est la variance estimée ?

$$\begin{aligned}
\widehat{\mathbb{V}\mathrm{ar}}(\hat{\eta}_i) &= \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_0) + \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_1 x_i) + 2\widehat{\mathbb{C}\mathrm{ov}}(\hat{\beta}_0, \hat{\beta}_1 x_i) \\
&= \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_0) + x_i^2 \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_1) + 2x_i \widehat{\mathbb{C}\mathrm{ov}}(\hat{\beta}_0, \hat{\beta}_1)
\end{aligned}$$

Selon les propriétés de la méthode du maximum de vraisemblance,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{\beta}, \frac{\mathcal{I}(\boldsymbol{\beta})^{-1}}{n}\right).$$

La matrice de variance-covariance estimée est donnée en R avec la commande suivante :

```
> (varcov <- summary(mod3)$cov.unscaled)
            (Intercept)        width
(Intercept)  2.51461812 -0.094581538
width       -0.09458154  0.003573437
```

On peut donc calculer la variance du prédicteur linéaire estimé :

$$\begin{aligned}\widehat{\mathbb{V}\mathrm{ar}}(\hat{\eta}_i) &= \widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_0) + x_i^2\widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}_1) + 2x_i\widehat{\mathbb{C}\mathrm{ov}}(\hat{\beta}_0, \hat{\beta}_1)\\ &= 2.51461812 + 26.5^2 \times 0.003573437 + 2 \times 26.5 \times -0.094581538\\ &= 0.01124285.\end{aligned}$$

Puisque $\hat{\eta}_i$ est une combinaison linéaire de variables aléatoires normalement distribuées (du moins, approximativement), alors

$$\hat{\eta}_i \approx \mathcal{N}(\eta_i, \widehat{\mathbb{V}\mathrm{ar}}(\hat{\eta}_i)).$$

Donc, un intervalle de confiance à 95% pour le prédicteur linéaire est

$$\hat{\eta}_i \pm z_{0.975}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat{\eta}_i)} = (-0.08788247,\, 0.32775659).$$

On peut donc trouver un intervalle de confiance sur la probabilité estimée :

$$\begin{aligned}&(1 - \exp(-\exp(-0.08788247)),\, 1 - \exp(-\exp(0.32775659)))\\ &= (0.5998311,\, 0.7503889).\end{aligned}$$

On forme le tableau de mauvaise classification en utilisant le seuil $\tau = 0.5$. D'abord, on calcule

$$\hat{Y}_i = \begin{cases} 1, & \text{si } \tau \geq 0.5,\\ 0, & \text{sinon.}\end{cases}$$

Par exemple,

```
> Ychap <- ifelse(fitted(mod3)>0.5,1,0)
```

Ensuite, on forme le tableau de mauvaise classification :

	Prédiction $\hat{Y}_i$	
Vrai Y_i	0	1
0	a	b
1	c	d

```
> aa <- sum((y==0)*(Ychap==0))
> dd <- sum((y==1)*(Ychap==1))
> cc <- sum((y==1)*(Ychap==0))
> bb <- sum((y==0)*(Ychap==1))
```

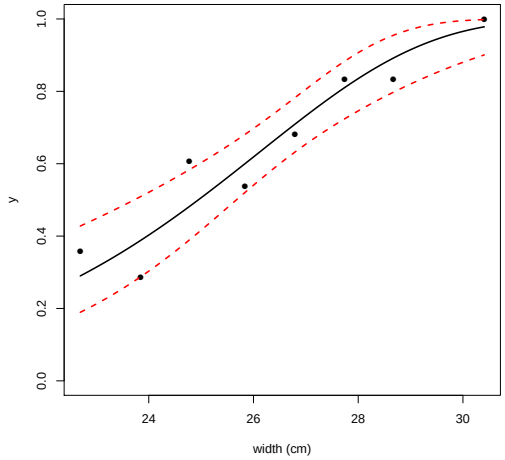

FIGURE 6.8: Prévisions et intervalle de confiance

```
> matrix(c(aa,bb,cc,dd),nrow=2)
     [,1] [,2]
[1,]   29   17
[2,]   33   94
```

La statistique de sensitivité est

$$\alpha(0.5) = \frac{d}{c+d} = \frac{94}{33+94} = 84.7\%.$$

Cela signifie que 85% des crabes qui ont un statellite sont catégorisés comme tel avec le modèle.

La statistique de spécificité est

$$\beta(0.5) = \frac{a}{a+b} = \frac{29}{29+17} = 46.8\%.$$

Cela signifie que 47% des crabes qui n'ont pas de statellite sont catégorisés comme tel avec le modèle.

Par contre, on ne veut pas nécessairement choisir $\tau = 0.5$, alors on trace la courbe ROC, soit $\alpha(\tau)$ en fonction de $1 - \beta(\tau)$, dans la Figure 6.9.

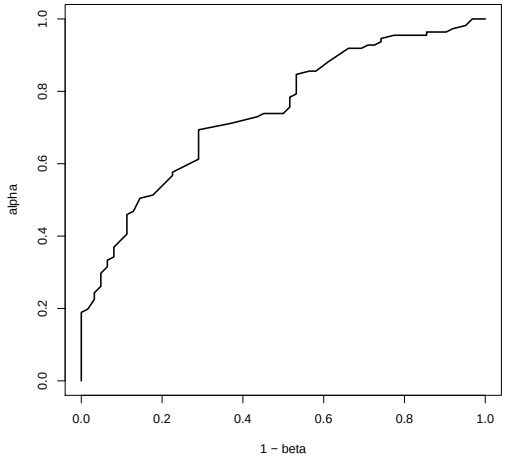

FIGURE 6.9: Courbe ROC

Modèle incluant l'effet de la couleur

On s'intéresse à l'effet de la couleur sur la probabilité d'avoir un satellite. Le modèle avec lien log-log complémentaire est

$$\ln(-\ln(1-\pi_i)) = \begin{cases} \beta_0 + \beta_1 x_i, & \text{si } C_i = 2 \\ \beta_0 + \beta_3^C + \beta_1 x_i, & \text{si } C_i = 3 \\ \beta_0 + \beta_4^C + \beta_1 x_i, & \text{si } C_i = 4 \\ \beta_0 + \beta_5^C + \beta_1 x_i, & \text{si } C_i = 5 \end{cases}$$

Dans ce modèle, la couleur translate la courbe, mais ne change pas sa forme.

```
> modcol <- glm(y~width+factor(color),family=binomial(link=cloglog))
> summary(modcol)
Call:
glm(formula = y ~ width + factor(color), family = binomial(link = cloglog))
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2174  -0.9769   0.4881   0.8913   1.9731
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.576039   1.696637  -4.465 7.99e-06 ***
width           0.292743   0.061393   4.768 1.86e-06 ***
factor(color)3  0.073797   0.407898   0.181    0.856
factor(color)4 -0.007385   0.438929  -0.017    0.987
factor(color)5 -0.881309   0.545422  -1.616    0.106
```

```
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 186.34  on 168  degrees of freedom
AIC: 196.34
Number of Fisher Scoring iterations: 6
```

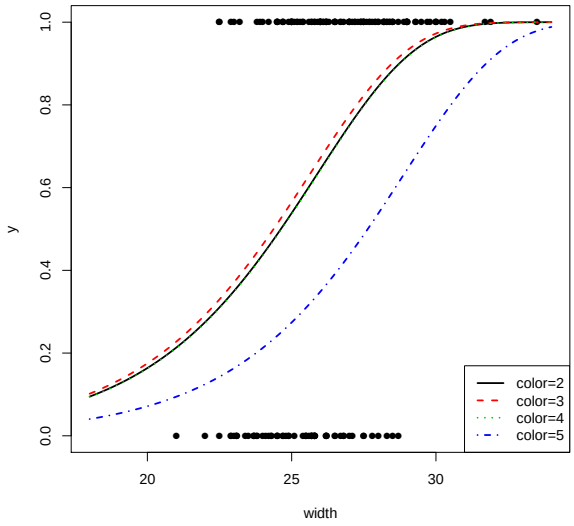

FIGURE 6.10: Ajustement modèle avec les variables exogènes `color` et `width`

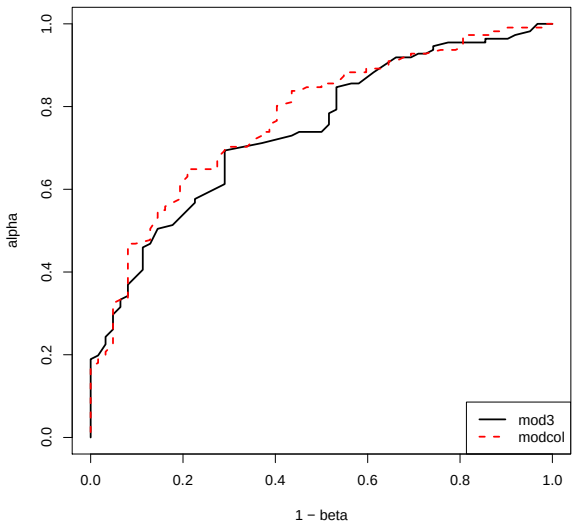

FIGURE 6.11: Courbes ROC

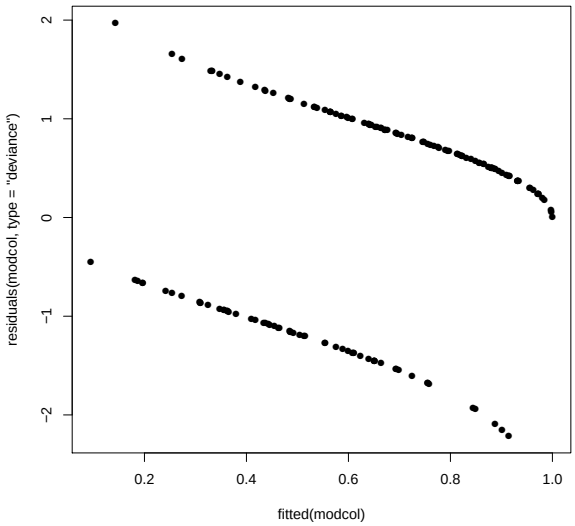

FIGURE 6.12: Résidus de déviance en fonction des moyennes prédites pour le modèle avec les variables exogènes `color` et `width`

Bibliographie

Agresti, A. (2013). *Categorical Data Analysis. Third Edition.* Wiley Series in Probability and Statistics. Wiley, New Jersey, USA.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22:134–137.

McCullagh, P. et Nelder, J. A. (1989). *Generalized linear models (Monographs on statistics and applied probability 37).* Chapman Hall, London.