Chapitre 5

Modélisation de données de comptage

La loi de Poisson (loi des petits nombres, Poisson (1837)) est très utile pour compter les occurences de certains évènements. Par exemple, cette loi a été appliquée dans la littérature pour compter le nombre de

1. soldats prussiens tués par ruade (von Bortkewisch (1898), voir Figure 5.1) ;
2. brevets (Hausmann, Hall et Griliches (1984)) ;
3. visites chez le médecin (Cameron et al (1988)) ;
4. faillites bancaires (Davutsyan (1989)) ;
5. publications chez les étudiants au doctorat (Long (1997)) ;
6. réclamations en assurance (Tröbliger (1961), Albrecht, Panjer,...).

Dans cette section, on s'intéresse aux données de comptage transversales. Cela signifie, par exemple, que nous ne suivons pas les assurés dans le temps, nous regardons seulement le nombre de réclamations pour une année.

5.1 Régression Poisson et modèles log-linéaires

Soit $Y_i \sim Poisson(\mu_i)$. Le lien canonique est $\theta_i = \ln(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$. On peut aussi utiliser le lien identité $g(t) = t$ et le lien racine carrée $g(t) = \sqrt{t}$. Toutefois, le lien log est préférable puisque cela permet de transformer la moyenne, qui prend des valeurs positives seulement, en prédicteur linéaire qui prend des valeurs dans $\mathbb{R}$.

De plus, si $g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$, alors

$$\frac{\partial}{\partial\beta_0}\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial\mu_i}{\partial\beta_0}\left(\frac{y_i}{\mu_i} - 1\right).$$

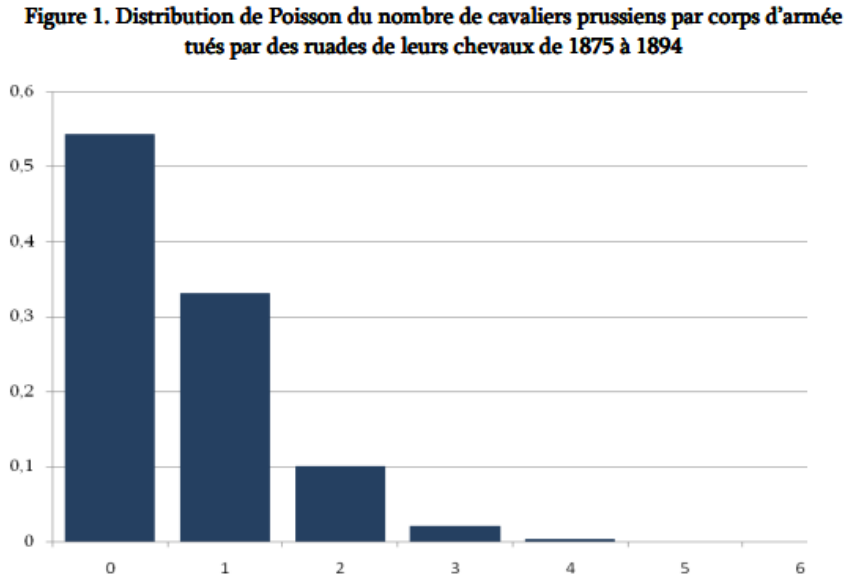

FIGURE 5.1: Selon von Bortkewisch (1898)

Sous le lien canonique, i.e. $g(\mu_i) = \ln(\mu_i)$:

$$\frac{\partial \mu_i}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \exp\left\{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right\} = \mu_i.$$

$$\Rightarrow \frac{\partial}{\partial \beta_0}\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mu_i)$$

Pour trouver l'estimateur du MV, on pose $\frac{\partial}{\partial \beta_0}\ell(\boldsymbol{\beta}) = 0$, ce qui implique que

$$\sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0 \Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i.$$

Finalement, la déviance est

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

On note que cette forme pour la déviance est valide seulement lorsque l'on utilise le lien canonique.

Exemple 5.1. *On s'intéresse au lien entre le nombre de cours de* Modèles Linéaires *où un étudiant était absent et son résultat à l'examen de mi-session. On suppose que le nombre de cours manqués suit une loi de Poisson avec paramètre*

$$\mu_i = \exp(\beta_0 + \beta_1 x_i),$$

où x_i est le résultat à l'examen pour l'étudiant i, $i = 1, ..., 20$. Les données sont présentées dans le tableau 5.1.

On désire utiliser l'algorithme de Newton-Raphson pour évaluer les paramètres selon la méthode du maximum de vraisemblance. La fonction de vraisemblance est

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = e^{-\sum_{i=1}^{n}\mu_i} \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!}.$$

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	90	0	11	57	0	21	48	0
2	93	0	12	41	0	22	45	1
3	47	3	13	31	2	23	99	0
4	79	0	14	51	0	24	74	0
5	82	0	15	55	0	25	41	3
6	97	0	16	35	0	26	88	0
7	77	0	17	97	0	27	31	0
8	56	1	18	28	1	28	87	0
9	67	0	19	51	0	29	80	0
10	80	0	20	48	0	30	86	0

TABLE 5.1: Résultats à l'examen (x_1) et nombre de cours manqués (y_i) pour l'exemple 5.1

La log-vraisemblance est donc

$$\begin{aligned}\ell(\beta_0,\beta_1) &= -\sum_{i=1}^{n}\mu_i + \sum_{i=1}^{n}\ln(\mu_i^{y_i}) - \sum_{i=1}^{n}\ln(y_i!)\\ &= -\sum_{i=1}^{n}\exp(\beta_0+\beta_1 x_i) + \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) + \textit{constante.}\end{aligned}$$

On utilise l'algorithme de Newton-Raphson. On a

$$\begin{aligned}\frac{\partial \ell(\beta_0,\beta_1)}{\partial \beta_0} &= \sum_{i=1}^{n}(y_i - \exp(\beta_0+\beta_1 x_i))\\ \frac{\partial \ell(\beta_0,\beta_1)}{\partial \beta_1} &= \sum_{i=1}^{n} x_i(y_i - \exp(\beta_0+\beta_1 x_i)),\end{aligned}$$

alors

$$\dot{\ell}(\boldsymbol{\beta}) = \begin{pmatrix}\sum_{i=1}^{n}(y_i - \exp(\beta_0+\beta_1 x_i))\\ \sum_{i=1}^{n} x_i(y_i - \exp(\beta_0+\beta_1 x_i))\end{pmatrix}.$$

Aussi,

$$\begin{aligned}\frac{\partial^2 \ell(\beta_0,\beta_1)}{\partial \beta_0^2} &= -\sum_{i=1}^{n}\exp(\beta_0+\beta_1 x_i)\\ \frac{\partial^2 \ell(\beta_0,\beta_1)}{\partial \beta_1^2} &= -\sum_{i=1}^{n} x_i^2\exp(\beta_0+\beta_1 x_i)\\ \frac{\partial^2 \ell(\beta_0,\beta_1)}{\partial \beta_1 \partial \beta_0} &= -\sum_{i=1}^{n} x_i\exp(\beta_0+\beta_1 x_i),\end{aligned}$$

alors

$$\ddot{\ell}(\boldsymbol{\beta}) = \begin{pmatrix}-\sum_{i=1}^{n}\exp(\beta_0+\beta_1 x_i) & -\sum_{i=1}^{n} x_i\exp(\beta_0+\beta_1 x_i)\\ -\sum_{i=1}^{n} x_i\exp(\beta_0+\beta_1 x_i) & -\sum_{i=1}^{n} x_i^2\exp(\beta_0+\beta_1 x_i)\end{pmatrix}.$$

On rappelle que chaque itération de l'algorithme de Newton-Raphson sera de la forme

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + \{-\ddot{l}(\hat{\beta}^{(m-1)})\}^{-1}\dot{\ell}(\hat{\beta}^{(m-1)}).$$

On choisit les valeurs de départ $\hat{\beta}_0 = \begin{pmatrix}2\\ -0.005\end{pmatrix}$. *La méthode et les résultats sont illustrés dans le code en* `R` *:*

```
> hbeta <- c(2,-0.005)
>
> for (i in 1:100)
+ {
+     mu <- exp(hbeta[1]+hbeta[2]*x)
+     dotl <- c(sum(y-mu),sum(x*(y-mu)))
+     a <- -sum(mu)
+     b <- -sum(x*mu)
+     d <- -sum(x^2*mu)
+     ddotl <- matrix(c(a,b,b,d),nrow=2)
+
+     vieux <- hbeta
+     (hbeta <- hbeta+solve(-ddotl)%*%dotl)
+
+     loglik <- -sum(exp(hbeta[1]+hbeta[2]*x))+sum(y*(hbeta[1]+hbeta[2]*x))-sum(gamma(y+1))
+     tol <- max(abs(sum(hbeta-vieux)),abs(sum(dotl)))
+     ifelse(tol<0.00005,break,i <- i+1)
+ }
>
> list(estimate=hbeta,std.error=sqrt(solve(-ddotl)[c(1,4)]),n.iter=i,loglik=loglik,tol=tol)
$estimate
            [,1]
[1,]  2.31538724
[2,] -0.06490573

$std.error
[1] 0.9613357 0.0220684

$n.iter
[1] 9

$loglik
[1] -56.06285

$tol
[1] 6.386731e-07
```

Plus simplement, on peut utiliser la fonction `glm` *en* `R` *:*

```
> summary(glm(y~x,family=poisson))
Call: glm(formula = y ~ x, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6458  -0.7434  -0.3356  -0.2000   2.4416
```

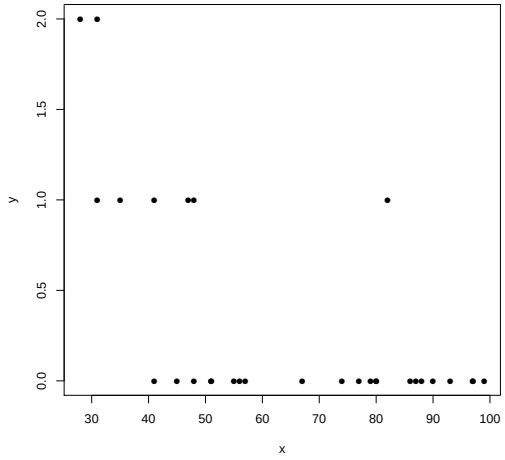

FIGURE 5.2: Résultats à l'examen (x_1) et nombre de cours manqués (y_i)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.31539    0.96134   2.409  0.01602 *
x           -0.06491    0.02207  -2.941  0.00327 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 38.029  on 29  degrees of freedom
Residual deviance: 24.082  on 28  degrees of freedom
AIC: 42.679
Number of Fisher Scoring iterations: 6
```

On peut aussi calculer la statistique X^2 de Pearson :

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Pour la Poisson, $V(\hat{\mu}_i) = \hat{\mu}_i$, et on trouve

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = 33.10.$$

Cela est calculé facilement en `R` *:*

`sum((y-fitted(modele))^2/fitted(modele))`

Les graphiques de résidus sont présentés dans la Figure 5.3.

□

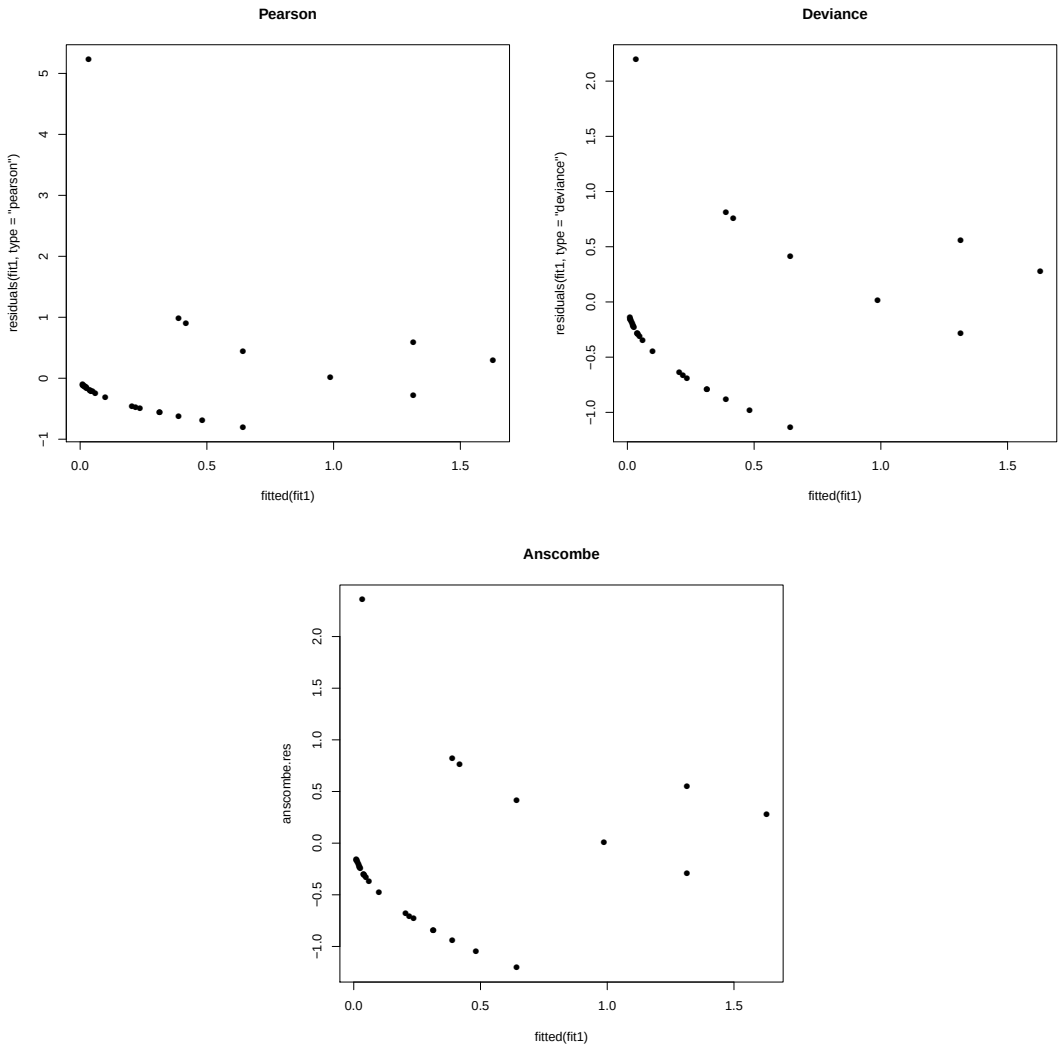

FIGURE 5.3: Résidus en fonction des moyennes ajustées

5.2 Terme "Offset"

Dans un contexte de modélisation du nombre de réclamations en assurance IARD, il arrive que les durées des contrats varient d'une observation à l'autre. Par exemple, un assuré peut résilier sa police après 1/2 année même s'il a subi une réclamation. On désire utiliser toutes les données pour ne pas biaiser les estimations. Pour ce faire, on utilise un modèle avec **offset**. Pour l'assuré i, soit Y_i le nombre de réclamations et t_i la durée de l'exposition au risque, qui est connue. On modélise l'espérance du *taux* de réclamation $E\left[Y_i/t_i\right]$:

$$\begin{aligned} \ln\left(\frac{\mu_i}{t_i}\right) &= \mathbf{x}_i\boldsymbol{\beta} \\ \ln(\mu_i) &= \mathbf{x}_i\boldsymbol{\beta} + \ln(t_i). \end{aligned}$$

Le terme $\ln(t_i)$ est un offset. On peut aussi le voir comme une variable explicative dont le coefficient est connu et est égal à 1.

Plus généralement, l'utilisation d'un terme offset permet d'inclure dans le modèle une variation systématique dans $E[Y_i]$ qui ne dépend pas des variables explicatives.

Exemple 5.2. On considère un modèle de Poisson pour le nombre de réclamations (`n.acc`). Deux variables qualitatives sont possibles : l'âge (groupé) et le sexe. Le modèle avec les effets principaux et les interactions, noté `Age*Sexe` ou `Age+Sexe+Age:Sexe` est construit de la façon suivante, avec le lien logarithmique :

$$\log(\mu_i) = \begin{cases} \alpha + \log(t_i) & , \text{ age}_i = \texttt{<25}, \text{ sexe}_i = H \\ \alpha + \beta_2^{age} + \log(t_i) & , \text{ age}_i = \texttt{25-39}, \text{ sexe}_i = H \\ \alpha + \beta_3^{age} + \log(t_i) & , \text{ age}_i = \texttt{>39}, \text{ sexe}_i = H \\ \alpha + \beta_2^{sexe} + \log(t_i) & , \text{ age}_i = \texttt{<25}, \text{ sexe}_i = F \\ \alpha + \beta_2^{age} + \beta_2^{sexe} + \gamma_{2,2} + \log(t_i) & , \text{ age}_i = \texttt{25-39}, \text{ sexe}_i = F \\ \alpha + \beta_3^{age} + \beta_2^{sexe} + \gamma_{3,2} + \log(t_i) & , \text{ age}_i = \texttt{>39}, \text{ sexe}_i = F \end{cases}.$$

Ce modèle peut être écrit de façon plus compacte. Les coefficients β_j^{age} et β_k^{sexe} sont les coefficients pour les effets principaux, alors que les coefficients $\gamma_{j,k}$ représentent les interactions. Le modèle pour la moyenne d'un assuré dont l'âge est de niveau j et le sexe de niveau k est donc

$$\log(\mu_{j,k}) = \alpha + \beta_j^{age} + \beta_k^{sexe} + \gamma_{j,k}, j = 1, 2, 3 \text{ et } k = 1, 2.$$

Par contre, il ne serait pas possible d'identifier tous les paramètres de ce modèle, il faut ajouter les contraintes d'identification :

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

Les données sont présentées dans le tableau 5.2.

On ajuste le modèle aux données :

i	age	sexe	t	n.acc	i	age	sexe	t	n.acc
1	<25	H	0.98	5	28	<25	F	0.54	4
2	25-39	F	0.54	0	29	25-39	H	0.63	0
3	>39	H	0.86	1	30	>39	F	0.90	0
4	<25	F	0.53	0	31	<25	H	0.59	3
5	25-39	H	0.60	2	32	25-39	F	0.67	0
6	>39	F	0.56	0	33	>39	H	0.73	3
7	<25	H	0.52	2	34	<25	F	0.75	1
8	25-39	F	0.93	1	35	25-39	H	0.82	0
9	>39	H	0.93	1	36	>39	F	0.72	0
10	<25	F	0.86	3	37	<25	H	0.75	4
11	25-39	H	0.52	2	38	25-39	F	0.89	0
12	>39	F	0.51	0	39	>39	H	0.81	1
13	<25	H	0.52	0	40	<25	F	0.68	0
14	25-39	F	0.89	0	41	25-39	H	0.79	1
15	>39	H	0.90	2	42	>39	F	0.78	1
16	<25	F	0.55	0	43	<25	H	0.89	6
17	25-39	H	0.66	1	44	25-39	F	0.87	0
18	>39	F	0.91	1	45	>39	H	0.81	0
19	<25	H	0.90	1	46	<25	F	0.87	4
20	25-39	F	0.90	0	47	25-39	H	0.68	2
21	>39	H	0.55	1	48	>39	F	0.59	1
22	<25	F	0.80	1	49	<25	H	0.67	1
23	25-39	H	0.76	2	50	25-39	F	0.87	3
24	>39	F	0.97	0	51	>39	H	0.78	2
25	<25	H	0.58	2	52	<25	F	0.61	1
26	25-39	F	0.71	3	53	25-39	H	0.68	4
27	>39	H	0.93	0	54	>39	F	0.70	1

TABLE 5.2: Données (simulées) pour l'exemple 5.2

```
> mod.offset.full <- glm(n.acc~age*sexe+offset(log(t)),family=poisson,data=dat)
> summary(mod.offset.full)

Call:
glm(formula = n.acc ~ age * sexe + offset(log(t)), family = poisson,
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9748  -1.1238  -0.3055   0.6628   2.0594

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.3218     0.2041   6.475 9.47e-11 ***
age25-39        -0.4975     0.3363  -1.479   0.1390
age>39          -0.9117     0.3641  -2.504   0.0123 *
sexeF           -0.5056     0.3363  -1.504   0.1327
age25-39:sexeF  -0.3564     0.5722  -0.623   0.5333
age>39:sexeF    -0.4112     0.6738  -0.610   0.5417
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 90.204  on 53  degrees of freedom
Residual deviance: 67.663  on 48  degrees of freedom
AIC: 167.73

Number of Fisher Scoring iterations: 6

> anova(mod.offset2)
Analysis of Deviance Table

Model: poisson, link: log

Response: n.acc

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev
NULL                        53     90.204
age       2  13.7228        51     76.481
sexe      1   8.2258        50     68.255
age:sexe  2   0.5920        48     67.663
```

```
> qchisq(0.99,c(1,2))
[1] 6.634897 9.210340
```

On trouve donc que le terme d'interaction n'est pas significatif. Le modèle contenant seulement les effets principaux, soit `age+sexe`, est une simplification adéquate du modèle complet. On a donc que

$$\log(\mu_i) = \begin{cases} \alpha & , \text{age}_i = \texttt{<25} \quad , \text{sexe}_i = H \\ \alpha + \beta_2^{age} & , \text{age}_i = \texttt{25-39} \quad , \text{sexe}_i = H \\ \alpha + \beta_3^{age} & , \text{age}_i = \texttt{>39} \quad , \text{sexe}_i = H \\ \alpha + \beta_2^{sexe} & , \text{age}_i = \texttt{<25} \quad , \text{sexe} = F \\ \alpha + \beta_2^{age} + \beta_2^{sexe} & , \text{age}_i = \texttt{25-39} \quad , \text{sexe}_i = F \\ \alpha + \beta_3^{age} + \beta_2^{sexe} & , \text{age}_i = \texttt{>39} \quad , \text{sexe}_i = F \end{cases}.$$

De façon plus succinte, le modèle pour la moyenne d'un assuré dont l'âge est de niveau j et le sexe de niveau k est donc

$$\log(\mu_{j,k}) = \alpha + \beta_j^{age} + \beta_k^{sexe}, j = 1, 2, 3 \text{ et } k = 1, 2,$$

avec les contraintes d'identification :

$$\beta_1^{age} = \beta_1^{sexe} = 0.$$

```
> mod.offset <- glm(n.acc~age+sexe+offset(log(t)),family=poisson,data=dat)
> summary(mod.offset)

Call:
glm(formula = n.acc ~ age + sexe + offset(log(t)), family = poisson,
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0378  -1.1957  -0.2119   0.5746   2.1457

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3845     0.1811   7.644 2.11e-14 ***
age25-39     -0.6224     0.2721  -2.287 0.022183 *
age>39       -1.0414     0.3049  -3.415 0.000638 ***
sexeF        -0.6862     0.2461  -2.789 0.005294 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 90.204  on 53  degrees of freedom
Residual deviance: 68.255  on 50  degrees of freedom
AIC: 164.32
```

```
Number of Fisher Scoring iterations: 5
```

5.2.1 Approximation de la Binomiale par une Poisson

Lorsqu'on a $Y \sim Bin(m, \pi)$ avec m très grand et π près de 0, alors la loi de Y ressemble à une loi de Poisson($m\pi$). On peut donc utiliser le modèle

$$\ln(\mu_i) = \ln(m_i) + \ln(\pi_i),$$

où $\ln(m_i)$ est un terme offset et $\ln(\pi_i) = \mathbf{x}_i\boldsymbol{\beta}$.

Exemple 5.3. Les données dans le tableau 5.3 montrent le nombre de décès dû au cancer du poumon par groupe d'âge (9 groupes différents), et selon quatre différents types de fumeurs :

1. Ne fume pas
2. Fume seulement la pipe ou des cigares
3. Fume des cigarettes et la pipe ou des cigares
4. Fume des cigarettes seulement.

On considère le modèle `fumeur+age`. Soit $j = 1, .., 4$ et $k = 1, ..., 9$, alors

$$\ln(\mu_{j,k}) = \ln(m_{j,k}) + \alpha + \beta_j^{fumeur} + \beta_k^{age},$$

avec la contrainte d'identification $\beta_1^{fumeur} = \beta_1^{age} = 0$ et $m_{j,k}$ est la population (en milliers) pour le groupe de fumeur j et le groupe d'âge k. On considère également trois autres modèles liés au premier, soient le modèle nul, noté `1`, et les modèles `fumeur` et `age`.

Code et résultats :

		Âge								
Fumeur		40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80+
1	Morts	18	22	19	55	117	170	179	120	120
1	Pop	656	359	249	632	1067	897	668	361	274
2	Morts	2	4	3	38	113	173	212	243	253
2	Pop	145	104	98	372	846	949	824	667	537
3	Morts	149	169	193	576	1001	901	613	337	189
3	Pop	4531	3030	2267	4682	6052	3880	2033	871	345
4	Morts	124	140	187	514	778	689	432	214	63
4	Pop	3410	2239	1851	3270	3791	2421	1195	436	113

TABLE 5.3: Données pour l'exemple 5.3

```
> mod1 <- glm(Mort~offset(log(Pop)),family=poisson,data=poumon)
> mod.fume <- glm(Mort~offset(log(Pop))+Fumeur,family=poisson,data=poumon)
> mod.age <- glm(Mort~offset(log(Pop))+Age,family=poisson,data=poumon)
> mod.tot <- glm(Mort~offset(log(Pop))+Fumeur+Age,family=poisson,data=poumon)
> anova(mod1,mod.fume,mod.age,mod.tot)
Analysis of Deviance Table

Model 1: Mort ~ offset(log(Pop))
Model 2: Mort ~ offset(log(Pop)) + Fumeur
Model 3: Mort ~ offset(log(Pop)) + Age
Model 4: Mort ~ offset(log(Pop)) + Fumeur + Age
  Resid. Df Resid. Dev Df Deviance
1        35     4056.0
2        32     3910.7  3    145.3
3        27      191.7  5   3719.0
4        24       21.5  3    170.2
> summary(mod.tot)

Call:
glm(formula = Mort ~ offset(log(Pop)) + Fumeur + Age, family = poisson,
    data = poumon)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.06055  -0.54773   0.06431   0.29963   1.48348

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.68002    0.06824 -53.929  < 2e-16 ***
Fumeur2      0.04781    0.04699   1.017    0.309
Fumeur3      0.21796    0.03869   5.633 1.77e-08 ***
Fumeur4      0.41696    0.03991  10.447  < 2e-16 ***
Age45-49     0.55388    0.07999   6.924 4.38e-12 ***
Age50-54     0.98039    0.07682  12.762  < 2e-16 ***
Age55-59     1.37946    0.06526  21.138  < 2e-16 ***
Age60-64     1.65423    0.06257  26.439  < 2e-16 ***
Age65-69     1.99817    0.06279  31.824  < 2e-16 ***
Age70-74     2.27141    0.06435  35.296  < 2e-16 ***
Age75-79     2.55858    0.06778  37.746  < 2e-16 ***
Age80+       2.84692    0.07242  39.310  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 4055.984  on 35  degrees of freedom
Residual deviance:   21.487  on 24  degrees of freedom
AIC: 285.51

Number of Fisher Scoring iterations: 4
```

□

5.3 Tableau de contingence

Un tableau de contingence (contingency table) est un moyen de représenter des données de comptage. La forme la plus simple est lorsqu'il y a deux variables explicatives qualitatives, et qu'il n'y a pas de variable réponse évidente autre que le nombre dans chaque cellule du tableau. On peut utiliser un GLM Poisson pour modéliser le nombre dans chaque cellule du tableau et ainsi comprendre si les effets des variables explicatives sont reliés ou non.

Plus spécifiquement, on s'intéresse à trois structures de modèles. Pour fins d'illustration, soit les trois variables qualitatives A, B, C avec J, K et L niveaux, respectivement.

Modèle d'indépendance : $A + B + C$, avec

$$\ln(\mu_{jkl}) = \alpha + \beta_j^A + \beta_k^B + \beta_l^C,$$

et $\beta_1^A = \beta_1^B = \beta_1^C = 0$. On compte $1 + (J-1) + (K-1) + (L-1)$ paramètres dans ce modèle. Les effets des variables explicatives sont indépendants, il n'y a pas de termes d'interaction.

Modèle d'indépendance partielle : $A + B * C = A + B + C + B.C$, avec

$$\ln(\mu_{jkl}) = \alpha + \beta_j^A + \beta_k^B + \beta_l^C + \gamma_{kl}^{BC},$$

$\beta_1^A = \beta_1^B = \beta_1^C = 0$ et $\gamma_{1l}^{BC} = \gamma_{k1}^{BC} = 0$. Dans ce modèle, l'effet du facteur A est indépendant de l'effet conjoint des facteurs (B, C). Pour un j fixé, la contribution des facteurs (B, C) est toujours

$$\beta_k^B + \beta_l^C + \gamma_{kl}^{BC},$$

peu importe la valeur de j choisie.

Modèle d'indépendance conditionelle : $A * B + B * C = A + B + C + A.B + B.C$, avec

$$\ln(\mu_{jkl}) = \alpha + \beta_j^A + \beta_k^B + \beta_l^C + \gamma_{jk}^{AB} + \gamma_{kl}^{BC},$$

$\beta_1^A = \beta_1^B = \beta_1^C = 0$ et $\gamma_{1k}^{AB} = \gamma_{j1}^{AB} = \gamma_{1l}^{BC} = \gamma_{k1}^{BC} = 0$. Si on conditionne sur B, les facteurs A et C influencent la variable réponse de façon indépendante. On peut tester si l'indépendance conditionelle est une hypothèse adéquate en utilisant l'analyse de déviance des modèles

$$A * B + B * C \text{ et } A * B * C.$$

	Fumeur		Non-Fumeur	
Age	Décès	Vie	Décès	Vie
18-24	2	53	1	61
25-34	3	121	5	152
35-44	14	95	7	114
45-54	27	103	12	66
55-64	51	64	40	81
65-74	29	7	101	28
75+	13	0	64	0

TABLE 5.4: Données pour l'exemple 5.5

Exercice 5.4. *Déterminer le nombre de paramètres dans les modèles d'indépendance partielle et d'indépendance conditionelle décrits ci-haut.* □

Exemple 5.5. *Cet exemple est tiré de Faraway (2005), d'après l'étude de Appleton, French, and Vanderpump (1996). On s'intéresse aux effets de la cigarette sur la durée de vie des femmes. 1314 femmes fumeuses et non-fumeuses ont été catégorisées dans sept groupes d'âges. Lors du suivi vingt ans plus tard, les chercheurs ont noté quels sujets étaient décédés ou vivants. Les femmes qui avaient cessé de fumer pendant l'étude ont été exclues des données, qui sont présentées dans le tableau 5.4. On considère un modèle de Poisson pour cette table de contingence à trois facteurs (*`age, smoke, dead`*).*

On ajuste d'abord le modèle saturé `age*smoke*dead`*. Ce modèle n'a aucun degré de liberté car le nombre de paramètre est égal au nombre de données, et le modèle reproduit exactement les observations.*

Code et résultats :

```
> library(faraway)
> data(femsmoke)
> satur<-glm(y~smoker*dead*age,family=poisson,data=femsmoke)
> summary(satur)

Call:
glm(formula = y ~ smoker * dead * age, family = poisson, data = femsmoke)

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[26]  0  0  0

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             6.931e-01  7.071e-01   0.980 0.326959
```

```
smokerno                  -6.931e-01  1.225e+00  -0.566 0.571426
deadno                     3.277e+00  7.203e-01   4.550 5.38e-06 ***
age25-34                   4.055e-01  9.129e-01   0.444 0.656923
age35-44                   1.946e+00  7.559e-01   2.574 0.010047 *
age45-54                   2.603e+00  7.328e-01   3.552 0.000383 ***
age55-64                   3.239e+00  7.208e-01   4.493 7.02e-06 ***
age65-74                   2.674e+00  7.311e-01   3.658 0.000254 ***
age75+                     1.872e+00  7.596e-01   2.464 0.013727 *
smokerno:deadno            8.337e-01  1.239e+00   0.673 0.501027
smokerno:age25-34          1.204e+00  1.426e+00   0.844 0.398485
smokerno:age35-44         -4.371e-15  1.309e+00   0.000 1.000000
smokerno:age45-54         -1.178e-01  1.273e+00  -0.093 0.926278
smokerno:age55-64          4.502e-01  1.243e+00   0.362 0.717172
smokerno:age65-74          1.941e+00  1.243e+00   1.562 0.118321
smokerno:age75+            2.287e+00  1.262e+00   1.812 0.069937 .
deadno:age25-34            4.200e-01  9.276e-01   0.453 0.650685
deadno:age35-44           -1.362e+00  7.751e-01  -1.758 0.078824 .
deadno:age45-54           -1.938e+00  7.521e-01  -2.577 0.009960 **
deadno:age55-64           -3.050e+00  7.444e-01  -4.097 4.18e-05 ***
deadno:age65-74           -4.699e+00  8.344e-01  -5.631 1.79e-08 ***
deadno:age75+             -2.914e+01  6.965e+04   0.000 0.999666
smokerno:deadno:age25-34 -1.116e+00  1.443e+00  -0.773 0.439232
smokerno:deadno:age35-44  4.174e-02  1.330e+00   0.031 0.974964
smokerno:deadno:age45-54 -4.679e-01  1.296e+00  -0.361 0.718160
smokerno:deadno:age55-64 -3.552e-01  1.268e+00  -0.280 0.779372
smokerno:deadno:age65-74 -6.953e-01  1.326e+00  -0.524 0.600044
smokerno:deadno:age75+   -2.428e+00  9.851e+04   0.000 0.999980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.1939e+03  on 27  degrees of freedom
Residual deviance: 3.0336e-10  on  0  degrees of freedom
AIC: 190.19

Number of Fisher Scoring iterations: 21

> anova(satur)
Analysis of Deviance Table

Model: poisson, link: log

Response: y
```

```
Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                                27    1193.94
smoker          1    17.16        26    1176.78
dead            1   261.27        25     915.50
age             6   180.50        19     735.00
smoker:dead     1     9.20        18     725.80
smoker:age      6    93.51        12     632.30
dead:age        6   629.92         6       2.38
smoker:dead:age 6     2.38         0       0.00
> mod2<-update(satur,~.-smoker:dead:age)
> summary(mod2)

Call:
glm(formula = y ~ smoker + dead + age + smoker:dead + smoker:age +
    dead:age, family = poisson, data = femsmoke)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.70006  -0.11004  -0.00002   0.12254   0.67272

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.54284    0.58736   0.924 0.355384
smokerno           -0.29666    0.25324  -1.171 0.241401
deadno              3.43271    0.59014   5.817 6.00e-09 ***
age25-34            0.92902    0.68381   1.359 0.174273
age35-44            1.94048    0.62486   3.105 0.001900 **
age45-54            2.76845    0.60657   4.564 5.02e-06 ***
age55-64            3.37507    0.59550   5.668 1.45e-08 ***
age65-74            2.86586    0.60894   4.706 2.52e-06 ***
age75+              2.02211    0.64955   3.113 0.001851 **
smokerno:deadno     0.42741    0.17703   2.414 0.015762 *
smokerno:age25-34   0.11752    0.22091   0.532 0.594749
smokerno:age35-44   0.01268    0.22800   0.056 0.955654
smokerno:age45-54  -0.56538    0.23585  -2.397 0.016522 *
smokerno:age55-64   0.08512    0.23573   0.361 0.718030
smokerno:age65-74   1.49088    0.30039   4.963 6.93e-07 ***
smokerno:age75+     1.89060    0.39582   4.776 1.78e-06 ***
deadno:age25-34    -0.12006    0.68655  -0.175 0.861178
deadno:age35-44    -1.34112    0.62857  -2.134 0.032874 *
deadno:age45-54    -2.11336    0.61210  -3.453 0.000555 ***
deadno:age55-64    -3.18077    0.60057  -5.296 1.18e-07 ***
```

```
deadno:age65-74     -5.08798    0.61951  -8.213  < 2e-16 ***
deadno:age75+      -27.31727 8839.01146  -0.003 0.997534
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1193.9378  on 27  degrees of freedom
Residual deviance:    2.3809  on  6  degrees of freedom
AIC: 180.58

Number of Fisher Scoring iterations: 18

> drop1(mod2,test="Chisq")
Single term deletions

Model:
y ~ smoker + dead + age + smoker:dead + smoker:age + dead:age
            Df Deviance    AIC    LRT Pr(>Chi)
<none>              2.38 180.58
smoker:dead  1     8.33 184.52   5.95  0.01475 *
smoker:age   6    92.63 258.83  90.25  < 2e-16 ***
dead:age     6   632.30 798.49 629.92  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> mod3<-update(mod2,~.-smoker:dead)
> summary(mod3)

Call:
glm(formula = y ~ smoker + dead + age + smoker:age + dead:age,
    family = poisson, data = femsmoke)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.30657  -0.26480  -0.00003   0.26643   1.20822

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.34377    0.58563   0.587 0.557199
smokerno            0.11980    0.18523   0.647 0.517785
deadno              3.63759    0.58490   6.219 5.00e-10 ***
age25-34            0.91760    0.68737   1.335 0.181895
age35-44            1.95402    0.62882   3.107 0.001887 **
age45-54            2.84979    0.60950   4.676 2.93e-06 ***
age55-64            3.44819    0.59868   5.760 8.43e-09 ***
```

```
age65-74             3.00134    0.61023   4.918 8.73e-07 ***
age75+               2.22118    0.64799   3.428 0.000609 ***
smokerno:age25-34    0.11616    0.22078   0.526 0.598789
smokerno:age35-44   -0.01536    0.22749  -0.068 0.946172
smokerno:age45-54   -0.63063    0.23414  -2.693 0.007074 **
smokerno:age55-64   -0.06894    0.22643  -0.304 0.760765
smokerno:age65-74    1.15649    0.26427   4.376 1.21e-05 ***
smokerno:age75+      1.47413    0.35617   4.139 3.49e-05 ***
deadno:age25-34     -0.10756    0.68613  -0.157 0.875435
deadno:age35-44     -1.33977    0.62810  -2.133 0.032920 *
deadno:age45-54     -2.17125    0.61128  -3.552 0.000382 ***
deadno:age55-64     -3.17171    0.59999  -5.286 1.25e-07 ***
deadno:age65-74     -4.94977    0.61512  -8.047 8.49e-16 ***
deadno:age75+      -26.30450 5776.51889  -0.005 0.996367
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1193.9378  on 27  degrees of freedom
Residual deviance:    8.3269  on  7  degrees of freedom
AIC: 184.52

Number of Fisher Scoring iterations: 17
```

On conclut que le modèle d'indépendance conditionnelle est adéquat pour ces données. □

Exemple 5.6. *La base de données* `esoph` *du package* `datasets` *en* `R` *contient des observations du nombre de cas de cancer de l'oesophage, en fonction de*

- *l'âge :*
  1. *25-34*
  2. *35-44*
  3. *45-54*
  4. *55-64*
  5. *65-74*
  6. *75+*
- *la consommation d'alcool :*
  1. *0-39 grammes/jour*
  2. *40-79 grammes/jour*
  3. *80-119 grammes/jour*
  4. *120+ grammes/jour*

- *et la consommation de tabac :*
  1. *0-9 grammes/jour*
  2. *10-19 grammes/jour*
  3. *20-29 grammes/jour*
  4. *30+ grammes/jour*

On suppose que le nombre de cas de cancer de l'oesophage dans chaque catégorie suit une loi de Poisson. On utilise le lien canonique. Pour le groupe d'âge j, $j = 1, ..., 6$, la consommation d'alcool k, $k = 1, ..., 4$ et la consommation de tabac l, $l = 1, ..., 4$, on a

$$\ln(\mu_{j,k,l}) = \alpha + \beta_j^{age} + \beta_k^{alc} + \beta_l^{tab},$$

avec la contrainte d'identification $\beta_1^{age} = \beta_1^{alc} = \beta_1^{tab} = 0$.

Code et résultats

```
> library(datasets)
> fit1<-glm(ncases~agegp+alcgp+tobgp,family=poisson,data=esoph)
> summary(fit1)

Call:
glm(formula = ncases ~ agegp + alcgp + tobgp, family = poisson,
    data = oesop)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9090  -0.8011  -0.2969   0.4882   2.5841

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.008109   0.189930  -0.043 0.965946
agegp.L      2.351080   0.633962   3.709 0.000208 ***
agegp.Q     -2.713488   0.573609  -4.731 2.24e-06 ***
agegp.C     -0.091883   0.433582  -0.212 0.832172
agegp^4      0.036029   0.291798   0.123 0.901733
agegp^5     -0.091050   0.175969  -0.517 0.604862
alcgp.L      0.218055   0.164895   1.322 0.186039
alcgp.Q     -0.553297   0.149818  -3.693 0.000222 ***
alcgp.C      0.377895   0.133162   2.838 0.004542 **
tobgp.L     -0.620613   0.151185  -4.105 4.04e-05 ***
tobgp.Q      0.176095   0.152489   1.155 0.248168
tobgp.C      0.175952   0.154042   1.142 0.253358
---
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 262.926  on 87  degrees of freedom
Residual deviance:  78.395  on 76  degrees of freedom
AIC: 272.1

Number of Fisher Scoring iterations: 6

> drop1(fit1,test="Chisq")
Single term deletions

Model:
ncases ~ agegp + alcgp + tobgp
       Df Deviance    AIC     LRT  Pr(>Chi)
<none>      78.395 272.10
agegp   5  223.466 407.17 145.071 < 2.2e-16 ***
alcgp   3  101.841 289.54  23.446 3.260e-05 ***
tobgp   3  100.564 288.26  22.169 6.016e-05 ***
---
```

□

5.4 Sous-dispersion et surdispersion

Avec la loi Poisson, on suppose que $E[Y_i|x_i] = Var(Y_i|x_i)$. Ce postulat est assez contraignant.

On dit que les données sont **sous-dispersées** ou qu'il y a présence de sous-dispersion si

$$E[Y_i|x_i] > Var(Y_i|x_i).$$

La sous-dispersion peut être détectée si la déviance réduite divisée par ses degrés de liberté $\frac{D(y,\hat{\mu})}{\hat{\phi}\, dl}$ est largement inférieure à 1 (disons, $<$0.6) ou si $\frac{X^2}{dl} < 0.6$. On peut souvent régler un problème de sous-dispersion en utilisant la distribution binomiale ou la Poisson "gonflée à 0".

Il est assez fréquent que les données d'assurance soit **surdispersées**, c'est à dire que

$$E[Y_i|x_i] < Var(Y_i|x_i).$$

Une indication de ce problème est lorsque la Déviance réduite est supérieure aux degrés de liberté de façon significative (ex : $D(y,\hat{\mu})/dl > 1.7$) ou si on a $X^2 > 1.7dl$. Dans ce cas, le modèle de Poisson n'est pas adéquat. Plusieurs modèles peuvent être utilisés afin de résoudre ce problème, mais la plupart ne seront

pas abordés dans le cours puisqu'ils sont basés sur la généralisation de l'équation de score de GLM, et sur la notion de quasi-vraisemblance.

On suppose que $Y|Z = z \sim Poisson(\mu z)$, où Z est une variable aléatoire non-négative. Alors, en conditionnant, on trouve les moments de Y :

$$\begin{aligned} E[Y] &= E[E[Y|Z]] = E[\mu Z] = \mu E[Z] \\ Var(Y) &= E[Var(Y|Z)] + Var(E[Y|Z]) \\ &= E[\mu Z] + Var(\mu Z) = \mu E[Z] + \mu^2 Var(Z). \end{aligned}$$

Si $Z \sim Gamma(\theta_z, \theta_z)$, i.e. $E[Z] = 1$ et $Var(Z) = 1/\theta_z$, alors

$$\begin{aligned} E[Y] &= E[E[Y|Z]] = E[\mu Z] = \mu \\ Var(Y) &= \mu E[Z] + \mu^2 Var(Z) = \mu + \frac{\mu^2}{\theta_z}. \end{aligned}$$

En fait, on peut montrer que $Y \sim BinNeg(\mu, \theta_z)$, soit

$$f_Y(y) = \frac{\Gamma(\theta_z + y)}{\Gamma(\theta_z) y!} \left(\frac{\mu}{\mu + \theta_z}\right)^y \left(\frac{\theta_z}{\mu + \theta_z}\right)^{\theta_z}.$$

On peut facilement adapter ce modèle pour utiliser la théorie des GLMs, en posant, par exemple

$$\ln(\mu_i) = x_i \boldsymbol{\beta},$$

et en utilisant le maximum de vraisemblance pour estimer les paramètres.

Remarque 5.7. *En* `R`*, la fonction* `glm.nb` *du package* `MASS` *permet d'estimer les paramètres* $\boldsymbol{\beta}$ *et* θ_z *de ce modèle.*

Les modèles Poisson et Binomiale négative sont liés. On note que

$$\lim_{\theta \to \infty} Var(Y) = \lim_{\theta \to \infty} \mu + \frac{\mu^2}{\theta} = \mu$$

et on retrouve le modèle Poisson. On peut donc tester avec le TRV si le modèle Poisson est une simplification adéquate du modèle NB. Soit $\varphi = 1/\theta_z$, on teste

$$H_0 : \varphi = 0 \text{ et } H_1 : \varphi > 0.$$

Par contre, puisque 0 est la borne inférieure du domaine du paramètre φ, il faut apporter un petit changement au test habituel. Si $Q \sim \chi^2_{(1)}$, alors

$$P[2(\ell^{Pois}(\hat{\beta}_n) - \ell^{NB}(\hat{\beta}_n, \hat{\varphi})) > x] = \frac{1}{2} P[Q > x].$$

Remarque 5.8. *On peut aussi considérer d'autres modèles où l'hétérogénéité, plutôt que de suivre une loi Gamma, suit une loi inverse-gaussienne ou lognormale. Le modèle de Poisson 'gonflé à zéro' est aussi un bon candidat pour modéliser le nombre de sinistres.*

Exemple 5.9. ***Exemple du nombre d'espèces de plantes dans les îles Galapagos*** Cet exemple utilise des données de Faraway (2005). Il y a 30 îles différentes et 6 variables dans la base de données. On s'intéresse à la relation entre le nombre d'espèces de plantes sur une île et plusieurs facteurs géograpiques.

- `Species` : Nombre d'espèces de plantes trouvées sur l'île,
- `Area` : l'aire de l'île (km carrés),
- `Elevation` : le point le plus haut de l'île (m),
- `Nearest` : la distance jusqu'à l'île la plus proche (km),
- `Scruz` : la distance de l'île de Santa Cruz (km),
- `Adjacent` : l'aire de l'île adjacente (km carrés).

On considère d'abord un modèle de Poisson avec les effets principaux seulement.

```
> mod1 <- glm(Species~.,family=poisson,data=gala)
> summary(mod1)
Call: glm(formula = Species ~ ., family = poisson, data = gala)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68
Number of Fisher Scoring iterations: 5
```

Tous les termes sont hautement significatifs, mais la déviance est très élevée. Peut-être que l'on peut améliorer le modèle en ajouter des termes d'interaction de deuxième ordre.

```
> mod2 <- glm(Species~(Area+Elevation+Nearest+Scruz+Adjacent)^2,
+family=poisson,data=gala)
> drop1(mod2,test="Chisq")
Single term deletions
Model:
Species ~ (Area + Elevation + Nearest + Scruz + Adjacent)^2
                  Df Deviance    AIC    LRT  Pr(>Chi)
```

```
<none>                  317.78 510.61
Area:Elevation      1   358.16 548.99 40.379 2.092e-10 ***
Area:Nearest        1   354.96 545.79 37.178 1.078e-09 ***
Area:Scruz          1   365.41 556.24 47.627 5.154e-12 ***
Area:Adjacent       1   400.50 591.33 82.720 < 2.2e-16 ***
Elevation:Nearest   1   323.03 513.86  5.248  0.021978 *
Elevation:Scruz     1   322.99 513.82  5.215  0.022396 *
Elevation:Adjacent  1   384.89 575.72 67.108 2.571e-16 ***
Nearest:Scruz       1   358.85 549.68 41.070 1.469e-10 ***
Nearest:Adjacent    1   325.52 516.35  7.739  0.005404 **
Scruz:Adjacent      1   323.23 514.07  5.456  0.019499 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec la fonction `drop1`, on observe que certains termes d'interaction ne sont pas significatifs à 99%. Par exemple, on teste

$$H_0 : \gamma^{elevation,scruz} = 0 \text{ versus } H_1 : \gamma^{elevation,scruz} \neq 0.$$

On trouve que la statistique du test de rapport de vraisemblance est la différence entre les déviances des modèles :

$$\Delta Deviance = 322.99 - 317.78 = 5.215.$$

Le nombre de degrés de libertés est 1, et on trouve que

$$P[\chi^2_{(1)} > 5.215] = 0.022396.$$

On décide donc d'enlever le terme d'interaction `Elevation:Scruz`.

```
> mod3 <- update(mod2,~.-Elevation:Scruz)
> drop1(mod3,test="Chisq")
Single term deletions
Model:
Species ~ Area + Elevation + Nearest + Scruz + Adjacent + Area:Elevation +
    Area:Nearest + Area:Scruz + Area:Adjacent + Elevation:Nearest +
    Elevation:Adjacent + Nearest:Scruz + Nearest:Adjacent + Scruz:Adjacent
                  Df Deviance    AIC    LRT  Pr(>Chi)
<none>                 322.99 513.82
Area:Elevation     1   360.02 548.85 37.023 1.167e-09 ***
Area:Nearest       1   395.40 584.23 72.403 < 2.2e-16 ***
Area:Scruz         1   411.98 600.81 88.990 < 2.2e-16 ***
Area:Adjacent      1   406.84 595.67 83.848 < 2.2e-16 ***
Elevation:Nearest  1   393.94 582.77 70.949 < 2.2e-16 ***
Elevation:Adjacent 1   385.00 573.83 62.005 3.426e-15 ***
Nearest:Scruz      1   381.76 570.59 58.766 1.776e-14 ***
Nearest:Adjacent   1   326.33 515.16  3.341   0.06758 .
```

```
Scruz:Adjacent      1   329.47 518.30  6.480   0.01091 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec la fonction `drop1`, on observe que certains termes d'interaction ne sont pas significatifs à 99%. Par exemple, on teste

$$H_0 : \gamma^{nearest,adjacent} = 0 \text{ versus } H_1 : \gamma^{nearest,adjacent} \neq 0.$$

On trouve que la statistique du test de rapport de vraisemblance est la différence entre les déviances des modèles :

$$\Delta Deviance = 326.33 - 322.99 = 3.341.$$

Le nombre de degrés de libertés est 1, et on trouve que

$$P[\chi^2_{(1)} > 5.215] = 0.06758.$$

On décide donc d'enlever le terme d'interaction `Nearest:Adjacent`.

```
> mod4 <- update(mod3,~.-Nearest:Adjacent)
> drop1(mod4,test="Chisq")
Single term deletions

Model:
Species ~ Area + Elevation + Nearest + Scruz + Adjacent + Area:Elevation +
    Area:Nearest + Area:Scruz + Area:Adjacent + Elevation:Nearest +
    Elevation:Adjacent + Nearest:Scruz + Scruz:Adjacent
                  Df Deviance    AIC    LRT  Pr(>Chi)
<none>                 326.33 515.16
Area:Elevation     1   360.02 546.85 33.684 6.484e-09 ***
Area:Nearest       1   397.64 584.47 71.303 < 2.2e-16 ***
Area:Scruz         1   412.63 599.46 86.296 < 2.2e-16 ***
Area:Adjacent      1   408.48 595.31 82.149 < 2.2e-16 ***
Elevation:Nearest  1   394.91 581.74 68.579 < 2.2e-16 ***
Elevation:Adjacent 1   391.93 578.76 65.597 5.532e-16 ***
Nearest:Scruz      1   403.50 590.33 77.169 < 2.2e-16 ***
Scruz:Adjacent     1   338.39 525.22 12.055 0.0005165 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec la fonction `drop1`, on observe que tous les termes d'interaction sont significatifs à 99%. On ne peut donc plus réduire le modèle. Cela signifierait que, si le modèle est adéquat, tous les facteurs géographiques considérés ont un impact sur le nombre d'espèces de plantes que l'on retrouve sur une île.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-8.8730  -1.3351   0.1126   1.4358   5.6093
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.781e+00  1.377e-01  12.941  < 2e-16 ***
Area                4.334e-03  1.049e-03   4.130 3.63e-05 ***
Elevation           6.157e-03  3.554e-04  17.324  < 2e-16 ***
Nearest             4.479e-02  7.702e-03   5.816 6.03e-09 ***
Scruz               1.499e-02  1.811e-03   8.281  < 2e-16 ***
Adjacent            1.191e-03  2.669e-04   4.462 8.12e-06 ***
Area:Elevation     -6.464e-06  1.114e-06  -5.804 6.49e-09 ***
Area:Nearest        2.362e-04  2.874e-05   8.217  < 2e-16 ***
Area:Scruz         -1.056e-04  1.162e-05  -9.091  < 2e-16 ***
Area:Adjacent       1.588e-05  1.769e-06   8.977  < 2e-16 ***
Elevation:Nearest  -1.152e-04  1.382e-05  -8.336  < 2e-16 ***
Elevation:Adjacent -1.015e-05  1.202e-06  -8.437  < 2e-16 ***
Nearest:Scruz      -5.651e-04  6.675e-05  -8.467  < 2e-16 ***
Scruz:Adjacent      4.538e-05  1.289e-05   3.520 0.000431 ***
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  326.33  on 16  degrees of freedom
AIC: 515.16
Number of Fisher Scoring iterations: 5
```

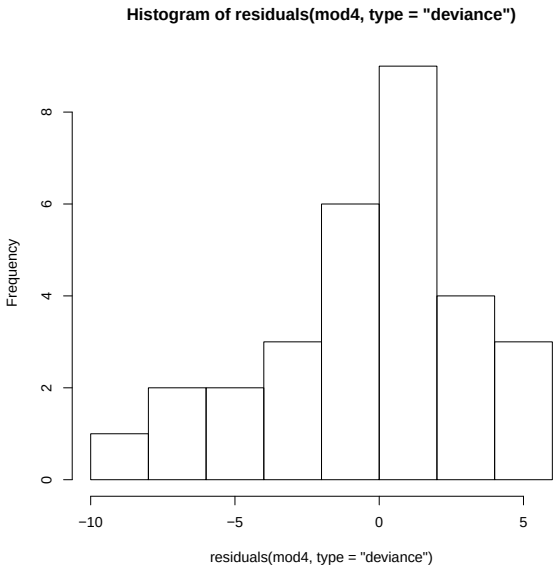

FIGURE 5.4: Résidus pour l'exemple des espèces de plantes dans les îles Galapagos

On observe plusieurs problèmes dans le graphique 5.4 :

1. Les résidus de déviance sont compris entre -8.8 et 5.6. L'étendue est beaucoup plus grande que celle d'une loi Normale centrée réduite.
2. La déviance est de 326.33 sur 16 degrés de liberté. Le modèle n'est pas adéquat parce que 326.33/16 est beaucoup plus grand que 1.
3. Le X^2 de Pearson est
$$X^2 = \sum_{i=1}^{30} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = 276.2253,$$
ce qui est également beaucoup plus élevé que 16.

Il y a présence de surdispersion dans les données ! Le modèle de Poisson n'étant pas adéquat pour ces données de comptage, il est nécessaire de le généraliser. On utilise donc un modèle NB :

$$Y_i|Z = z \sim Poisson(\mu_i z),\ Z \sim \Gamma(\theta_z, \theta_z).$$

On pose
$$\log(\mu_i) = \eta_i$$
et on estime les paramètres avec `R`.

```
> modnb1 <- glm.nb(Species~.,data=gala)
> summary(modnb1)
Call:
glm.nb(formula = Species ~ ., data = gala, init.theta = 1.674602286,
    link = log)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1344  -0.8597  -0.1476   0.4576   1.8416
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.9065247  0.2510344  11.578  < 2e-16 ***
Area        -0.0006336  0.0002865  -2.211 0.027009 *
Elevation    0.0038551  0.0006916   5.574 2.49e-08 ***
Nearest      0.0028264  0.0136618   0.207 0.836100
Scruz       -0.0018976  0.0028096  -0.675 0.499426
Adjacent    -0.0007605  0.0002278  -3.338 0.000842 ***
(Dispersion parameter for Negative Binomial(1.6746) family taken to be 1)
    Null deviance: 88.431  on 29  degrees of freedom
Residual deviance: 33.196  on 24  degrees of freedom
AIC: 304.22
Number of Fisher Scoring iterations: 1
              Theta:  1.675
          Std. Err.:  0.442
 2 x log-likelihood:  -290.223
```

Ici, on observe que plusieurs variables semblent superflues ! On tente d'enlever les variables `Nearest` et `Scruz`. On teste

$$H_0 : \beta^{scruz} = \beta^{nearest} = 0 \text{ versus } H_1 : \beta^{scruz} \neq 0 \text{ ou } \beta^{nearest} \neq 0.$$

On trouve que la statistique du test de rapport de vraisemblance est :

$$2(l_{H_0} - l_{H_1}) = -290.2228 + 290.5934 = 0.3706144.$$

Le nombre de degrés de libertés est 2, et on trouve que

$$P[\chi^2_{(2)} > 0.3706] = 0.830849.$$

On décide donc d'enlever les termes `Scruz` et `Nearest`.

```
> modnb2 <- update(modnb1,~.-Nearest-Scruz)
> anova(modnb1,modnb2)
Likelihood ratio tests of Negative Binomial Models

Response: Species
                                          Model    theta Resid. df
1                   Area + Elevation + Adjacent 1.651523        26
2 Area + Elevation + Nearest + Scruz + Adjacent 1.674602        24
     2 x log-lik.   Test    df  LR stat.  Pr(Chi)
1       -290.5934
2       -290.2228 1 vs 2     2 0.3706144 0.830849
```

```
> summary(modnb2)
Call:
glm.nb(formula = Species ~ Area + Elevation + Adjacent, data = gala,
    init.theta = 1.651523226, link = log)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1251  -0.9963  -0.1226   0.5403   1.6754
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.8149003  0.2231452  12.615  < 2e-16 ***
Area        -0.0006449  0.0002804  -2.300 0.021459 *
Elevation    0.0039299  0.0006761   5.812 6.16e-09 ***
Adjacent    -0.0007943  0.0002196  -3.616 0.000299 ***
(Dispersion parameter for Negative Binomial(1.6515) family taken to be 1)
    Null deviance: 87.279  on 29  degrees of freedom
Residual deviance: 33.155  on 26  degrees of freedom
AIC: 300.59
Number of Fisher Scoring iterations: 1
              Theta:  1.652
          Std. Err.:  0.434
 2 x log-likelihood:  -290.593
```

5.4.1 Poisson gonflée à zéro

Exemple 5.10. ***Nombre de poissons pêchés dans un parc national*** Un biologiste de la faune veut modéliser le nombre de poissons qui sont pêchés dans un parc national. On a des données pour 250 groupes de visiteurs, qui doivent donner les informations suivantes :

persons : le nombre de personnes dans le groupe,

child : le nombre d'enfants dans le groupe,

camper : si les gens sont venus en véhicule motorisé ou non,

count : le nombre de poissons pêchés.

Certaines personnes ne pêchent pas, alors que d'autres n'attrappent aucun poisson. On utilise un modèle Poisson avec lien logarithmique :

$$E[Y_i] = \exp(\beta_0 + \beta_1 \text{persons}_i + \beta_2 \text{child}_i + \beta_3 \text{camper}_i).$$

```
> mod1 <- glm(count~child+camper+persons,family=poisson,data=zinb)
> summary(mod1)
Call: glm(formula = count ~ child + camper + persons,
    family = poisson, data = zinb)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.8096  -1.4431  -0.9060  -0.0406  16.1417
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.98183    0.15226  -13.02   <2e-16 ***
child       -1.68996    0.08099  -20.87   <2e-16 ***
camper       0.93094    0.08909   10.45   <2e-16 ***
persons      1.09126    0.03926   27.80   <2e-16 ***
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 2958.4  on 249  degrees of freedom
Residual deviance: 1337.1  on 246  degrees of freedom
AIC: 1682.1
Number of Fisher Scoring iterations: 6
```

Les données sont surdispersées ! Il y a peut-être trop de zéros, causé par les groupes qui ne pêchent pas. On tente de modéliser ces données ave un modèle de Poisson gonflé à zéro :

$$f_{Y_i}(y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i}, & \text{si } y_i = 0 \\ (1-\pi_i)\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, & \text{si } y_i = 1, 2, 3, ... \end{cases},$$

où π_i représente la probabilité que le groupe ne pêche pas, alors que, sachant que le groupe pêche, le nombre de poissons suit une loi Poisson avec paramètre λ_i. On modélise π_i avec le lien logistique :

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_0 + \gamma_1 \text{persons}_i + \gamma_2 \text{child}_i + \gamma_3 \text{camper}_i.$$

On modélise le paramètre de la Poisson avec le lien log :

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{persons}_i + \beta_2 \text{child}_i + \beta_3 \text{camper}_i.$$

```
> m2 <- zeroinfl(count ~ child + camper + persons, data = zinb)
> summary(m2)
Call:zeroinfl(formula = count ~ child + camper + persons, data = zinb)
Pearson residuals:
     Min       1Q   Median       3Q      Max
-3.05440 -0.74336 -0.44275 -0.07559 27.99301
Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.79826    0.17081  -4.673 2.96e-06 ***
child       -1.13666    0.09299 -12.224  < 2e-16 ***
camper       0.72425    0.09314   7.776 7.51e-15 ***
persons      0.82904    0.04395  18.862  < 2e-16 ***
Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6636     0.5155   3.227  0.00125 **
child         1.9046     0.3261   5.840 5.21e-09 ***
camper       -0.8336     0.3527  -2.364  0.01808 *
persons      -0.9228     0.1992  -4.632 3.62e-06 ***
Number of iterations in BFGS optimization: 14
Log-likelihood: -752.7 on 8 Df

> m3 <- zeroinfl(count ~ child + camper + persons| persons+child, data = zinb)
> summary(m3)
Call: zeroinfl(formula = count ~ child + camper + persons | persons + child,
    data = zinb)
Pearson residuals:
      Min        1Q    Median        3Q       Max
-2.625840 -0.680271 -0.420227 -0.005486 25.589403
Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.84523    0.17220  -4.909 9.18e-07 ***
child       -1.14725    0.09383 -12.227  < 2e-16 ***
camper       0.75973    0.09446   8.043 8.80e-16 ***
persons      0.83374    0.04394  18.976  < 2e-16 ***
Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9813     0.4269   2.298   0.0215 *
persons      -0.8454     0.1919  -4.404 1.06e-05 ***
child         1.8227     0.3200   5.696 1.22e-08 ***
Number of iterations in BFGS optimization: 13
Log-likelihood: -755.5 on 7 Df
```

Avec les modèles liés, on peut utiliser un TRV. On teste

$$H_0 : \gamma_3 = 0 \text{ versus } H_1 : \gamma_3 \neq 0.$$

On trouve que la statistique du test de rapport de vraisemblance est :

$$2(l_{H_0} - l_{H_1}) = 2(-752.7 + 755.5) = 0.3706144.$$

Le nombre de degrés de libertés est 1, et on trouve que

$$P[\chi^2_{(1)} > 5.6] = 0.01796048.$$

On décide donc d'enlever la variable explicative camper du modèle pour π_i.

On peut utiliser les critères AIC ou BIC pour comparer les modèles Poisson et Poisson gonflée à zéro.

- Poisson : AIC = 1682.1
- Poisson gonflée à zéro : AIC $= -2l + 2p = 2 \times 755.5 + 2 \times 7 = 1525$

Selon ce critère, le modèle Poisson gonflée à zéro est préférable.

Bibliographie

Faraway, J. J. (2005). *Extending the linear model with R : generalized linear, mixed effects and nonparametric regression models.* CRC press., http ://www.maths.bath.ac.uk/ jjf23/ELM/.