ACT-2003 Modèles linéaires en actuariat
Exercices
Modélisation de données de comptage

Marie-Pier Côté

Automne 2018

1. (Suite de l'exemple 19 dans les notes de cours). Ajuster un modèle de Poisson avec lien logarithmique au données `esoph` du package `datasets` en R. À partir du modèle avec effets principaux et les interactions de second ordre `agegp+alcgp+tobgp+agegp:alcgp+agegp:tobgp+alcgp:tobgp`, faire une analyse de déviance pour trouver le modèle le plus approprié. Y a-t-il une interaction qui est significative dans le modèle ? Expliquer.

2. Montrer que si $Y|Z = z \sim Poisson(\mu z)$, et $Z \sim Gamma(\theta_z, \theta_z)$, alors $Y \sim BinNeg(\mu, \theta_z)$, soit

$$f(y) = \frac{\Gamma(\theta_z + y)}{\Gamma(\theta_z) y!} \left(\frac{\mu}{\mu + \theta_z}\right)^y \left(\frac{\theta_z}{\mu + \theta_z}\right)^{\theta_z}, y = 0, 1, \ldots$$

3. On suppose que Y_i suit une Poisson avec $g(\mu_i) = \beta_0 + \beta_1 x_i$, où $x_i = 1$, pour $i = 1, \ldots, n_A$ (groupe A) et $x_i = 0$, pour $i = n_A + 1, \ldots, n_A + n_B$ (groupe B). Montrer que, pour toute fonction de lien g continue, l'estimation du GLM par maximum de vraisemblance implique que les moyennes ajustées $\hat{\mu}_A$ et $\hat{\mu}_B$ sont égales aux moyennes empiriques dans l'échantillon.
   Indice : La dérivée de la réciproque d'une fonction continue g est $\frac{1}{g' \circ g^{-1}}$.

4. Dans une expérience, on s'intéresse au taux d'imperfection pour deux procédés utilisés pour fabriquer des plaquettes de silicium dans des puces électroniques. Le traitement A a été appliqué pour dix plaquettes et les nombres d'imperfections sont

$$8, 7, 6, 6, 3, 4, 7, 2, 3, 4.$$

   Le traitement B a été appliqué sur dix autres plaquettes et les nombres d'imperfections sont

$$9, 9, 8, 14, 8, 13, 11, 5, 7, 6.$$

   On traite les données de comptage comme des variables Poisson indépendantes, avec moyennes μ_A et μ_B. [1]

1. Cet exercice est tiré de Agresti (2013).

a) Ajuster le modèle

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

où

$$x_i = \begin{cases} 0, & \text{si traitement A,} \\ 1, & \text{si traitement B.} \end{cases}$$

Montrer que $\exp(\beta_1) = \mu_B/\mu_A$ et interpréter la valeur de l'estimateur du paramètre.

b) Tester $H_0 : \mu_A = \mu_B$ avec le test de Wald. Iterpréter.

c) Construire un intervalle de confiance à 95% pour μ_B/μ_A.

d) Y a-t-il présence de surdispersion ? Expliquer.

e) Ajuster le modèle Binomiale Négative avec lien logarithmique. Que peut-on remarquer ?

f) Ajuster les modèles Poisson et Binomiale Négative aux 20 données sans inclure la variable explicative x. Comparer les résultats et comparer les intervalles de confiance pour la moyenne de la variable réponse. Commenter.

5. Le tableau 1 dénombre les applications aux études graduées à l'Université Berkeley en Californie, pour l'automne 1973. On y voit les décisions d'admission par sexe et par département.

a) Effectuer une régression Poisson avec lien canonique sur le nombre de personnes admises, en utilisant le logarithme du nombre total de personnes qui ont appliqué comme terme offset. Si on utilise seulement le sexe comme variable explicative, est-ce que le sexe a un impact significatif sur le taux d'acceptation ?

b) Si on ajoute le département comme variable explicative dans le modèle en a), est-ce que le sexe a toujours un impact significatif sur le taux d'acceptation ? Qu'est-ce que cela signifie ?

c) Est-ce que l'intéraction entre le sexe et le département est une variable significative dans le modèle ? Que peut-on conclure ?

d) Est-ce que le modèle Poisson est adéquat pour ces données ? Utiliser la déviance et la statistique de Pearson.

e) Refaire les questions a) à c) en utilisant un modèle binomial avec lien logistique, en supposant que m_i est le nombre total de personnes qui ont appliqué.

	Hommes		Femmes	
Département	Admis	Non admis	Admis	Non admis
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

TABLE 1: Données pour l'exercice sur les admissions aux études graduées. Source : P. Bickel et al. (1975). *Science* **187** : 398 − 403.

6. Le fichier de données `MNLung.csv` (séparé avec des virgules) contient des données sur le nombre de décès dûs au cancer du poumon dans 87 régions au Minnesota, pour les hommes et les femmes. L'objectif de l'étude pour laquelle les données ont été recueillies était d'examiner si l'exposition au gaz radon est relié à un changement dans le taux de morbidité standardisé. La base de données contient sept colonnes :

   County : Nom de la région ;

   ID : Numéro d'identifcation de la région dans la base de données ;

   YM : Nombre de décès dûs au cancer du poumon chez les hommes sur une période de 5 ans ;

   EM : Espérance du nombre de cas YM, basé sur des facteurs démographiques ;

   YF : Nombre de décès dûs au cancer du poumon chez les femmes sur une période de 5 ans ;

   EF : Espérance du nombre de cas YF, basé sur des facteurs démographiques ;

   Radon : Mesure moyenne de l'exposition au radon dans chaque région pour la période de 5 ans.

   Le taux de morbidité standardisé (SMR) pour la région i est défini comme

   $$SMR_i = \frac{Y_i}{E_i}.$$

   En utilisant un modèle linéaire généralisé Poisson approprié, répondre aux question suivantes :

   (a) Y a-t-il des preuves dans ces données que le radon est associé avec un changement dans le SMR ?

   (b) Y a-t-il une différence entre le SMR pour les hommes et les femmes, lorsque l'on inclut ou pas la variable explicative radon dans le modèle ?

   (c) Donner les prévisions pour le SMR, avec la mesure d'incertitude, pour des hommes dans une région hypothétique où l'exposition moyenne au radon est de 6 unités.

   (d) Commenter sur la validité du modèle de Poisson pour ces données.

Solutions

1. On ajuste d'abord le modèle avec les effets principaux et les interactions.

```
> library(datasets)
> fit1 <- glm(ncases~factor(agegp)*(factor(alcgp)+factor(tobgp))+factor(alcgp):factor(tobgp),
+family=poisson,data=esoph)
Warning message:
glm.fit: fitted rates numerically 0 occurred
> anova(fit1)
Analysis of Deviance Table

Model: poisson, link: log

Response: ncases

Terms added sequentially (first to last)

                           Df Deviance Resid. Df Resid. Dev
NULL                                          87    262.926
factor(agegp)               5  138.256        82    124.670
factor(alcgp)               3   24.106        79    100.564
factor(tobgp)               3   22.169        76     78.395
factor(agegp):factor(alcgp) 15  32.417        61     45.979
factor(agegp):factor(tobgp) 15  18.109        46     27.870
factor(alcgp):factor(tobgp)  9   7.658        37     20.212
Warning message:
glm.fit: fitted rates numerically 0 occurred
> qchisq(0.95,9) ## rejette alcgp:tobgp
[1] 16.91898
> qchisq(0.95,15) ## rejette agegp:tobgp mais conserve agegp:alcgp
[1] 24.99579
```

On trouve donc que l'interaction entre la consommation d'alcool et de tabac n'est pas significative parce que

$$\Delta Deviance = 7.658 < \chi^2_{(9;0.95)} = 16.92.$$

Cela signifie que le modèle `agegp*(alcgp+tobgp)` est une simplification adéquate du modèle `agegp+alcgp+tobgp+agegp.alcgp+agegp.tobgp+alcgp.tobgp`. De plus, on peut enlever l'interaction entre l'âge et la consommation de tabac :

$$\Delta Deviance = 18.109 < \chi^2_{(15;0.95)} = 25.$$

Cela signifie que le modèle `agegp*alcgp+tobgp` est une simplification adéquate du modèle

`agegp*(alcgp+tobgp)`. Toutefois, on ne peut pas enlever l'autre terme d'interaction car

$$\Delta Deviance = 32.417 > \chi^2_{(15;0.95)} = 25.$$

Si on tente de remettre l'interaction entre la consommation d'alcool et de tabac dans le modèle, on trouve qu'elle n'est toujours pas significative :

```
> fit3 <- update(fit1,~.-factor(agegp):factor(tobgp))
> anova(fit2,fit3)
Analysis of Deviance Table

Model 1: ncases ~ agegp * alcgp + tobgp
Model 2: ncases ~ factor(agegp) + factor(alcgp) + factor(tobgp) + factor(agegp):factor(alcgp) +
    factor(alcgp):factor(tobgp)
  Resid. Df Resid. Dev Df Deviance
1        61     45.979
2        52     38.973  9    7.006
```

Par conséquent, le modèle final est `agegp*alcgp+tobgp`. Cela signifie que l'effet de consommer de l'alcool sur l'occurence du cancer de l'oesophage est différent pour chaque groupe d'âge.

2. Intégrer la densité conditionnelle Poisson sur z. Comme c'est plus agréable à faire à la mais qu'à taper, je vous laisse le soin de réussir par vous-même.

3. On note $n = n_A + n_B$. La vraisemblance pour ce GLM est

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1) &= \prod_{i=1}^{n} \exp(y_i \log(\mu_i) - \mu_i + \text{cte}) \\ &= \prod_{i=1}^{n} \exp(y_i \log(g^{-1}(\eta_i)) - g^{-1}(\eta_i) + \text{cte}).\end{aligned}$$

La log-vraisemblance est donc :

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i \log(g^{-1}(\eta_i)) - g^{-1}(\eta_i) + \text{cte}).$$

On dérive par rapport à β_0 et β_1 :

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

On égalise à 0 pour obtenir le système d'équations à résoudre.

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i) g'(g^{-1}(\hat{\eta}_i))} \\
0 &= \sum_{i=1}^{n} \frac{x_i (y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i) g'(g^{-1}(\hat{\eta}_i))}
\end{aligned}$$

On utilise que $x_i = 0 \forall i \in (n_A + 1, ..., n_A + n_B)$:

$$\begin{aligned}
0 &= \sum_{i=1}^{n_A + n_B} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i) g'(g^{-1}(\hat{\eta}_i))} \\
0 &= \sum_{i=1}^{n_A} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i) g'(g^{-1}(\hat{\eta}_i))} \\
\Rightarrow 0 &= \sum_{i=n_A+1}^{n_A + n_B} \frac{(y_i - g^{-1}(\hat{\eta}_i))}{g^{-1}(\hat{\eta}_i) g'(g^{-1}(\hat{\eta}_i))}.
\end{aligned}$$

Aussi, $\forall i \in (1, ..., n_A), \hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1$, ce qui ne dépend pas de i. Le dénominateur ne dépend pas de i et peut sortir de la somme et s'annuler. De même, $\forall i \in (n_A + 1, ..., n_A + n_B), \hat{\eta}_i = \hat{\beta}_0$, ce qui ne dépend pas de i. Le dénominateur ne dépend pas de i et peut sortir de la somme et s'annuler. On obtient donc les équations :

$$\begin{aligned}
0 &= \sum_{i=1}^{n_A} (y_i - g^{-1}(\hat{\eta}_i)) \\
0 &= \sum_{i=n_A+1}^{n_A + n_B} (y_i - g^{-1}(\hat{\eta}_i)).
\end{aligned}$$

Finalement, $g^{-1}(\hat{\eta}_i) = \hat{\mu}_i$ par définition. Alors

$$0 = \sum_{i=1}^{n_A}(y_i - \hat{\mu}_A) \Rightarrow \sum_{i=1}^{n_A} y_i = n_A\hat{\mu}_A \Rightarrow \frac{\sum_{i=1}^{n_A} y_i}{n_A} = \hat{\mu}_A$$
$$0 = \sum_{i=n_A+1}^{n_A+n_B}(y_i - \hat{\mu}_B) \Rightarrow \sum_{i=n_A+1}^{n_A+n_B} y_i = n_B\hat{\mu}_B \Rightarrow \frac{\sum_{i=n_A+1}^{n_A+n_B} y_i}{n_B} = \hat{\mu}_B.$$

□

4. a) Avec ce modèle, on a que

$$\mu_A = \exp(\beta_0)$$
$$\mu_B = \exp(\beta_0 + \beta_1) = \mu_A \exp(\beta_1),$$

ce qui implique que $\exp(\beta_1) = \mu_B/\mu_A$. On ajuste le modèle en R, et on vérifie que cela est bien vrai :

```
> y <- c( 8,7,6,6,3,4,7,2,3,4,9,9,8,14,8,13,11,5,7,6)
> x <- rep(0:1,each=10)
> fit1 <- glm(y~x,family=poisson)
> summary(fit1)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5280  -0.7622  -0.1699   0.6938   1.5399

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
x             0.5878     0.1764   3.332 0.000861 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27.857  on 19  degrees of freedom
Residual deviance: 16.268  on 18  degrees of freedom
AIC: 94.349

Number of Fisher Scoring iterations: 4
> log(mean(y[which(x==1)])/mean(y[which(x==0)]))
[1] 0.5877867
```

b) Puisque $\exp(\beta_1) = \mu_B/\mu_A$, alors si $H_0 : \mu_A = \mu_B$ est vraie, $\beta_1 = 0$. On peut utiliser la statistique de Wald directement, on trouve que le seuil observé du test est 0.000861. On rejette donc l'hypothèse nulle à un niveau de confiance de 99%, ce qui implique que les moyennes diffèrent de façon significative.

c) Un I.C. à 95% pour β_1 est

```
> fit1$coef[2]+c(-1,1)*qnorm(0.975)*summary(fit1)$coefficients[2,2]
[1] 0.2420820 0.9334913
```

Alors, un I.C. pour μ_B/μ_A est $(\exp(0.2421), \exp(0.93349)) = (1.273899, 2.543373)$.

d) Il n'y a pas d'indications de surdispersion, puisque la déviance est 16.26 sur 18 degrés de liberté, et $16.26/18 < 1$.

e) Quand on ajuste une binomiale négative à ces données, on trouve que θ_z tend vers l'infini, donc le modèle Poisson est une simplification adéquate du modèle NB. En fait, les estimations des paramètres β_0 et β_1 sont exactement les mêmes que celles obtenues dans le modèle Poisson.

```
> library(MASS)
> fit2 <- glm.nb(y~x)
Warning messages:
1: In theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace >  :
  iteration limit reached
2: In theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace >  :
  iteration limit reached
> summary(fit2)

Call:
glm.nb(formula = y ~ x, init.theta = 113420.6921, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5280  -0.7622  -0.1699   0.6937   1.5398

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
x             0.5878     0.1764   3.332 0.000861 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(113420.7) family taken to be 1)

    Null deviance: 27.855  on 19  degrees of freedom
Residual deviance: 16.267  on 18  degrees of freedom
AIC: 96.349
```

```
Number of Fisher Scoring iterations: 1


              Theta:  113421
          Std. Err.:  4076959
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -90.349
```

f) Dans ce cas, on remarque que, bien que l'estimation du paramètre est égale pour les deux modèles, l'écart-type diffère. Aussi, le modèle de Poisson ne semble plus adéquat, car $Deviance/dl = 27.857/19 > 1$, alors que le modèle NB s'ajuste bien aux données. Cela montre que lorsqu'une variable explicative importante n'est pas observée, le modèle de Poisson peut perdre sa validité pour des données de comptage. La variable explicative manquante introduit de la sur-dispersion dans les données, ce qui est capturé efficacement avec la loi NB.

```
> fit3 <- glm(y~1,family=poisson)
> fit4 <- glm.nb(y~1)
> summary(fit3)

Call:
glm(formula = y ~ 1, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2336  -0.9063   0.0000   0.4580   2.3255

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.94591    0.08451   23.02   <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: 27.857  on 19  degrees of freedom
Residual deviance: 27.857  on 19  degrees of freedom
AIC: 103.94

Number of Fisher Scoring iterations: 4

> summary(fit4)

Call:
glm.nb(formula = y ~ 1, init.theta = 18.2073559, link = log)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9810  -0.7836   0.0000   0.3859   1.9033

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.94591    0.09944   19.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(18.2074) family taken to be 1)

    Null deviance: 20.279  on 19  degrees of freedom
Residual deviance: 20.279  on 19  degrees of freedom
AIC: 104.77

Number of Fisher Scoring iterations: 1


              Theta:  18.2
          Std. Err.:  21.0

 2 x log-likelihood:  -100.767
> exp(fit3$coef[1]+c(-1,1)*qnorm(0.975)*summary(fit3)$coefficients[1,2])
[1] 5.931421 8.261090
> exp(fit4$coef[1]+c(-1,1)*qnorm(0.975)*summary(fit4)$coefficients[1,2])
[1] 5.760386 8.506374
```

5. a) On y va

```
> sex <- rep(0:1,each=6)
> Dep <- rep(0:5,2)
> y <- c(512,353,120,138,53,22,89,17,202,131,94,24)
> no <- c(313,207,205,279,138,351,19,8,391,244,299,317)
> nb <- y+no
> fitpSex <- glm(y~factor(sex)+offset(log(nb)),family=poisson)
> summary(fitpSex)

Call:
glm(formula = y ~ factor(sex) + offset(log(nb)), family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-14.1129   -3.6826   -0.2719    3.7437    8.0834
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.80926    0.02889 -28.011  < 2e-16 ***
factor(sex)1 -0.38298    0.05128  -7.468 8.15e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 551.69  on 11  degrees of freedom
Residual deviance: 493.56  on 10  degrees of freedom
AIC: 573.76

Number of Fisher Scoring iterations: 4
```

On trouve donc que la valeur-p du test de Wald $H_0 : \beta^{SEX} = 0$ est 8.15×10^{-14} ce qui est hautement significatif. Puisque le coefficient est négatif et que le niveau de base utilisé est "hommes", cela signifie que les femmes ont moins de chance d'être acceptées aux études graduées que les hommes.

b) On ajoute le département :

```
> fitp2 <- glm(y~factor(sex)+factor(Dep)+offset(log(nb)),family=poisson)
> summary(fitp2)

Call:
glm(formula = y ~ factor(sex) + factor(Dep) + offset(log(nb)),
    family = poisson)

Deviance Residuals:
       1         2         3         4         5         6         7         8
-0.68882  -0.01474   0.96655   0.02569   0.97713  -0.28371   1.77895   0.06756
       9        10        11        12
-0.71131  -0.02632  -0.68503   0.28254

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.44677    0.04148 -10.771   <2e-16 ***
factor(sex)1  0.05859    0.06166   0.950    0.342
factor(Dep)1 -0.01391    0.06625  -0.210    0.834
factor(Dep)2 -0.63911    0.07660  -8.344   <2e-16 ***
factor(Dep)3 -0.66125    0.07675  -8.615   <2e-16 ***
factor(Dep)4 -0.97250    0.09836  -9.887   <2e-16 ***
factor(Dep)5 -2.32388    0.15468 -15.024   <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: 551.6926  on 11  degrees of freedom
Residual deviance:   6.6698  on  5  degrees of freedom
AIC: 96.868

Number of Fisher Scoring iterations: 3
```

Dans ce modèle, le résultat du test de Wald pour le coefficient de la variable Sexe est différent. Puisque le seuil observé du test est 34.5%, on ne peut pas rejeter l'hypothèse nulle que $\beta^{SEX} = 0$. Cela signifie que le sexe n'est pas un facteur qui influence le taux d'admission aux études graduées lorsqu'on prend en considération le département. Il en est ainsi car les femmes appliquent plus souvent que les hommes dans des départements où il est plus difficile d'être admis.

c) À l'aide de l'analyse de la déviance, on trouve que l'interaction n'est pas significative :

$$\Delta Deviance = 6.67 < \chi^2(0.95, 5) = 11.07.$$

```
> fitp <- glm(y~factor(sex)*factor(Dep)+offset(log(nb)),family=poisson)
> anova(fitp)
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


                      Df Deviance Resid. Df Resid. Dev
NULL                                     11     551.69
factor(sex)            1    58.13        10     493.56
factor(Dep)            5   486.89         5       6.67
factor(sex):factor(Dep)  5     6.67         0       0.00
> qchisq(0.95,5) ## reject interaction
[1] 11.0705
```

d) Le modèle final est celui avec une seule variable explicative dichotomique : le département. La déviance pour ce modèle est 7.5706, ce qui est légèrement supérieur à 6, le nombre de degrés de liberté. Toutefois, puisque $Deviance/dl \approx 1.26$, cela n'est pas très alarmant, et il n'y a pas de raison de supposer que le modèle de Poisson est inadéquat. La statistique de Pearson est 8.03, ce qui est aussi une valeur attendue pour la loi chi-carrée avec 6 degrés de liberté.

```
> fitpDep <- glm(y~factor(Dep)+offset(log(nb)),family=poisson)
> summary(fitpDep)
```

```
Call:
glm(formula = y ~ factor(Dep) + offset(log(nb)), family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8481  -0.4187   0.1160   0.4595   2.2321

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.43981    0.04079 -10.782   <2e-16 ***
factor(Dep)1 -0.01830    0.06608  -0.277    0.782
factor(Dep)2 -0.60784    0.06906  -8.801   <2e-16 ***
factor(Dep)3 -0.64004    0.07336  -8.725   <2e-16 ***
factor(Dep)4 -0.93966    0.09201 -10.212   <2e-16 ***
factor(Dep)5 -2.30243    0.15298 -15.050   <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: 551.6926  on 11  degrees of freedom
Residual deviance:   7.5706  on  6  degrees of freedom
AIC: 95.769

Number of Fisher Scoring iterations: 4
> sum((y-fitted(fitpDep))^2/fitted(fitpDep))
[1] 8.025236
> pchisq(8.025236,6)
[1] 0.7637397
```

e) On recommence et on obtient exactement les mêmes conclusions :

```
> fitbSex <- glm(cbind(y,nb-y)~factor(sex),family=binomial)
> summary(fitbSex)

Call:
glm(formula = cbind(y, nb - y) ~ factor(sex), family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-16.7915   -4.7613   -0.4365    5.1025   11.2022

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  -0.22013    0.03879  -5.675 1.38e-08 ***
factor(sex)1 -0.61035    0.06389  -9.553  < 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: 877.06  on 11  degrees of freedom
Residual deviance: 783.61  on 10  degrees of freedom
AIC: 856.55

Number of Fisher Scoring iterations: 4
> fitb2 <- glm(cbind(y,nb-y)~factor(sex)+factor(Dep),family=binomial)
> summary(fitb2)

Call:
glm(formula = cbind(y, nb - y) ~ factor(sex) + factor(Dep), family = binomial)

Deviance Residuals:
      1        2        3        4        5        6        7        8
-1.2487  -0.0560   1.2533   0.0826   1.2205  -0.2076   3.7189   0.2706
      9       10       11       12
-0.9243  -0.0858  -0.8509   0.2052

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.58205    0.06899   8.436   <2e-16 ***
factor(sex)1  0.09987    0.08085   1.235    0.217
factor(Dep)1 -0.04340    0.10984  -0.395    0.693
factor(Dep)2 -1.26260    0.10663 -11.841   <2e-16 ***
factor(Dep)3 -1.29461    0.10582 -12.234   <2e-16 ***
factor(Dep)4 -1.73931    0.12611 -13.792   <2e-16 ***
factor(Dep)5 -3.30648    0.16998 -19.452   <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: 877.056  on 11  degrees of freedom
Residual deviance:  20.204  on  5  degrees of freedom
AIC: 103.14

Number of Fisher Scoring iterations: 4
> fitb <- glm(cbind(y,nb-y)~factor(sex)*factor(Dep),family=binomial)
> anova(fitb)
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(y, nb - y)

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev
NULL                                       11     877.06
factor(sex)              1    93.45        10     783.61
factor(Dep)              5   763.40         5      20.20
factor(sex):factor(Dep)  5    20.20         0       0.00
> qchisq(0.95,5)
[1] 11.0705
```

6. Cette solution est en anglais, si vous avez des questions, vous pouvez écrire sur le forum.
   If $Y_i \sim Poisson(E_i\lambda_i)$, then, using the canonical link,

$$\log(\mu_i) = log(E_i) + log(\lambda_i),$$

   where λ_i is the mean SMR for observation i. $\log(E_i)$, the natural logarithm of the expected count of lung cancer based on the demographics of the county, is passed to the `glm` function as an offset factor.

   The data for males and females are concatenated to create a model with one covariate, Radon exposure, and one factor predictor, Sex, which takes 2 levels (0 for males and 1 for females).

```
> Ytot <- c(YM,YF)
> Etot <- c(EM,EF)
> Sex <- c(rep(0,length(YM)),rep(1,length(YF))) ## 1 if female
> Radontot <- rep(Radon,2)

> modsex <- glm(Ytot~Sex+offset(log(Etot)),family=poisson)
> modsexrad <- glm(Ytot~Radontot+Sex+offset(log(Etot)),family=poisson)

> anova(modsex,modsexrad)
Analysis of Deviance Table

Model 1: Ytot ~ Sex + offset(log(Etot))
Model 2: Ytot ~ Radontot + Sex + offset(log(Etot))
  Resid. Df Resid. Dev Df Deviance
1       172     410.27
2       171     364.05  1    46.22
```

```
> qchisq(0.99,1)
[1] 6.634897
```

As shown above, the analysis of deviance shows strong evidence that the radon exposure influences the number of lung cancer in a particular county :

$$\Delta Deviance = 46.22 > \chi^2_{(1;0.99)} = 6.6349.$$

b) The null model and the model including sex only are fitted.

```
> modtot <- glm(Ytot~1+offset(log(Etot)),family=poisson)
> modsex <- glm(Ytot~Sex+offset(log(Etot)),family=poisson)
> anova(modtot,modsex)
Analysis of Deviance Table

Model 1: Ytot ~ 1 + offset(log(Etot))
Model 2: Ytot ~ Sex + offset(log(Etot))
  Resid. Df Resid. Dev Df  Deviance
1       173     410.28
2       172     410.27  1 0.0093398
> qchisq(0.95,1)
[1] 3.841459
```

The analysis of deviance shows that $\Delta Deviance = 0.0093398 < \chi^2_{(1;0.95)} = 3.8415$. Hence, the null model is an appropriate simplification of the model including the factor Sex, so the factor is not significant. However, below is the R output for the analysis of deviance when the covariate Radon (known to be significant from a) is included in the model. If we first consider the model with main effects and interactions, we see that $\Delta Deviance = 8.823 > \chi^2_{(1;0.99)}$, meaning that the model with main effects only is not an adequate simplification of the model with main effects and interactions. Thus, the factor predictor Sex is significant in the model through its interaction with the covariate Radon. Note that even if the main effect of the Sex does not appear to be significant, it is kept in the model by convention.

```
> modtotrad <- glm(Ytot~Radontot+offset(log(Etot)),family=poisson)
> modsexrad <- glm(Ytot~Radontot+Sex+offset(log(Etot)),family=poisson)
> modsexradINT <- glm(Ytot~Radontot*Sex+offset(log(Etot)),family=poisson)
> anova(modtot,modtotrad,modsexrad,modsexradINT)
Analysis of Deviance Table

Model 1: Ytot ~ 1 + offset(log(Etot))
Model 2: Ytot ~ Radontot + offset(log(Etot))
Model 3: Ytot ~ Radontot + Sex + offset(log(Etot))
Model 4: Ytot ~ Radontot * Sex + offset(log(Etot))
  Resid. Df Resid. Dev Df Deviance
1       173     410.28
```

```
2       172     364.06  1   46.219
3       171     364.05  1    0.011
4       170     355.23  1    8.823
```

c) The predictions are obtained using the command

```
predict(modsexradINT,data.frame(Radontot=6,Sex=0,Etot=1),type="response",se.fit=TRUE)
```

If the model Sex*Radon is used, we find

$$\hat{SMR}_{Sex=0,Radon=6} = 0.9708183,$$

with a standard error of 0.01415307.

d) The model Sex*Radon has a deviance of 355.23 on 170 degrees of freedom. A heuristic check for the validity of the model is to calculate the estimated dispersion parameter

$$\hat{\phi} = \frac{355.23}{170} = 2.089$$

and to compare it with 1, the dispersion parameter implied in the Poisson model. This check suggests the presence of overdispersion in the data as $\hat{\phi}$ is greater than 1. Fitting the quasipoisson model also leads to the same conclusion : the estimated dispersion parameter is 1.98311, closer to 2. Thus, we can conclude that the Poisson model is not adequate, we might consider fitting a Negative Binomial model to capture the overdispersion.