Laboratori del corso di Statistica (cp)

3. Modelli per Tabelle di Frequenze

Leonardo Egidi

AA 2021/2022

Tutti i dati per questa esercitazione sono nell'archivio `glm.Rdata`, che carichiamo con il comando

```
load("glm.Rdata")
```

1 Dati sul cancro della pelle

```
skin
```

```
##    Counts  Type  Site
## 1      22 Melan HeadN
## 2      16 Super HeadN
## 3      19 Nodul HeadN
## 4      11 Indet HeadN
## 5       2 Melan Trunk
## 6      54 Super Trunk
## 7      33 Nodul Trunk
## 8      17 Indet Trunk
## 9      10 Melan Extrm
## 10    115 Super Extrm
## 11     73 Nodul Extrm
## 12     28 Indet Extrm
```

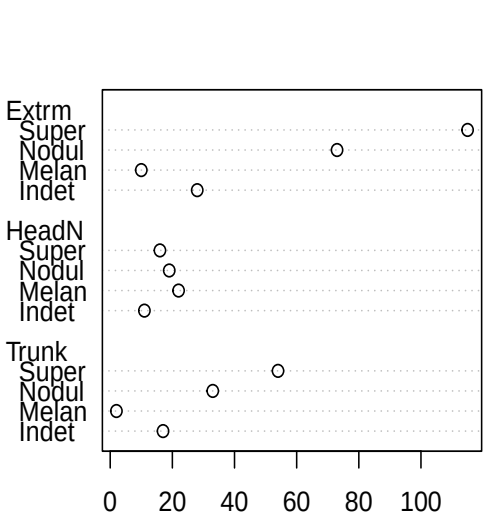
Si tratta di una tabella a doppia entrata con le due variabili tipo di tumore `Type` e sito del medesimo `Site`, possiamo riorganizzare i dati in forma tabellare con il comando `xtabs`.

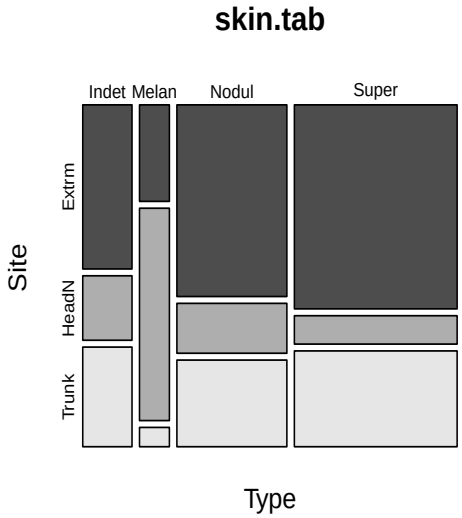
```
skin.tab=xtabs(skin,formula=Counts~Type+Site)
skin.tab
```

```
##        Site
## Type    Extrm HeadN Trunk
##   Indet    28    11    17
##   Melan    10    22     2
##   Nodul    73    19    33
##   Super   115    16    54
```

I dati possono essere rappresentati graficamente con un diagramma a punti o un mosaico

```
dotchart(skin.tab)
mosaicplot(skin.tab,color=TRUE)
```



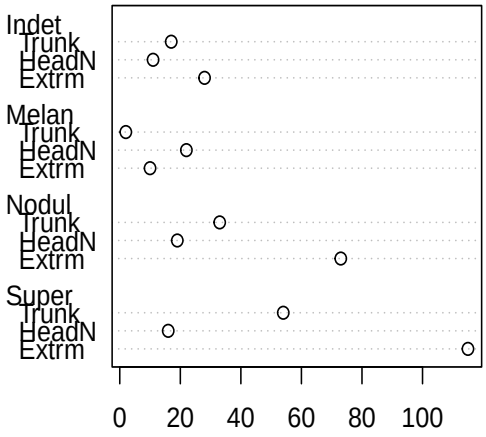


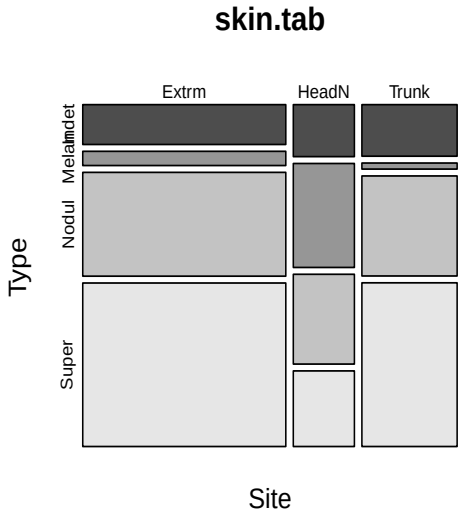
Notiamo che si possono cambiare i ruoli delle due variabili

```
skin.tab=xtabs(skin,formula=Counts~Site+Type)
skin.tab

##        Type
## Site    Indet Melan Nodul Super
##   Extrm    28    10    73   115
##   HeadN    11    22    19    16
##   Trunk    17     2    33    54

par(mfrow=c(1,2))
dotchart(skin.tab)
mosaicplot(skin.tab,color=TRUE)
```





Stimando un modello con effetti principali si può valutare l'ipotesi di indipendendenza tra le due variabili.

```
skin.fit <- glm(Counts~Type+Site,family=poisson,data=skin)
skin.fit

##
## Call:  glm(formula = Counts ~ Type + Site, family = poisson, data = skin)
##
## Coefficients:
## (Intercept)    TypeMelan    TypeNodul    TypeSuper    SiteHeadN
##      3.4544      -0.4990       0.8030       1.1950      -1.2010
##   SiteTrunk
##     -0.7571
##
## Degrees of Freedom: 11 Total (i.e. Null);  6 Residual
## Null Deviance:     295.2
## Residual Deviance: 51.8  AIC: 122.9

summary(skin.fit)

##
## Call:
## glm(formula = Counts ~ Type + Site, family = poisson, data = skin)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -3.0453  -1.0741   0.1297   0.5857   5.1354
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.4544     0.1406  24.561  < 2e-16 ***
## TypeMelan    -0.4990     0.2174  -2.295   0.0217 *
## TypeNodul     0.8030     0.1608   4.993 5.93e-07 ***
## TypeSuper     1.1950     0.1525   7.835 4.69e-15 ***
## SiteHeadN    -1.2010     0.1383  -8.683  < 2e-16 ***
## SiteTrunk    -0.7571     0.1177  -6.431 1.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 295.203  on 11  degrees of freedom
## Residual deviance:  51.795  on  6  degrees of freedom
## AIC: 122.91
##
## Number of Fisher Scoring iterations: 5
```

Il fatto che i coefficienti siano significativamente diversi da zero indica semplicemente che non è uniforme la distribuzione dei tumori né per tipo né per sito: ad esempio sono più frequenti quelli di tipo `Super` rispetto agli altri tipi, e sono più frequenti quelli alle estremità che quelli alla testa o al tronco.
Il modello così costruito assume vi sia indipendenza tra tipo e sito; un indicatore di quanto questa ipotesi sia compatibile con i dati è nella bontà o meno dell'adattamento. Notiamo allora che la devianza è pari a 52 con 6 g.d.l., un valore troppo alto, che corrisponde a un valore p per l'ipotesi di indipendenza pari a

```
1-pchisq(51.8,6)
```

```
## [1] 2.045725e-09
```

Ricordiamo che qui possiamo ricorrere a questo tipo di test (devianza diviso coefficiente di dispersione, in questo caso $\phi = 1$) siccome stiamo considerando un modello di Poisson con soli fattori come covariate. Dobbiamo quindi concludere che le variabili `Type` e `Site` non sono indipendenti. Avremmo potuto fare un test simile tramite la statistica χ^2 di Pearson.

```
chi2Pearson=sum((skin$Counts-fitted(skin.fit))^2/fitted(skin.fit))
chi2Pearson
```

```
## [1] 65.81293
```

```
1-pchisq(chi2Pearson,6)
```

```
## [1] 2.943201e-12
```

che conduce alla stessa conclusione, seppure i valori delle funzioni test non siano uguali.
Infine si noti che il modello con interazioni produce il modello saturo (senza gradi di libertà):

```
fitSaturo=glm(Counts~Type*Site,poisson,skin)
summary(fitSaturo)
```

```
##
## Call:
## glm(formula = Counts ~ Type * Site, family = poisson, data = skin)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           3.3322     0.1890  17.632  < 2e-16 ***
## TypeMelan            -1.0296     0.3684  -2.795 0.005192 **
## TypeNodul             0.9583     0.2223   4.311 1.63e-05 ***
## TypeSuper             1.4127     0.2107   6.704 2.03e-11 ***
## SiteHeadN            -0.9343     0.3558  -2.626 0.008649 **
## SiteTrunk            -0.4990     0.3075  -1.623 0.104612
## TypeMelan:SiteHeadN   1.7228     0.5216   3.303 0.000957 ***
## TypeNodul:SiteHeadN  -0.4117     0.4393  -0.937 0.348618
## TypeSuper:SiteHeadN  -1.0380     0.4448  -2.334 0.019602 *
## TypeMelan:SiteTrunk  -1.1104     0.8334  -1.332 0.182713
## TypeNodul:SiteTrunk  -0.2950     0.3722  -0.792 0.428092
## TypeSuper:SiteTrunk  -0.2570     0.3489  -0.736 0.461479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  2.9520e+02  on 11  degrees of freedom
## Residual deviance: -1.1102e-15  on  0  degrees of freedom
## AIC: 83.111
##
## Number of Fisher Scoring iterations: 3
```

Nel modello saturo l'adattamento è perfetto, di conseguenza la devianza è nulla.

```
predict(fitSaturo,type="response")
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12
##  22  16  19  11   2  54  33  17  10 115  73  28
```

```
skin$Counts
```

```
##  [1]  22  16  19  11   2  54  33  17  10 115  73  28
```

2 Neonati e cure antenatali

I dati `babies` riguardano la sopravvivenza di bimbi nati in due diverse cliniche da madri con due diversi tipi di cura antenatale.

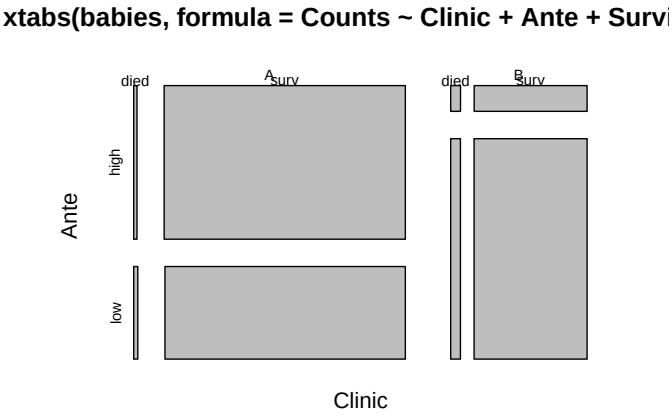
```
babies
```

```
##   Counts Clinic Ante Survival
## 1    176      A  low     surv
## 2    293      A high     surv
## 3    197      B  low     surv
## 4     23      B high     surv
## 5      3      A  low     died
## 6      4      A high     died
## 7     17      B  low     died
## 8      2      B high     died
```

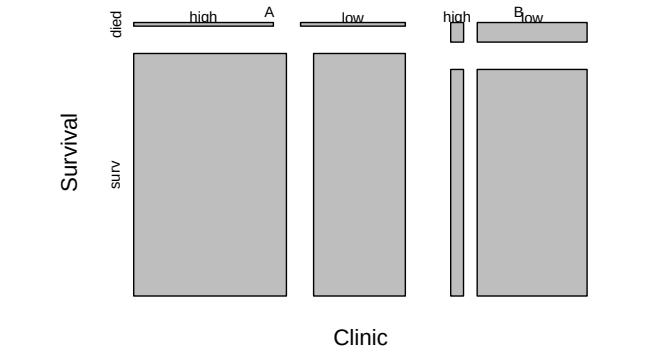
```
ftable(xtabs(babies,formula=Counts~Clinic+Ante+Survival))
```

```
##             Survival died surv
## Clinic Ante
## A      high             4  293
##        low              3  176
## B      high             2   23
##        low             17  197
```

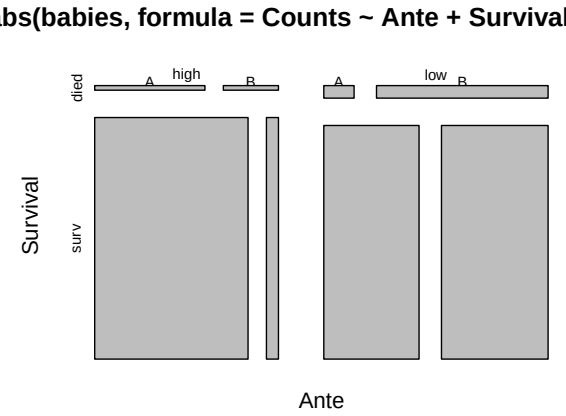
```
plot(xtabs(babies,formula=Counts~Clinic+Ante+Survival))
```




```
plot(xtabs(babies,formula=Counts~Clinic+Survival+Ante))
```




```
plot(xtabs(babies,formula=Counts~Ante+Survival+Clinic))
```



Stimiamo un modello con i soli effetti principali:

```
babies.fit <- glm(Counts~Clinic+Ante+Survival,poisson,data=babies)
babies.fit

##
## Call:  glm(formula = Counts ~ Clinic + Ante + Survival, family = poisson,
##     data = babies)
##
## Coefficients:
##  (Intercept)       ClinicB       Antelow  Survivalsurv
##       2.0535       -0.6890        0.1993        3.2771
##
## Degrees of Freedom: 7 Total (i.e. Null);  4 Residual
## Null Deviance:     1066
## Residual Deviance: 211.5  AIC: 259.7
```

Chiaramente il modello non va bene. Comunque, dovremmo includere i termini che sono fissati dal disegno: in questo caso è probabile che le frequenze di `Clinic`×`Ante` fossero fissate, quindi un modello più appropriato è

```
babies.fit2 <- glm(Counts~Clinic*Ante+Survival,poisson,data=babies)
babies.fit2

##
## Call:  glm(formula = Counts ~ Clinic * Ante + Survival, family = poisson,
##     data = babies)
##
## Coefficients:
##     (Intercept)          ClinicB          Antelow     Survivalsurv
##          2.3795          -2.4749          -0.5063           3.2771
```

```
## ClinicB:Antelow
##          2.6534
##
## Degrees of Freedom: 7 Total (i.e. Null);  3 Residual
## Null Deviance:     1066
## Residual Deviance: 17.83  AIC: 68.01
```

Però, anche in questo caso la devianza residua è grande rispetto i gradi di libertà.
Stimiamo allora il modello saturo per confrontare i contributi alla devianza.

```
babies.fit3 <- glm(Counts~Clinic*Ante*Survival,poisson,data=babies)
anova(babies.fit3)

## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Counts
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev
## NULL                                    7    1066.43
## Clinic               1    80.06         6     986.36
## Ante                 1     7.06         5     979.30
## Survival             1   767.82         4     211.48
## Clinic:Ante          1   193.65         3      17.83
## Clinic:Survival      1    17.75         2       0.08
## Ante:Survival        1     0.04         1       0.04
## Clinic:Ante:Survival 1     0.04         0       0.00
```

L'analisi della devianza suggerisce che le interazioni che includono sia `Ante` che `Survival` non siano necessarie.
Semplifichiamo allora il modello

```
babies.fit4 <- glm(Counts~Clinic*Ante + Clinic*Survival,poisson,babies)
babies.fit4

##
## Call:  glm(formula = Counts ~ Clinic * Ante + Clinic * Survival, family = poisson,
##     data = babies)
##
## Coefficients:
##          (Intercept)              ClinicB              Antelow
```

```
##               1.4742               -0.7874               -0.5063
##         Survivalsurv       ClinicB:Antelow  ClinicB:Survivalsurv
##               4.2047                2.6534               -1.7555
##
## Degrees of Freedom: 7 Total (i.e. Null);  2 Residual
## Null Deviance:     1066
## Residual Deviance: 0.08229  AIC: 52.26
```

Con questo modello la devianza residua è piccola rispetto ai gradi di libertà. Il modello implica che la sopravvivenza dipende dalla clinica ma non dal tipo di cura.
Si parla in questo caso di indipendenza condizionata: cura e sopravvivenza sono indipendenti condizionatamente alla clinica.
Possiamo confermare questa circostanza anche guardando semplicemente alla tabella dei dati

```
tabbabies=xtabs(Counts~Ante+Survival+Clinic,babies)
ftable(tabbabies)

##               Clinic   A   B
## Ante Survival
## high died              4   2
##      surv            293  23
## low  died              3  17
##      surv            176 197
```

Otteniamo le probabilità di sopravvivenza condizionatamente alla clinica e per i due livelli di trattamento antenatale con

```
prop.table(tabbabies,c(1,3))

## , , Clinic = A
##
##       Survival
## Ante         died       surv
##   high 0.01346801 0.98653199
##   low  0.01675978 0.98324022
##
## , , Clinic = B
##
##       Survival
## Ante         died       surv
##   high 0.08000000 0.92000000
##   low  0.07943925 0.92056075
```

Più formalmente possiamo testare l'ipotesi di indipendenza per ciascuna delle tabelle a due vie che si ottengono condizionandosi rispetto alla clinica

```
xtabs(Counts~Ante+Survival,babies,subset=Clinic=="A")
```

```
##       Survival
## Ante   died surv
##   high    4  293
##   low     3  176
```

```
summary(xtabs(Counts~Ante+Survival,babies,subset=Clinic=="A"))
```

```
## Call: xtabs(formula = Counts ~ Ante + Survival, data = babies, subset = Clinic ==
##     "A")
## Number of cases in table: 476
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.08352, df = 1, p-value = 0.7726
##  Chi-squared approximation may be incorrect
```

```
xtabs(Counts~Ante+Survival,babies,subset=Clinic=="B")
```

```
##       Survival
## Ante   died surv
##   high    2   23
##   low    17  197
```

```
summary(xtabs(Counts~Ante+Survival,babies,subset=Clinic=="B"))
```

```
## Call: xtabs(formula = Counts ~ Ante + Survival, data = babies, subset = Clinic ==
##     "B")
## Number of cases in table: 239
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 9.619e-05, df = 1, p-value = 0.9922
##  Chi-squared approximation may be incorrect
```

In un modo o nell'altro i conteggi non mostrano scostamenti significativi dall'indipendenza.
Per vedere che il modello sopra implica l'indipendenza condizionata (esatta) si scriva il modello nella forma

$$\log \mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ik} + \epsilon_{jk}$$

dove col triplo indice si indica la frequenza di osservazioni nelle classi determinate dalle variabili esplicative, in particolare

- l'indice i indica la cura prenatale, $i = 1$ per `high`, $i = 2$ per `low`;
- l'indice j indica la sopravvivenza, $j = 1$ per `died`, $j = 2$ per `surv`;
- l'indice k indica la clinica, $k = 1$ per `A`, $k = 2$ per `B`;

alla specificazione sopra vanno aggiunti i vincoli

$$\alpha_1 = \beta_1 = \gamma_1 = 0$$

e

$$\delta_{1k} = \delta_{i1} = \epsilon_{1k} = \epsilon_{j1} = 0 \ \ \forall i, j, k$$

In pratica sono non nulli solo α_2, β_2, γ_2, δ_{22} e ϵ_{22} che corrispondono ai coefficienti stimati in `babies.fit4`.
Con ciò condizionatamente a `Clinic` i logaritmi delle frequenze attese della tabella a doppia entrata per sopravvivenza e cura (`Survival` e *Ante*) sono

`Clinic=A`	`died`	`surv`
`high`	μ	$\mu + \beta_2$
`low`	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_2$

`Clinic=B`	`died`	`surv`
`high`	$\mu + \gamma_2$	$\mu + \beta_2 + \gamma_2 + \epsilon_{22}$
`low`	$\mu + \alpha_2 + \gamma_2 + \delta_{22}$	$\mu + \alpha_2 + \beta_2 + \gamma_2 + \delta_{22} + \epsilon_{22}$

e pertanto gli *odds-ratio* della sopravvivenza rispetto la cura condizionatamente alla clinica sono

$$OR_{AS|Clinic=A} = \frac{\frac{\exp\{\mu+\beta_2\}}{\exp\{\mu\}}}{\frac{\exp\{\mu+\alpha_2+\beta_2\}}{\exp\{\mu+\alpha_2\}}} = 1$$

$$OR_{AS|Clinic=B} = \ldots = 1$$

che corrisponde appunto all'indipendenza condizionata[1].

[1] Alternativamente, il modello può essere scritto nella forma

$$\log \mu_i = \lambda_0 + \lambda_A x_{iA} + \lambda_S x_{iS} + \lambda_C x_{iC} + \lambda_{AC} x_{iAC} + \lambda_{CS} x_{iCS}$$

dove

$$x_{iA} = \begin{cases} 1 & \text{se } \texttt{Ante=low} \\ 0 & \text{altrimenti} \end{cases}, \quad x_{iS} = \begin{cases} 1 & \text{se } \texttt{Survival=surv} \\ 0 & \text{altrimenti} \end{cases}, \quad x_{iC} = \begin{cases} 1 & \text{se } \texttt{Clinic=B} \\ 0 & \text{altrimenti} \end{cases}$$

$$x_{iAC} = \begin{cases} 1 & \text{se } \texttt{Ante=low} \text{ e } \texttt{Clinic=B} \\ 0 & \text{altrimenti} \end{cases}, \quad x_{iCS} = \begin{cases} 1 & \text{se } \texttt{Survival=surv} \text{ e } \texttt{Clinic=B} \\ 0 & \text{altrimenti} \end{cases}$$

condizionatamente a `Clinic` i logaritmi delle frequenze attese della tabella a doppia entrata per sopravvivenza e cura (`Survival` e *Ante*) sono

`Clinic=A`	`died`	`surv`
`low`	$\lambda_0 + \lambda_A$	$\lambda_0 + \lambda_A + \lambda_S$
`high`	λ_0	$\lambda_0 + \lambda_S$

`Clinic=B`	`died`	`surv`
`low`	$\lambda_0 + \lambda_C + \lambda_A + \lambda_{AC}$	$\lambda_0 + \lambda_C + \lambda_A + \lambda_S + \lambda_{AC} + \lambda_{CS}$
`high`	$\lambda_0 + \lambda_C$	$\lambda_0 + \lambda_C + \lambda_S + \lambda_{CS}$

3 Donne fumatrici

Nell'ambito di uno studio ventennale, sono state osservate alcune donne rispetto al fatto che siano fumatrici e alla sopravvivenza (per i venti anni di durata dello studio).

```
library(faraway)

## Warning:  package 'faraway' was built under R version 4.1.2

data(femsmoke)
femsmoke

##       y smoker dead   age
## 1     2    yes  yes 18-24
## 2     1     no  yes 18-24
## 3     3    yes  yes 25-34
## 4     5     no  yes 25-34
## 5    14    yes  yes 35-44
## 6     7     no  yes 35-44
## 7    27    yes  yes 45-54
## 8    12     no  yes 45-54
## 9    51    yes  yes 55-64
## 10   40     no  yes 55-64
## 11   29    yes  yes 65-74
## 12  101     no  yes 65-74
## 13   13    yes  yes   75+
## 14   64     no  yes   75+
## 15   53    yes   no 18-24
## 16   61     no   no 18-24
## 17  121    yes   no 25-34
## 18  152     no   no 25-34
## 19   95    yes   no 35-44
## 20  114     no   no 35-44
## 21  103    yes   no 45-54
## 22   66     no   no 45-54
## 23   64    yes   no 55-64
## 24   81     no   no 55-64
```

e pertanto gli *odds-ratio* della sopravvivenza rispetto la cura sono

$$OR_{AS|Clinic=A} = \frac{\frac{\exp\{\lambda_0+\lambda_S\}}{\exp\{\lambda_0\}}}{\frac{\exp\{\lambda_0+\lambda_A+\lambda_S\}}{\exp\{\lambda_0+\lambda_A\}}} = 1$$

$$OR_{AS|Clinic=B} = \frac{\frac{\exp\{\lambda_0+\lambda_C+\lambda_S+\lambda_{CS}\}}{\exp\{\lambda_0+\lambda_C\}}}{\frac{\exp\{\lambda_0+\lambda_C+\lambda_A+\lambda_S+\lambda_{AC}+\lambda_{CS}\}}{\exp\{\lambda_0+\lambda_C+\lambda_A+\lambda_{AC}\}}} = 1$$

```
## 25   7    yes   no 65-74
## 26  28     no   no 65-74
## 27   0    yes   no   75+
## 28   0     no   no   75+
```

Se in prima battuta ignoriamo la suddivisione in classi di età, e consideriamo la tabella marginale

```
margSD=xtabs(y~smoker+dead,data=femsmoke)
margSD

##       dead
## smoker yes  no
##    yes 139 443
##    no  230 502

prop.table(margSD,1)

##       dead
## smoker       yes        no
##    yes 0.2388316 0.7611684
##    no  0.3142077 0.6857923
```

dovremmo concludere che il fumo ha un effetto benefico sulla sopravvivenza, si può calcolare un test del χ^2 che farà concludere che tale effetto non è ascrivibile al caso.
Osserviamo però che se la probabilità di morte è calcolata condizionatamente alla classe di età, questa è maggiore per le fumatrici in tutte le classi tranne la classe 25-34.

```
tavola=xtabs(y~smoker+dead+age,data=femsmoke)
prop.table(tavola,c(1,3))[,1,]

##       age
## smoker      18-24      25-34      35-44      45-54      55-64
##    yes 0.03636364 0.02419355 0.12844037 0.20769231 0.44347826
##    no  0.01612903 0.03184713 0.05785124 0.15384615 0.33057851
##       age
## smoker      65-74        75+
##    yes 0.80555556 1.00000000
##    no  0.78294574 1.00000000
```

Il fatto che marginalmente l'associazione tra fumo e sopravvivenza sia contraria si deve alla prevalenza di fumatrici nelle diverse classi di età, si nota infatti che vi sono molte più fumatrici tra le giovani.

```
prop.table(xtabs(y~smoker+age,data=femsmoke),2)
```

```
##        age
## smoker     18-24     25-34     35-44     45-54     55-64     65-74
##    yes 0.4700855 0.4412811 0.4739130 0.6250000 0.4872881 0.2181818
##    no  0.5299145 0.5587189 0.5260870 0.3750000 0.5127119 0.7818182
##        age
## smoker       75+
##    yes 0.1688312
##    no  0.8311688
```

Stimiamo dapprima il modello con i soli effetti marginali, questo risulta poco soddisfacente in termini di devianza residua,

```
mod1=glm(y~smoker+dead+age,family=poisson,data=femsmoke)
summary(mod1)
```

```
##
## Call:
## glm(formula = y ~ smoker + dead + age, family = poisson, data = femsmoke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.9306  -5.3175  -0.5514   2.4229  11.1895
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.67778    0.10702  25.021  < 2e-16 ***
## smokerno     0.22931    0.05554   4.129 3.64e-05 ***
## deadno       0.94039    0.06139  15.319  < 2e-16 ***
## age25-34     0.87618    0.11003   7.963 1.67e-15 ***
## age35-44     0.67591    0.11356   5.952 2.65e-09 ***
## age45-54     0.57536    0.11556   4.979 6.40e-07 ***
## age55-64     0.70166    0.11307   6.206 5.45e-10 ***
## age65-74     0.34377    0.12086   2.844  0.00445 **
## age75+      -0.41837    0.14674  -2.851  0.00436 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1193.9  on 27  degrees of freedom
## Residual deviance:  735.0  on 19  degrees of freedom
## AIC: 887.2
##
```

```
## Number of Fisher Scoring iterations: 6
```

All'opposto, stimiamo il modello saturo e analizziamo le devianze delle componenti, risulta che l'interazione a tre ha un contributo trascurabile, quella tra `dead` e `smoker` ha invece un contributo modesto ma potenzialmente rilevante.

```
modS=glm(y~age*dead*smoker,family=poisson,data=femsmoke)
anova(modS)

## 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
## age              6   180.50        21    1013.44
## dead             1   261.27        20     752.16
## smoker           1    17.16        19     735.00
## age:dead         6   633.17        13     101.83
## age:smoker       6    93.51         7       8.33
## dead:smoker      1     5.95         6       2.38
## age:dead:smoker  6     2.38         0       0.00
```

Consideriamo quindi i modelli con tutte le interazioni a due, `mod3`, e il modello che omette quella tra `dead` e `smoker`, `mod2`. Entrambi hanno una devianza residua soddisfacente, si noti però che da `mod3` si può concludere che l'interazione tra `dead` e `smoker` ha un coefficiente significativamente diverso da zero.

```
mod2=glm(y~age*smoker+dead*age,family=poisson,data=femsmoke)
mod3=glm(y~age*smoker+dead*age+smoker*dead,family=poisson,data=femsmoke)
summary(mod2)

##
## Call:
## glm(formula = y ~ age * smoker + 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
## 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             0.11980    0.18523   0.647 0.517785
## deadno               3.63759    0.58490   6.219 5.00e-10 ***
## age25-34:smokerno    0.11616    0.22078   0.526 0.598789
## age35-44:smokerno   -0.01536    0.22749  -0.068 0.946172
## age45-54:smokerno   -0.63063    0.23414  -2.693 0.007074 **
## age55-64:smokerno   -0.06894    0.22643  -0.304 0.760765
## age65-74:smokerno    1.15649    0.26427   4.376 1.21e-05 ***
## age75+:smokerno      1.47413    0.35617   4.139 3.49e-05 ***
## age25-34:deadno     -0.10756    0.68613  -0.157 0.875435
## age35-44:deadno     -1.33977    0.62810  -2.133 0.032920 *
## age45-54:deadno     -2.17125    0.61128  -3.552 0.000382 ***
## age55-64:deadno     -3.17171    0.59999  -5.286 1.25e-07 ***
## age65-74:deadno     -4.94977    0.61512  -8.047 8.49e-16 ***
## age75+:deadno      -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
```

```
summary(mod3)
```

```
##
## Call:
## glm(formula = y ~ age * smoker + dead * age + smoker * dead,
##     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
## 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            -0.29666    0.25324  -1.171 0.241401
## deadno               3.43271    0.59014   5.817 6.00e-09 ***
## age25-34:smokerno    0.11752    0.22091   0.532 0.594749
## age35-44:smokerno    0.01268    0.22800   0.056 0.955654
## age45-54:smokerno   -0.56538    0.23585  -2.397 0.016522 *
## age55-64:smokerno    0.08512    0.23573   0.361 0.718030
## age65-74:smokerno    1.49088    0.30039   4.963 6.93e-07 ***
## age75+:smokerno      1.89060    0.39582   4.776 1.78e-06 ***
## age25-34:deadno     -0.12006    0.68655  -0.175 0.861178
## age35-44:deadno     -1.34112    0.62857  -2.134 0.032874 *
## age45-54:deadno     -2.11336    0.61210  -3.453 0.000555 ***
## age55-64:deadno     -3.18077    0.60057  -5.296 1.18e-07 ***
## age65-74:deadno     -5.08798    0.61951  -8.213  < 2e-16 ***
## age75+:deadno      -27.31727 8839.01146  -0.003 0.997534
## smokerno:deadno      0.42741    0.17703   2.414 0.015762 *
## ---
## 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
```

4 Boy scout

Si considerino i dati riportati nella seguente tabella in cui un campione di individui è classificato secondo l'essere o meno stato boy scout (B), l'aver avuto problemi con la legge (L) e lo status socio-economico (S):

Status socioeconomico	Boy scout	Problemi legali Sì	No
Basso	Sì	12	37
	No	42	154
Medio	Sì	19	140
	No	22	128
Alto	Sì	8	192
	No	3	51

1. Individuare un modello log-lineare opportuno.
2. Calcolare le frequenze attese.
3. Si costruisca la marginale (B-L) e la si esamini. Perché e fuorviante sostenere che fare il boy scout conduce a un tasso di delinquenza minore? Si risponda al quesito anche interpretando la struttura del modello log-lineare individuato.

```
Count=c(12, 37, 42, 154, 19, 140, 22, 128,8, 192, 3, 51)
Legal=gl(2,1,12,labels=c("S","N"))
Boyscout=gl(2,2,12,labels=c("S","N"))
Socioec=gl(3,4,12,labels=c("Basso","Medio","Alto"))
d=data.frame(Count,Legal,Boyscout,Socioec)
ftable(xtabs(d,formula=Count~Socioec+Boyscout+Legal))
```

```
##                  Legal   S   N
## Socioec Boyscout
## Basso   S               12  37
##         N               42 154
## Medio   S               19 140
##         N               22 128
## Alto    S                8 192
##         N                3  51
```

```
bs=glm(Count~Socioec*Boyscout+Socioec*Legal+Boyscout*Legal,family=poisson,data=d)
summary(bs)
```

```
##
## Call:
## glm(formula = Count ~ Socioec * Boyscout + Socioec * Legal +
##     Boyscout * Legal, family = poisson, data = d)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
##  0.54708  -0.29073  -0.27338   0.14535  -0.29116   0.11000   0.28227
##        8         9        10        11        12
```

```
## -0.11430  -0.17295   0.03606   0.30595  -0.06967
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.32271    0.24294   9.561  < 2e-16 ***
## SocioecMedio             0.68779    0.25565   2.690  0.00714 **
## SocioecAlto             -0.18274    0.37119  -0.492  0.62251
## BoyscoutN                1.45685    0.24813   5.871 4.32e-09 ***
## LegalN                   1.33563    0.24817   5.382 7.37e-08 ***
## SocioecMedio:BoyscoutN -1.43710    0.19705  -7.293 3.03e-13 ***
## SocioecAlto:BoyscoutN  -2.68021    0.22500 -11.912  < 2e-16 ***
## SocioecMedio:LegalN      0.58621    0.23941   2.449  0.01434 *
## SocioecAlto:LegalN       1.77929    0.37199   4.783 1.73e-06 ***
## BoyscoutN:LegalN        -0.08997    0.24070  -0.374  0.70857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 722.84954  on 11  degrees of freedom
## Residual deviance:   0.79897  on  2  degrees of freedom
## AIC: 86.043
##
## Number of Fisher Scoring iterations: 4
```

```
bs2=glm(Count~Socioec*Boyscout+Socioec*Legal,family=poisson,data=d)
summary(bs2)
```

```
##
## Call:
## glm(formula = Count ~ Socioec * Boyscout + Socioec * Legal, family = poisson,
##     data = d)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
##  0.35868  -0.19519  -0.18343   0.09695  -0.46446   0.17813   0.46215
##        8         9        10        11        12
## -0.18436  -0.22770   0.04779   0.41421  -0.09222
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.3795     0.1867  12.747  < 2e-16 ***
## SocioecMedio             0.6696     0.2496   2.683  0.00730 **
## SocioecAlto             -0.2207     0.3561  -0.620  0.53548
## BoyscoutN                1.3863     0.1597   8.680  < 2e-16 ***
```

```
## LegalN                   1.2633     0.1541   8.197 2.47e-16 ***
## SocioecMedio:BoyscoutN  -1.4446     0.1961  -7.365 1.77e-13 ***
## SocioecAlto:BoyscoutN   -2.6956     0.2214 -12.174  < 2e-16 ***
## SocioecMedio:LegalN      0.6141     0.2278   2.696  0.00701 **
## SocioecAlto:LegalN       1.8319     0.3446   5.315 1.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 722.84954  on 11  degrees of freedom
## Residual deviance:   0.93903  on  3  degrees of freedom
## AIC: 84.183
##
## Number of Fisher Scoring iterations: 4
```

```
bs3=glm(Count~Socioec+Boyscout*Legal,family=poisson,data=d)
summary(bs3)
```

```
##
## Call:
## glm(formula = Count ~ Socioec + Boyscout * Legal, family = poisson,
##     data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.2394  -2.2105  -0.0217   1.8107   6.4423
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.47026    0.16878  14.636  < 2e-16 ***
## SocioecMedio      0.23208    0.08554   2.713  0.00667 **
## SocioecAlto       0.03608    0.08955   0.403  0.68704
## BoyscoutN         0.54113    0.20141   2.687  0.00722 **
## LegalN            2.24723    0.16838  13.346  < 2e-16 ***
## BoyscoutN:LegalN -0.64379    0.21513  -2.993  0.00277 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 722.85  on 11  degrees of freedom
## Residual deviance: 212.70  on  6  degrees of freedom
## AIC: 289.94
##
```

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

## marginale B-L
d.array=xtabs(d,formula=Count~Socioec+Boyscout+Legal)
dim(d.array)

## [1] 3 2 2

margBL=apply(d.array,c(2,3),sum)
d.margBL=as.data.frame.table(margBL)
mod.margBL=glm(Freq~Boyscout*Legal,family=poisson,data=d.margBL)
```

5 Larve

Larve di mosche *drosophila fruit fly* vengono immerse in un recipiente che contiene DDT. Le stesse sono poi classificate secondo il sesso della mosca, la posizione nel recipiente e la presenza o meno di segni di avvelenamento:

```
ddt

##    Counts    Sex          Site Poisoned
## 1      55   Male     In medium  Healthy
## 2      34 Female     In medium  Healthy
## 3      23   Male Medium margin  Healthy
## 4      15 Female Medium margin  Healthy
## 5       7   Male     Vial wall  Healthy
## 6       3 Female     Vial wall  Healthy
## 7       8   Male Top of medium  Healthy
## 8       5 Female Top of medium  Healthy
## 9       6   Male     In medium Poisoned
## 10     17 Female     In medium Poisoned
## 11      1   Male Medium margin Poisoned
## 12      5 Female Medium margin Poisoned
## 13      4   Male     Vial wall Poisoned
## 14      5 Female     Vial wall Poisoned
## 15      3   Male Top of medium Poisoned
## 16      3 Female Top of medium Poisoned
```

Si organizzino i dati in tabelle di frequenza: si possono osservare delle regolarità nei dati?
Si stimi un modello GLM per questi dati, assumendo che il disegno sperimentale fissasse la posizione nel recipiente e il sesso.
Cosa si conclude?

6 Titanic

I dati `Titanic` descrivono caratteristiche dei passeggeri della famosa nave. Si veda `help(Titanic)` per una descrizione dei dati. Si costruisca un modello GLM per analizzare i dati in un modo appropriato.