Tema 6: Modelos Log-Lineales para tablas de Contingencia

Introducción

Habitualmente, se suelen estudiar las tablas de contingencia calculando estadísticos del tipo χ^2 para contrastar independencia entre las variables. Cuando hay más variables involucradas, una posibilidad es repetir el análisis por parejas para las distintas sub-tablas y determinar las interacciones o asociaciones entre las variables.

Pero otra alternativa posible es aplicar modelos *log–lineales*, que son un caso particular de los GLM para datos distribuidos como una distribución multinomial o como una Poisson.

Los modelos log–lineales se usan para analizar la relación entre dos, tres o más variables categóricas en una tabla de contingencia. Todas las variables que se analizan se consideran como variables respuesta, es decir, no se hace distinción entre variables *independientes* y *dependientes*. Es por ello que en estos modelos solo se estudia asociación entre las variables.

Los modelos se representan mediante las frecuencias esperadas y se tienen en cuenta las asociaciones o interacciones entre las variables. Los patrones de asociación entre las variables pueden describirse en términos de los *odds* y las *razones de odds*.

Se parte de una tabla de contingencia $I \times J$ en la que se estudian n individuos. Cuando las respuestas son independientes, las probabilidades conjuntas de cada casilla π_{ij} se obtienen como el producto de las marginales de filas y columnas

$$\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$$

para $i = 1, \ldots, I$, $j = 1, \ldots, J$.

Las probabilidades π_{ij} son los parámetros de una distribución multinomial, pero en los modelos log-lineales se usan frecuencias esperadas $\mu_{ij} = n\pi_{ij}$ en lugar de las probabilidades

π_{ij}. También se pueden considerar distribuciones de Poisson con valores esperados μ_{ij}.

Asumiendo independencia, se tiene que $\mu_{ij} = n\pi_{i+} \cdot \pi_{+j}$ para todo i y j.

Modelos log-lineales de independencia para tablas de contingencia

Dado que $\mu_{ij} = n\pi_{i+} \cdot \pi_{+j}$ para todo i y j, si se toman logaritmos:

$$\begin{aligned} \log(\mu_{ij}) &= \log(n) + \log(\pi_{i+}) + \log(\pi_{+j}) = \\ &= \lambda + \lambda_i^X + \lambda_j^Y \end{aligned}$$

donde se denomina a λ_i^X el efecto *fila*, λ_j^Y el efecto *columna*. Este modelo se denomina **log-lineal de independencia**.

La interpretación de los parámetros es más sencilla para respuestas binarias. Por ejemplo, en el modelo de independencia para una tabla $I \times 2$, donde las columnas corresponden a la respuesta Y, en cada fila i el logit de π_i para $Y = 1$ es

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) = \log(\mu_{i1}) - \log(\mu_{i2}) = \\ &= \left(\lambda + \lambda_i^X + \lambda_1^Y\right) - \left(\lambda + \lambda_i^X + \lambda_2^Y\right) = \lambda_1^Y - \lambda_2^Y \end{aligned}$$

lo cual no depende de i, es decir, el logit de Y no depende del nivel de X (de la fila), lo que corresponde al caso simple en que $\text{logit}(\pi_i) = cte$.

Así, la probabilidad de clasificar algo en una columna particular es constante a lo largo de las filas.

Identificabilidad y restricciones sobre los parámetros

En una tabla de 2×2, por ejemplo, el modelo de independencia especifica 5 parámetros y por lo tanto es redundante.

Como en el caso de los modelos *ANOVA*, se pueden imponer restricciones para los parámetros para evitar redundancias entre ellos. Por ejemplo, se puede imponer que para el primer nivel de cada variable el parámetro sea 0 o bien se puede imponer que la suma de los parámetros correspondientes a una variable sea 0.

Por ejemplo, en una tabla 2×2:

$$\begin{aligned} \lambda_1^X + \lambda_2^X &= 0 \\ \lambda_1^Y + \lambda_2^Y &= 0 \end{aligned}$$

y se cumple que la diferencia entre dos efectos principales es la misma.

Ejemplo

Se tiene una muestra de personas en donde se les ha preguntado si creían en la vida después de la muerte. El número de personas que respondía **sí** fue 1339 de entre 1639 de raza *blanca*, 260 de entre 315 de raza de *color* y 88 de 110 clasificadas como *otros*.

Se usa un modelo log-lineal de independencia sobre la correspondiente tabla 3×2 y se fija un nivel de los efectos en 0.

```
Raza = c("White", "White", "Black", "Black",
"Others", "Others")
Cree = c("SI", "NO", "SI", "NO", "SI", "NO")
datos = c(1339, 300, 260, 55, 88, 22)
Raza = as.factor(Raza)
Cree = as.factor(Cree)

fit = glm(datos ~ Raza + Cree, family=poisson(link="log"))
summary(fit)
```

```
Call:
glm(formula = datos ~ Raza + Cree, family = poisson(link = "log"))

Deviance Residuals:
       1         2         3         4         5         6
-0.01717   0.03631   0.15781  -0.33688  -0.20194   0.41917

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.05242    0.07309   55.44   <2e-16 ***
RazaOthers  -1.05209    0.11075   -9.50   <2e-16 ***
RazaWhite    1.64927    0.06152   26.81   <2e-16 ***
CreeSI       1.49846    0.05697   26.30   <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: 2849.21758  on 5  degrees of freedom
Residual deviance:    0.35649  on 2  degrees of freedom
AIC: 49.437

Number of Fisher Scoring iterations: 3
```

Se obtiene que $\lambda_1^Y = 1{,}4985$ y $\lambda_2^Y = 0$. Por tanto, el odds de la creencia en la vida después de la muerte fue de $\exp(1{,}4985) = 4{,}475$ para cada raza.

Modelo Saturado

Cuando las variables son dependientes satisfacen un modelo más complejo:

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

donde los parámetros λ_{ij}^{XY} reflejan la asociación entre X e Y. Este modelo describe cualquier conjunto de frecuencias observadas y es el modelo más general para una tabla de contingencia bivariante. El caso de independencia corresponde a $\lambda_{ij}^{XY} = 0$.

En tablas 2×2 existe una relación directa entre el logaritmo de la razón de odds y los parámetros de asociación λ_{ij}^{XY}.

$$\log\theta = \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

Si $\lambda_{ij}^{XY} = 0$ las razones de odds valen 1 y X e Y son independientes.

Ejemplo

Se podría ajustar un modelo saturado en la tabla de las creencias en la vida eterna:

```
fitT = glm(datos ~ Raza*Cree, family=poisson(link="log"))
summary(fitT)
```

```
Call:
glm(formula = datos ~ Raza * Cree, family = poisson(link = "log"))

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

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        4.00733    0.13484  29.719  < 2e-16 ***
RazaOthers        -0.91629    0.25226  -3.632 0.000281 ***
RazaWhite          1.69645    0.14668  11.566  < 2e-16 ***
CreeSI             1.55335    0.14842  10.466  < 2e-16 ***
RazaOthers:CreeSI -0.16705    0.28080  -0.595 0.551889
RazaWhite:CreeSI  -0.05745    0.16158  -0.356 0.722165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2.8492e+03  on 5  degrees of freedom
Residual deviance: 9.1038e-14  on 0  degrees of freedom
AIC: 53.081

Number of Fisher Scoring iterations: 3
```

Cada uno de los términos o coeficientes representa un incremento o decremento en el valor del logaritmo de los recuentos predichos. Así, si es igual a 1.696, se puede afirmar que el factor raza blanca incrementa el logaritmo de los recuentos en 1.696 cuando el resto de factores permanece constante.

A partir de los *p-valores* se observa que ninguna de las interacciones es significativa y que el mejor modelo es el de independencia.

Ejemplo

Más sobre la vida *eterna*... Se observan los resultados según el género de las personas:

	Cree vida eterna		
	Si	**No**	*Total*
Mujer	435	147	582
Hombre	375	134	509
Total	810	281	1091

En los datos de las creencias en la vida eterna la razón de odds es

$$\theta = \frac{435 \cdot 134}{147 \cdot 375} = 1{,}057$$

de modo que $\log \theta = 0{,}0558$ y, por tanto,

$$\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY} = 0{,}0558$$

Es decir, la razón de odds es casi 1: no parece haber diferencias de creencias según el género.

```
sexo = c("mujer", "hombre", "mujer", "hombre")
Cree = c("SI", "NO", "NO", "SI")
datos = c(435, 134, 147, 375)

sexo = as.factor(sexo)
Cree = as.factor(Cree)

fit = glm(datos ~ sexo * Cree, family=poisson(link="log"))
summary(fit)
```

```
Call:
glm(formula = datos ~ sexo * Cree, family = poisson(link = "log"))

Deviance Residuals:
[1]  0  0  0  0

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.89784    0.08639  56.697   <2e-16 ***
sexomujer         0.09259    0.11944   0.775    0.438
CreeSI            1.02909    0.10064  10.225   <2e-16 ***
sexomujer:CreeSI  0.05583    0.13868   0.403    0.687
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2.7269e+02  on 3  degrees of freedom
Residual deviance: 1.2790e-13  on 0  degrees of freedom
AIC: 37.245
```

El modelo saturado tiene $1+(I-1)+(J-1)+(I-1)(J-1) = IJ$ parámetros, es decir, tiene tantos parámetros como observaciones, dando un ajuste perfecto. En la práctica, normalmente se trata de usar modelos no saturados ya que dan lugar a interpretaciones más simples.

Los modelos log-lineales son modelos jerárquicos, es decir, incluyen todos los términos de orden menor que están presentes en un término de orden mayor. Así, si el modelo contiene λ_{ij}^{XY}, entonces también están presentes en el modelo los términos λ_i^X y λ_j^Y.

Cuando un modelo tiene interacciones hay que tener cuidado con la interpretación de los efectos individuales, como sucede también en los modelos *ANOVA* bifactoriales, ya que los efectos principales podrían estar enmascarados por la interacción.

Modelos log-lineales para tablas tridimensionales

Supongamos que tenemos tres variables categóricas X, Y y Z que tienen como valores posibles:

$$\begin{aligned} X &\Rightarrow 1, 2, \ldots, I \\ Y &\Rightarrow 1, 2, \ldots, J \\ Z &\Rightarrow 1, 2, \ldots, K \end{aligned}$$

En este caso, se pueden considerar distintos modelos log-lineales para las frecuencias esperadas por casilla.

Por ejemplo, el modelo

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

contiene un término XZ $\left(\lambda_{ik}^{XZ}\right)$, que modeliza la asociación entre X y Z, condicionada por Y. Este modelo también permite la asociación entre YZ, condicionada por X.

No hay un término de asociación entre XY: se especifica la independencia condicional entre X e Y, condicionada por Z. Es decir, X e Y son independientes en la tabla parcial correspondiente a cada nivel k (para todo k) de Z.

Este modelo se puede resumir como (XZ, YZ), notación que indica los términos de interacción más alta de las variables. Se interpreta como la asociación entre dos variables (X e Y) que desaparece cuando condicionamos por una tercera variable (Z).

Un modelo que permite todas las asociaciones condicionales de las tres variables por parejas es

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

que corresponde a *asociación homogénea* (XY, XZ, YZ). En este modelo los parámetros están directamente relacionados mediante las razones de odds condicionales.

En el caso de tablas $2 \times 2 \times K$ la razón de odds condicional de XY, que se denota como $\theta_{XY(k)}$, describe la asociación entre X e Y en las sub-tablas que se obtienen para

un k fijo. Así,

$$\log \theta_{XY(k)} = \log \left(\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \right) =$$

$$= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

El lado derecho de la expresión no depende de k de modo que la razón de odds es la misma para cualquier nivel de Z.

Otro modelo posible es el de *independencia mutua* que contiene sólo los términos individuales (X, Y, Z). Este trata a cada par de variables como independientes, tanto marginal como condicionalmente, aunque se presenta raramente en la realidad.

El modelo log-lineal más general para tablas tridimensionales es

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

que se denota por (XYZ) y es un modelo *saturado* que proporciona el ajuste perfecto.

Ejemplo

Se considera una muestra entre estudiantes de Ohio, donde se les preguntó si habían tomado alguna vez alcohol, tabaco o marihuana. Se denotan las variables como A para alcohol, C para tabaco (cigarrillos) y M para marihuana.

		Marihuana	
Alcohol	*Tabaco*	**Si**	**No**
Si	**Si**	911	538
	No	44	456
No	**Si**	3	43
	No	2	279

Se introducen los datos en R.

```
tabla = data.frame(expand.grid(
marihuana = factor(c("Yes","No"), levels=c("No","Yes")),
tabaco = factor(c("Yes","No"), levels=c("No","Yes")),
alcohol = factor(c("Yes","No") ,levels=c("No","Yes"))),
count = c(911,538,44,456,3,43,2,279))
```

Se ajusta un modelo log-lineal usando el comando `loglm`. Se ajusta el modelo saturado, luego el modelo de asociación homogénea y así sucesivamente hasta llegar al modelo de independencia marginal. Se obtienen tanto los estimadores de los parámetros como los valores predichos.

```
library(MASS)

# ACM
fitACM =
loglm(count ~ alcohol*tabaco*marihuana, data=tabla ,param=T, fit=T)

# AC, AM, CM
fitAC.AM.CM = update(fitACM, .~. - alcohol:tabaco:marihuana)

# AM, CM
fitAM.CM = update(fitAC.AM.CM, .~. - alcohol:tabaco)

# AC, M
fitAC.M = update(fitAC.AM.CM, .~. - alcohol:marihuana -
tabaco:marihuana)

# A, C, M
fitA.C.M = update(fitAC.M, .~. - alcohol:tabaco)
```

Se obtienen los recuentos estimados con la orden `fitted`. Para convertirlos en vectores se transponen mediante la orden `aperm` y luego se escriben en un *dataframe*.

```
data.frame(tabla[8:1,-4],
ACM = c(aperm(fitted(fitACM))),
AC.AM.CM = c(aperm(fitted(fitAC.AM.CM))),
AM.CM = c(aperm(fitted(fitAM.CM))),
AC.M = c(aperm(fitted(fitAC.M))),
A.C.M = c(aperm(fitted(fitA.C.M))))
```

```
  marihuana tabaco alcohol ACM   AC.AM.CM       AM.CM      AC.M     A.C.M
8        No     No      No 279 279.614402 179.8404255 162.47627  64.87990
7       Yes     No      No   2   1.383160   0.2395833 118.52373  47.32880
6        No    Yes      No  43  42.383882 142.1595745  26.59754 124.19392
5       Yes    Yes      No   3   3.616919   4.7604167  19.40246  90.59739
4        No     No     Yes 456 455.385598 555.1595745 289.10369 386.70007
3       Yes     No     Yes  44  44.616840  45.7604167 210.89631 282.09123
2        No    Yes     Yes 538 538.616118 438.8404255 837.82250 740.22612
1       Yes    Yes     Yes 911 910.383081 909.2395833 611.17750 539.98258
```

El modelo que predice mejor las observaciones es el de *asociación homogénea* (AC, AM , CM), es decir, el modelo denotado como `AC.AM.CM`.

Así, los parámetros estimados son:

```
fitAC.AM.CM$param
```

```
$(Intercept)
[1] 4.251537

$alcohol
       No       Yes
-1.503994  1.503994

$tabaco
        No        Yes
-0.2822777  0.2822777

$marihuana
       No       Yes
 1.196045 -1.196045

$alcohol.tabaco
       tabaco
alcohol         No        Yes
    No   0.5136255 -0.5136255
    Yes -0.5136255  0.5136255

$alcohol.marihuana
       marihuana
alcohol        No       Yes
    No   0.746502 -0.746502
    Yes -0.746502  0.746502

$tabaco.marihuana
      marihuana
tabaco         No        Yes
   No   0.7119739 -0.7119739
   Yes -0.7119739  0.7119739
```

De modo alternativo, se puede considerar el modelo obtenido mediante el comando `glm` con la familia de distribuciones de Poisson.

```
options(contrasts = c("contr.sum", "contr.poly"))

(fit.glm = glm(count ~ alcohol + tabaco + marihuana + alcohol:tabaco +
alcohol:marihuana + tabaco:marihuana, data=tabla,
family=poisson))
```

```
Call:  glm(formula = count ~ alcohol + tabaco + marihuana + alcohol:tabaco +
    alcohol:marihuana + tabaco:marihuana, family = poisson, data = tabla)

Coefficients:
        (Intercept)             alcohol1              tabaco1
             4.2515              -1.5040              -0.2823
         marihuana1     alcohol1:tabaco1  alcohol1:marihuana1
             1.1960               0.5136               0.7465
 tabaco1:marihuana1
             0.7120

Degrees of Freedom: 7 Total (i.e. Null);  1 Residual
Null Deviance:      2851
Residual Deviance: 0.374        AIC: 63.42
```

Relaciones entre modelos log-lineales y regresión logística

Los modelos log-lineales para tablas de contingencia modelizan la asociación entre variables categóricas, mientras que los modelos de regresión logística describen cómo una variable categórica de tipo respuesta depende de un grupo de variables explicativas. Aunque los modelos parecen diferentes, en realidad existen conexiones entre ellos.

En un modelo log-lineal, se pueden calcular los logits de una respuesta para ayudar a la interpretación del modelo. Además, un modelo logístico con variables explicativas categóricas tiene un equivalente en un modelo log-lineal.

Como ilustración, se puede considerar un modelo saturado para una tabla de dos variables,

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

Supongamos que Y es binaria. Consideramos a Y como una variable respuesta y a X como explicativa. Cuando X está en el nivel i,

$$\begin{aligned} \text{logit}\left(P(Y=1)\right) &= \log\left(\frac{nP\left(Y=1|X=i\right)}{nP\left(Y=2|X=i\right)}\right) = \\ &= \log(\mu_{i1}) - \log(\mu_{i2}) \\ &= \left(\lambda_1^Y - \lambda_2^Y\right) + \left(\lambda_{i1}^{XY} - \lambda_{i2}^{XY}\right) \end{aligned}$$

El primer término no depende de i, y el segundo término depende del nivel i de X. De este modo, el logit tiene la forma

$$\text{logit}\left(P(Y=1)\right) = \alpha + \beta_i^X$$

Cuando hay una variable respuesta única de tipo binaria, los modelos log-lineales que se pueden aplicar corresponden a modelos logísticos para esa respuesta. Cuando la respuesta tiene más de dos categorías, los correspondientes modelos log-lineales corresponden a modelos logit multinomiales con una categoría de referencia.

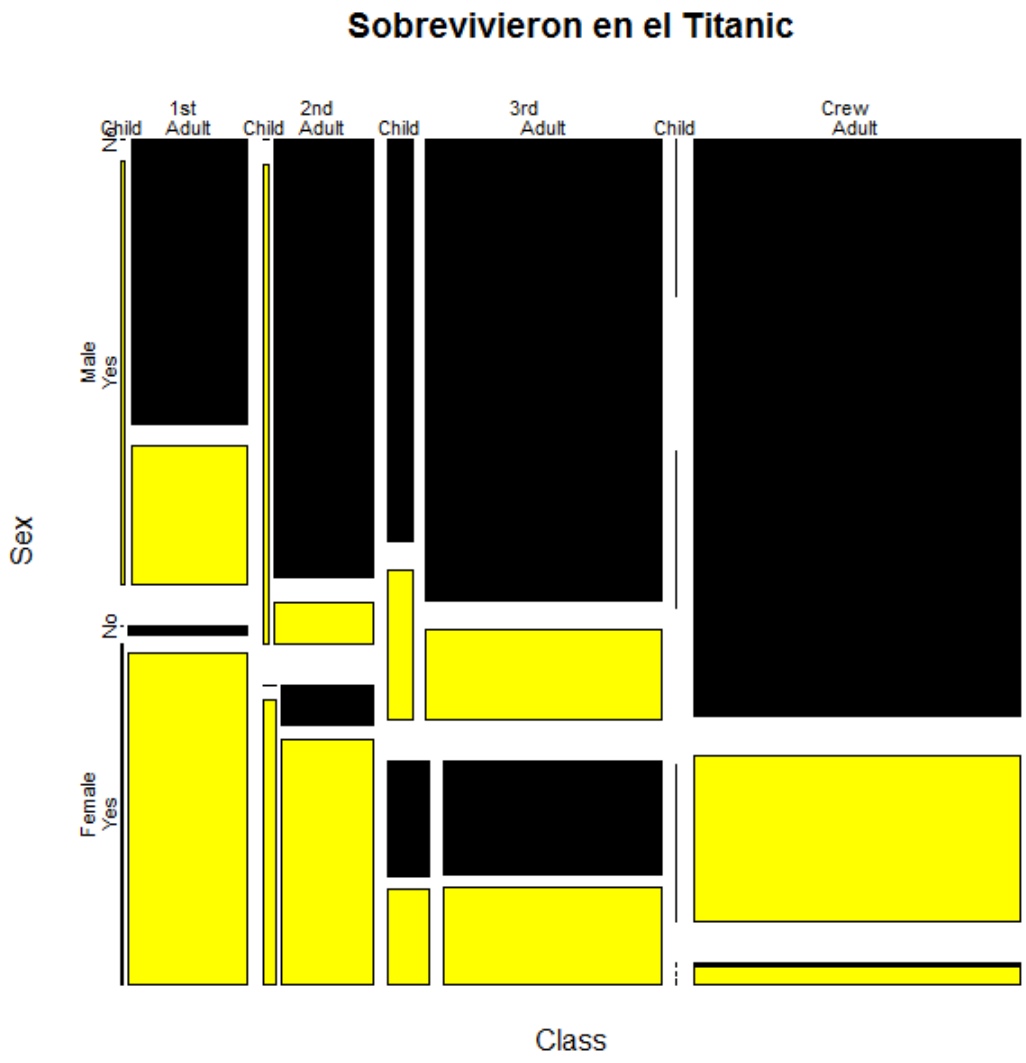
En resumen, los modelos log-lineales son más naturales cuando al menos dos variables son variables respuesta y queremos estudiar su estructura de asociación. De lo contrario, es mejor usar modelos logísticos.

Ejemplo

Se toman los datos relacionados con el hundimiento del *Titanic* en abril de 1912. El resultado se puede expresar en una tabla de dimensión 4.

Las variables son `Class` de los pasajeros (1, 2, 3, Tripulación), `Sex` de los pasajeros (Male, Female), `Age` de los pasajeros (Child, Adult), y `Survived` si los pasajeros sobrevivieron o no (No, Yes).

```
help(Titanic)
mosaicplot(Titanic, main="Sobrevivieron en el Titanic",
col=c("black","yellow"), off=c(5, 5, 5, 5))
```



```r
summary(Titanic)
```

```
Number of cases in table: 2201
Number of factors: 4
Test for independence of all factors:
        Chisq = 1637.4, df = 25, p-value = 0
        Chi-squared approximation may be incorrect
```

Se considera entonces un modelo log-lineal.

Tipos posibles de efectos

- $Class$: Hay más pasajeros en algunas clases que en otras.
- Sex: Hay más pasajeros en un sexo que en otro.
- Age: Hay más pasajeros en un grupo de edad que en otro.
- $Survived$: Hay más pasajeros o vivos o muertos que la alternativa.
- $Class \times Sex$: Class y Sex no son independientes.
- $Class \times Age$: Class y Age no son independientes.
- $Class \times Survived$: Class y Survived no son independientes.
- $Sex \times Age$: Sex y Age no son independientes.
- $Sex \times Survived$: Sex y Survived no son independientes.
- $Age \times Survived$: Age y Survived no son independientes.
- $Class \times Sex \times Age$, $Class \times Sex \times Survived$, $Class \times Age \times Survived$, $Sex \times Age \times Survived$: hay interacción triple entre las variables.
- $Class \times Sex \times Age \times Survived$: hay interacción cuádruple entre las variables.

```
library(MASS)

simple = loglm(~ Class + Sex + Age + Survived, data=Titanic)
simple
```

```
Statistics:
                     X^2 df P(> X^2)
Likelihood Ratio 1243.663 25        0
Pearson          1637.445 25        0
```

```
sat.model = loglm(~ Class * Sex * Age * Survived, data=Titanic)
sat.model
```

```
Statistics:
                X^2 df P(> X^2)
Likelihood Ratio   0  0        1
Pearson          NaN  0        1
```

El modelo con solo los efectos simples no es adecuado (se rechaza la hipótesis nula, ya que el *p-valor* es casi 0). Debe haber interacciones entre las variables. Por otro lado, en el otro extremo, el modelo saturado no es útil ya que predice completamente todas las frecuencias observadas.

Se pueden ir probando distintos modelos y comprobando su significación. Alternativamente se puede usar un procedimiento *stepwise*.

```
stepAIC(sat.model, direction="backward", trace=0)
```

```
Call:
loglm(formula = ~Class + Sex + Age + Survived + Class:Sex +
Class:Age + Sex:Age + Class:Survived + Sex:Survived + Age:Survived +
Class:Sex:Age + Class:Sex:Survived + Class:Age:Survived,
data = Titanic, evaluate = FALSE)

Statistics:
                     X^2 df  P(> X^2)
Likelihood Ratio 1.685479  4 0.7933536
Pearson               NaN  4       NaN
```

Se parte del modelo saturado y se usa el procedimiento *backward* con la orden `stepAIC`. Se toma finalmente el modelo con menor valor de AIC.

El modelo mejor, parece el que elimina la interacción de orden 4 y la interacción `Sex:Age:Survived`.

```
step.model = loglm(formula = ~Class + Sex + Age + Survived +
Class:Sex + Class:Age + Sex:Age + Class:Survived +
Sex:Survived + Age:Survived + Class:Sex:Age +
Class:Sex:Survived + Class:Age:Survived, data=Titanic)

print(step.model)
fitted(step.model)
```

```
Call:
loglm(formula = ~Class + Sex + Age + Survived + Class:Sex + Class:Age +
    Sex:Age + Class:Survived + Sex:Survived + Age:Survived +
    Class:Sex:Age + Class:Sex:Survived + Class:Age:Survived,
    data = Titanic)

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 1.685479  4 0.7933536
Pearson               NaN  4       NaN

, , Age = Child, Survived = No
      Sex
Class      Male   Female
  1st   0.00000  0.00000
  2nd   0.00000  0.00000
  3rd  37.43281 14.56719
  Crew  0.00000  0.00000

, , Age = Adult, Survived = No
      Sex
Class      Male  Female
  1st  118.0000  4.0000
  2nd  154.0000 13.0000
  3rd  384.5672 91.4328
  Crew 670.0000  3.0000

, , Age = Child, Survived = Yes
      Sex
Class      Male   Female
  1st   5.00000  1.00000
  2nd  10.98493 13.01507
  3rd  10.56718 16.43282
  Crew  0.00000  0.00000

, , Age = Adult, Survived = Yes
      Sex
Class       Male    Female
  1st   57.00000 140.00000
  2nd   14.02291  79.97709
  3rd   77.43281  73.56719
  Crew 192.00000  20.00000
```

Se puede hacer el análisis también con la función `glm`. Para ello hay que pasar los datos a un *dataframe*.

```r
ti = as.data.frame(Titanic)
glm.model = glm(Freq ~ Class*Age*Sex*Survived, data=ti,
family=poisson)

stepAIC(glm.model, direction="backward", trace=0)
```

```
Call:  glm(formula = Freq ~ Class + Age + Sex + Survived + Class:Age +
    Class:Sex + Age:Sex + Class:Survived + Age:Survived + Sex:Survived +
    Class:Age:Sex + Class:Age:Survived + Class:Sex:Survived,
    family = poisson, data = ti)

Coefficients:
                     (Intercept)                           Class2nd
                       -20.54742                           -0.76313
                         Class3rd                          ClassCrew
                        24.16996                           -1.97334
                         AgeAdult                          SexFemale
                        25.31810                           -5.89242
                      SurvivedYes                 Class2nd:AgeAdult
                        22.15686                            1.02940
               Class3rd:AgeAdult                ClassCrew:AgeAdult
                       -22.98853                            3.70993
              Class2nd:SexFemale                Class3rd:SexFemale
                         1.84450                            4.94864
             ClassCrew:SexFemale                AgeAdult:SexFemale
                         4.31897                            2.50803
            Class2nd:SsurvivedYes              Class3rd:SurvivedYes
                         1.55159                          -23.42165
           ClassCrew:SurvivedYes              AgeAdult:SurvivedYes
                       -23.73031                          -22.88449
           SexFemale:SurvivedYes       Class2nd:AgeAdult:SexFemale
                         4.28298                           -0.93211
     Class3rd:AgeAdult:SexFemale      ClassCrew:AgeAdult:SexFemale
                        -3.00077                           -6.34324
   Class2nd:AgeAdult:SurvivedYes     Class3rd:AgeAdult:SurvivedYes
                        -3.22185                           22.54657
  ClassCrew:AgeAdult:SurvivedYes    Class2nd:SexFemale:SurvivedYes
                        23.20816                           -0.06801
 Class3rd:SexFemale:SurvivedYes   ClassCrew:SexFemale:SurvivedYes
                        -2.89768                           -1.13608

Degrees of Freedom: 31 Total (i.e. Null);  4 Residual
Null Deviance:      4953
Residual Deviance: 1.685        AIC: 185.1
```

```r
step.glm = glm(formula = Freq ~ Class + Age + Sex + Survived +
Class:Age + Class:Sex + Age:Sex + Class:Survived +
Age:Survived + Sex:Survived + Class:Age:Sex +
Class:Age:Survived + Class:Sex:Survived,
family=poisson, data=ti)

anova(step.glm, test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: Freq

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  31     4953.1
Class               3   475.81        28     4477.3 < 2.2e-16 ***
Age                 1  2183.56        27     2293.8 < 2.2e-16 ***
Sex                 1   768.32        26     1525.4 < 2.2e-16 ***
Survived            1   281.78        25     1243.7 < 2.2e-16 ***
Class:Age           3   148.33        22     1095.3 < 2.2e-16 ***
Class:Sex           3   412.60        19      682.7 < 2.2e-16 ***
Age:Sex             1     6.09        18      676.6   0.01363 *
Class:Survived      3   180.90        15      495.7 < 2.2e-16 ***
Age:Survived        1    25.58        14      470.2 4.237e-07 ***
Sex:Survived        1   353.58        13      116.6 < 2.2e-16 ***
Class:Age:Sex       3     4.02        10      112.6   0.25916
Class:Age:Survived  3    35.66         7       76.9 8.825e-08 ***
Class:Sex:Survived  3    75.22         4        1.7 3.253e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(update(step.glm, . ~ . - Class:Age:Sex), test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: Freq

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  31     4953.1
Class               3   475.81        28     4477.3 < 2.2e-16 ***
Age                 1  2183.56        27     2293.8 < 2.2e-16 ***
Sex                 1   768.32        26     1525.4 < 2.2e-16 ***
Survived            1   281.78        25     1243.7 < 2.2e-16 ***
Class:Age           3   148.33        22     1095.3 < 2.2e-16 ***
Class:Sex           3   412.60        19      682.7 < 2.2e-16 ***
Age:Sex             1     6.09        18      676.6   0.01363 *
Class:Survived      3   180.90        15      495.7 < 2.2e-16 ***
Age:Survived        1    25.58        14      470.2 4.237e-07 ***
Sex:Survived        1   353.58        13      116.6 < 2.2e-16 ***
Class:Age:Survived  3    29.21        10       87.4 2.024e-06 ***
Class:Sex:Survived  3    65.43         7       22.0 4.066e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
print(update(step.glm, . ~ . - Class:Age:Sex))
```

```
Call:  glm(formula = Freq ~ Class + Age + Sex + Survived + Class:Age +
    Class:Sex + Age:Sex + Class:Survived + Age:Survived + Sex:Survived +
    Class:Age:Survived + Class:Sex:Survived, family = poisson,
    data = ti)

Coefficients:
                     (Intercept)                          Class2nd
                         -18.624                           -0.435
                         Class3rd                         ClassCrew
                          22.373                            0.985
                         AgeAdult                         SexFemale
                          23.395                           -3.508
                     SurvivedYes                 Class2nd:AgeAdult
                          19.312                            0.702
               Class3rd:AgeAdult                ClassCrew:AgeAdult
                         -21.205                            0.751
              Class2nd:SexFemale                Class3rd:SexFemale
                           0.912                            2.015
             ClassCrew:SexFemale                AgeAdult:SexFemale
                          -2.024                            0.124
           Class2nd:SurvivedYes              Class3rd:SurvivedYes
                           1.450                          -20.417
          ClassCrew:SurvivedYes              AgeAdult:SurvivedYes
                         -20.834                          -19.988
          SexFemale:SurvivedYes     Class2nd:AgeAdult:SurvivedYes
                           4.210                           -2.840
  Class3rd:AgeAdult:SurvivedYes    ClassCrew:AgeAdult:SurvivedYes
                          19.458                           20.260
 Class2nd:SexFemale:SurvivedYes    Class3rd:SexFemale:SurvivedYes
                          -0.398                           -2.799
ClassCrew:SexFemale:SurvivedYes
                          -1.063

Degrees of Freedom: 31 Total (i.e. Null);  7 Residual
Null Deviance:      4953
Residual Deviance: 21.95        AIC: 199.4
```

Ejemplo con SAS

Se ha realizado un análisis sobre el valor terapéutico del ácido ascórbico (vitamina C) en relación a su efecto sobre la gripe común. Se tiene una tabla 2×2 con los recuentos correspondientes para una muestra de 279 personas:

	Gripe	**No Gripe**	*Totales*
Placebo	31	109	140
Acido Ascorbico	17	122	139
Totales	48	231	279

Se aplica un modelo log-lineal y se analizan sus componentes.

```
OPTIONS nodate ls=75;
/* ODS listing file='/folders/myfolders/cosa.lst'; */
   ODS rtf file='/folders/myfolders/cosa.rtf' style=minimal
      startpage=no;
DATA aspirina;
INPUT tratamiento $ respuesta $ recuento;
DATALINES;
placebo gripe 31
placebo nogripe 109
ascorbico gripe 17
ascorbico nogripe 122
;
RUN;

PROC freq;
weight recuento;
tables tratamiento*respuesta/ chisq expected;
exact or;
RUN;

PROC genmod data=aspirina order=data;
class tratamiento respuesta;
model recuento = tratamiento respuesta
/link=log dist=poisson lrci type3;
RUN;
ODS rtf close;
/* ODS listing close; */
```

Procedimiento FREQ

Tabla de tratamiento por respuesta

tratamiento respuesta

Frecuencia| Esperado | Porcentaje| Pct fila | Pct col	gripe	nogripe	Total
ascorbic	17	122	139
	23.914	115.09	
	6.09	43.73	49.82
	12.23	87.77	
	35.42	52.81	
placebo	31	109	140
	24.086	115.91	
	11.11	39.07	50.18
	22.14	77.86	
	64.58	47.19	
Total	48	231	279
	17.20	82.80	100.00

Estadisticos para la tabla de tratamiento por respuesta

Estadistico	DF	Valor	Prob
Chi-cuadrado	1	4.8114	0.0283
Chi-cuadrado de ratio de verosimilitud	1	4.8717	0.0273
Chi-cuadrado adj. de continuidad	1	4.1407	0.0419
Chi-cuadrado Mantel-Haenszel	1	4.7942	0.0286
Coeficiente Phi		-0.1313	
Coeficiente de contingencia		0.1302	
V de Cramer		-0.1313	

Test exacto de Fisher

Celda (1,1) Frecuencia (F)	17
Alineado a la izquierda Pr <= F	0.0205
Alineado a la derecha Pr >= F	0.9910
Tabla de probabilidad (P)	0.0115
De dos caras Pr <= P	0.0385

Procedimiento FREQ

Estadisticos para la tabla de tratamiento por respuesta

Ratio de probabilidades y riesgos relativos

Estadistico	Valor	Limites de confianza al 95%	
Ratio de probabilidad	0.4900	0.2569	0.9343
Riesgo relativo (Columna 1)	0.5523	0.3209	0.9506
Riesgo relativo (Columna 2)	1.1273	1.0120	1.2558

Ratio de probabilidad

Ratio de probabilidad	0.4900
Lim. de confianza asintoticos	
95% Limite conf. inferior	0.2569
95% Limite conf. superior	0.9343
Limites conf. exactos	
95% Limite conf. inferior	0.2407
95% Limite conf. superior	0.9740

Tamano de la muestra = 279

```
                        Procedimiento GENMOD

                  Informacion de nivel de clase
             Clase           Niveles    Valores

             tratamiento           2    placebo ascorbic
             respuesta             2    gripe nogripe

              Criterios para valorar la bondad de ajuste

   Criterio                        DF           Valor        Valor/DF

   Desviacion                      1          4.8717          4.8717
   Desviacion escalada             1          4.8717          4.8717
   Chi-cuadrado de Pearson         1          4.8114          4.8114
   Pearson X2 escalado             1          4.8114          4.8114
   Verosimilitud log                        970.6299
   Verosimilitud log completa               -14.0019
   AIC (mejor mas pequeno)                   34.0038
   AICC (mejor mas pequeno)                     .
   BIC (mejor mas pequeno)                   32.1627

    Analisis de estimadores de parametros de maxima verosimilitud

                                           % de limites de confianza
                                                 Error 95de ratio de
   Parametro                 DF    Estimador    estandar verosimilitud

   Intercept                  1       4.7457       0.0891      4.5663

    Analisis de estimadores de parametros de maxima verosimilitud

                           % de limites
                            de confianza
                           95de ratio de  Chi-cuadrado
    Parametro             verosimilitud       de Wald    Pr > ChiSq

    Intercept                    4.9158         2836.81        <.0001

                        Procedimiento GENMOD

    Analisis de estimadores de parametros de maxima verosimilitud

                                           % de limites de confianza
                                                 Error 95de ratio de
   Parametro                 DF    Estimador    estandar verosimilitud

   tratamiento    placebo     1       0.0072       0.1197     -0.2277
   tratamiento    ascorbic    0       0.0000       0.0000      0.0000
   respuesta      gripe       1      -1.5712       0.1586     -1.8934
   respuesta      nogripe     0       0.0000       0.0000      0.0000
   Escala                     0       1.0000       0.0000      1.0000

    Analisis de estimadores de parametros de maxima verosimilitud


                             % de limites de confianza
                           95de ratio de  Chi-cuadrado
    Parametro             verosimilitud       de Wald    Pr > ChiSq

    tratamiento    placebo       0.2422            0.00        0.9523
    tratamiento    ascorbic      0.0000               .            .
    respuesta      gripe        -1.2702           98.11        <.0001
    respuesta      nogripe       0.0000               .            .
    Escala                       1.0000

NOTE: The scale parameter was held fixed.

                Estadisticos LR para analisis de tipo 3

                                       Chi-
             Fuente             DF    cuadrado    Pr > ChiSq

             tratamiento         1        0.00        0.9523
             respuesta           1      130.59        <.0001
```

Criterios para valorar la bondad de ajuste: da la información acerca la convergencia del algoritmo de máxima verosimilitud de los parámetros y el ajuste del modelo. Los valores obtenidos son iguales a los obtenidos con el contraste de la chi-cuadrado para independencia.

Analisis de estimadores de parametros de maxima verosimilitud: Se obtienen los estimadores de los parámetros. El procedimiento `PROC GENMOD` fija el último nivel de cada variable en cero.

$$\boxed{\hat{\lambda} = \mathbf{4.75},\ \hat{\lambda}_1^A = \mathbf{0.0072},\ \hat{\lambda}_2^A = \mathbf{0},\ \hat{\lambda}_1^B = \mathbf{-1.5712},\ \hat{\lambda}_2^B = \mathbf{0}}$$

Los recuentos estimados para cada casilla de tener gripe dado que se tomó vitamina C es, en este modelo:

$$\mu_{11} = \exp[\lambda + \lambda_2^A + \lambda_1^B] = \exp[4{,}745 + 0 - 1{,}5712] = 23{,}91$$

La razón de odds estimada de tener gripe es:

$$\exp[\lambda_1^B - \lambda_2^B] = \exp[-1{,}571 - 0] = 0{,}208$$

En *Estadisticos LR para analisis de tipo 3* se contrastan los efectos principales mediante la *razón de verosimilitudes* (LR).

La hipótesis nula H_0 es que no hay efectos principales, es decir que la distribución de las personas es igual a lo largo de los niveles de la respuesta. En este caso se rechaza la H_0 donde el estadístico `LR` es `130.59`, `df=1`, con un p-valor `<.0001`