Contajes (Sesión 2)

Modelos Lineales Generalizados

Grado de Estadística

23/11/2018

Índice

Clasificación de modelos

Explicative Variables	Response Variable				
	Dicothomic or Binary	*Polythomic*	*Counts (discrete)*	*Continuous*	
				Normal	*Time between events*
Dicothomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	Tests for 2 subpopulation means: t.test	Survival Analysis
Polythomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	ONEWAY, ANOVA	Survival Analysis
Continuous (covariates)	Logistic regression	*	Log-linear models	Multiple regression	Survival Analysis
Factors and covariates	Logistic regression	*	Log-linear models	Covariance Analysis	Survival Analysis
Random Effects	Mixed models	Mixed models	Mixed models	Mixed models	Mixed models

Introducción

- La relación entre los modelos log-lineales y los modelos de respuesta multinomial procede del hecho que la ley multinomial puede derivarse a partir de un conjunto de variables de Poisson condicionadas a un número total de observaciones fijado.
- El hecho de tener el número total de eventos fijado es lo que hace que estemos ante una distribución multinomial y no de Poisson.
- Ciertos modelos log-lineales son equivalentes a modelos de respuesta multinomial:
    - si parámetros de interés son los cocientes de las medias de las variables poissonianas
    - o equivalentemente, si los cocientes de las medias de Poisson respecto los totales
- Los modelos log-lineales vinculados a modelos multinomiales llevan un conjunto de parámetros molestos (nuisance parameters) vinculados a los totales parciales o totales de la tabla.
- Notas: no todos los modelos log-lineales son equivalentes a modelos multinomiales ni viceversa

Notación (I)

- $Y_1, \ldots, Y_L$: variables de Poisson independientes esperanzas de las anteriores variables
- Índices de tablas:
    - Filas: $i = 1, \ldots, I$
    - Columnas: $j = 1, \ldots, J$
    - Subtablas: $k = 1, \ldots, K$
- Factores: A, B, C, ...

Notación (II)

	FACTOR C											
	FACTOR B				FACTOR B				FACTOR B			
FACTOR A	C_1				...				C_K			
	B_1	...	B_J	TOTAL	B_1	...	B_J	TOTAL	B_1	...	B_J	TOTAL
A_1	Y_{111}	...	Y_{1J1}	Y_{1+1}	...	...	...	...	Y_{11K}	...	Y_{1JK}	Y_{1+K}
A_2	Y_{211}	...	Y_{2J1}	Y_{2+1}	...	...	...	...	Y_{21K}	...	Y_{2JK}	Y_{2+K}
...	...	...	...	...	...	...	...	...	...	...	...	...
A_I	Y_{I11}	...	Y_{IJ1}	Y_{I+1}	...	...	...	...	Y_{I1K}	...	Y_{IJK}	Y_{I+K}
TOTAL	Y_{+11}	...	Y_{+J1}	Y_{++1}	...	...	...	...	Y_{+1K}	...	Y_{+JK}	Y_{++K}

Total marginal univariante del factor A: $Y_{i++} = \sum_j \sum_k Y_{ijk}$.	Total marginal bivariante de los factores A y C: $Y_{i+k} = \sum_j Y_{ijk}$
Total marginal univariante del factor B: $Y_{+j+} = \sum_i \sum_k Y_{ijk}$	Total marginal bivariante de los factores B y C: $Y_{+jk} = \sum_i Y_{ijk}$
Total marginal univariante del factor C: $Y_{++k} = \sum_i \sum_j Y_{ijk}$	Total trivariante de los factores A, B y C: Y_{ijk} .
Total marginal bivariante de los factores A y B: $Y_{ij+} = \sum_k Y_{ijk}$	Total: $Y_{+++} = \sum_i \sum_j \sum_k Y_{ijk}$.

Tablas de 2 dimensiones (I x J)

- En tablas de dimensión 2, la hipótesis que las filas (A) y las columnas (B) son independientes puede formularse como que la probabilidad total es igual al producto de probabilidades marginales:

Valor fijado	**Total de observaciones**: μ	
H_0 (independencia)	Expresión	$\pi_{ij} = \pi_{i\cdot} \cdot \pi_{\cdot j}$ $E[Y_{ij}] = Y_{++} \cdot \pi_{i\cdot} \cdot \pi_{\cdot j}$
	Modelo	$log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j$
	Parámetros	$I + J - 1$
H_1 (dependencia)	Expresión	$\pi_{ij} \neq \pi_{i\cdot} \cdot \pi_{\cdot j}$ $E[Y_{ij}] \neq Y_{++} \cdot \pi_{i\cdot} \cdot \pi_{\cdot j}$
	Modelo	$log(\mu_{ij}) = \eta_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$
	Parámetros	$I \cdot J$

Tablas de 2 dimensiones (I x J)

- La relación (dependencia) entre los dos factores A y B puede resolverse realizando el contraste de las interacciones en el modelo log-lineal: una interacción significativa implica relación entre las variables.
- En el fondo, se realiza un contraste equivalente al que se realiza con el test de la $\chi 2$
- Observaciones para los modelos posteriores:
    - Los modelos log-lineales para el análisis de tablas de contingencia son jerárquicos, en el sentido que los términos de interacciones de orden superior, sólo se pueden incluir en el modelo si los términos de interacciones de orden inferior están presentes.
    - Los parámetros correspondientes a las constantes fijadas siempre deben incluirse en el modelo

Tablas de 3 dimensiones

- Total fijado (Ejemplo: encuesta sin cuotas sobre 3 factores)
  - Independencia total
  - Independencia por bloques
  - Independencia parcial
  - Asociación uniforme

- Total bivariante fijado (Ejemplo: encuesta con una cuota de hombres/mujeres)
  - Homogeneidad de filas para todas las subtablas
  - Homogeneidad por fila dentro de cada subtabla
  - Homogeneidad entre 2 factores para todas las combinaciones del otro factor

Tablas de 3 dim. Total fijado. Independencia total

- En tablas de dimensión 3, la hipótesis de **independencia total** entre las 3 respuestas - filas (A), columnas (B) y subtablas (C) con total fijado - se verifica a través del análisis de cualquier interacción.

Valor fijado	**Total de observaciones**: μ	
H_0 (independencia)	Expresión	$\pi_{ijk} = \pi_{i\cdot\cdot} \cdot \pi_{\cdot j\cdot} \cdot \pi_{\cdot\cdot k}$ $E[Y_{ijk}] = Y_{+++} \cdot \pi_{i\cdot\cdot} \cdot \pi_{\cdot j\cdot} \cdot \pi_{\cdot\cdot k}$
	Modelo	$log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k$
	Parámetros	$I + J + K - 2$
H_1 (dependencia)	Expresión	$\pi_{ijk} \neq \pi_{i\cdot\cdot} \cdot \pi_{\cdot j\cdot} \cdot \pi_{\cdot\cdot k}$ $E[Y_{ijk}] \neq Y_{+++} \cdot \pi_{i\cdot\cdot} \cdot \pi_{\cdot j\cdot} \cdot \pi_{\cdot\cdot k}$
	Modelo	Cualquiera con alguna interacción
Modelo saturado	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$
	Parámetros	$I \cdot J \cdot K$

Tablas de 3 dim. Total fijado. Independendencia total

- Ejemplo de independencia total

$C-$				$C+$				$Total$			
	B-	B+	Total		B-	B+	Total		B-	B+	Total
A-	7	181	188	A-	23	584	607	A-	30	765	795
A+	16	403	419	A+	53	1301	1354	A+	69	1704	1773
Total	23	584	607	Total	76	1885	1961	Total	99	2469	2568

$$\pi_{A_+} = \frac{1773}{2568} = 0.690$$

$$\pi_{B_+} = \frac{2469}{2568} = 0.961$$

$$\pi_{A_+B_+C_+} = \frac{1301}{2568} = 0.506 \simeq \pi_{A_+} \cdot \pi_{B_+} \cdot \pi_{C_+}$$

$$\pi_{C_+} = \frac{1961}{2568} = 0.764$$

Tablas de 3 dim. Total fijado. Independencia total (Ej. R)

- Modelo minimal

```
##
## Call:
## glm(formula = Y ~ A + B + C, family = poisson, data = df0)
##
## Deviance Residuals:
##         1          2          3          4          5          6
## -0.091312  -0.038955   0.024518   0.003539  -0.083749   0.111097
##         7          8
##  0.013191  -0.019941
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.98023    0.11058   17.91   <2e-16 ***
## AYes         0.80209    0.04268   18.79   <2e-16 ***
## BYes         3.21645    0.10250   31.38   <2e-16 ***
## CYes         1.17268    0.04645   25.25   <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: 3.8546e+03  on 7  degrees of freedom
## Residual deviance: 3.0397e-02  on 4  degrees of freedom
## AIC: 59.342
##
## Number of Fisher Scoring iterations: 3
```

- Modelo maximal

```
##
## Call:
## glm(formula = Y ~ A * B * C, family = poisson, data = df0)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.945910   0.377964   5.148 2.63e-07 ***
## AYes            0.826679   0.453163   1.824  0.06812 .
## BYes            3.252587   0.385204   8.444  < 2e-16 ***
## CYes            1.189584   0.431666   2.756  0.00585 **
## AYes:BYes      -0.026239   0.461913  -0.057  0.95470
## AYes:CYes       0.008119   0.517401   0.016  0.98748
## BYes:CYes      -0.018180   0.439969  -0.041  0.96704
## AYes:BYes:CYes -0.007571   0.527438  -0.014  0.98855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3.8546e+03  on 7  degrees of freedom
## Residual deviance: 2.7756e-13  on 0  degrees of freedom
## AIC: 67.312
##
## Number of Fisher Scoring iterations: 3
```

Tablas de 3 dim. Total fijado. Independencia por bloques

- En tablas de dimensión 3, la hipótesis de **independencia por bloques**, por ejemplo, del factor A (filas) de las otras 2 respuestas (columnas y subtablas) se verifica mediante la ausencia de otras interacciones que no sean las de estas 2 últimas respuestas.

Valor fijado	**Total de observaciones**: μ	
H_0 (ind. por bloques)	Expresión	$\pi_{ijk} = \pi_{i\cdot\cdot} \cdot \pi_{\cdot jk}$ $E[Y_{ijk}] = Y_{+++} \cdot \pi_{i\cdot\cdot} \cdot \pi_{\cdot jk}$
	Modelo	$log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \beta\gamma_{jk}$
	Parámetros	$I + JK - 1$
H_1 (dependencia)	Expresión	$\pi_{ijk} \neq \pi_{i\cdot\cdot} \cdot \pi_{\cdot jk}$ $E[Y_{ijk}] \neq Y_{+++} \cdot \pi_{i\cdot\cdot} \cdot \pi_{\cdot jk}$
	Modelo	Cualquiera con alguna interacción excepto BC
Modelo saturado	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$
	Parámetros	$I \cdot J \cdot K$

Tablas de 3 dim. Total fijado. Independencia por bloques (Ej)

- Ejemplo de independencia por bloques. A es independiente de B y C si no se considera C y B respectivamente.

$C-$				$C+$				$Total$			
	B-	B+	Total		B-	B+	Total		B-	B+	Total
A-	13	67	80	A-	1	13	14	A-	14	80	94
A+	30	148	178	A+	1	30	31	A+	31	178	209
Total	43	215	258	Total	2	43	45	Total	45	258	303

$$\pi_{A_+} = \frac{209}{303} = 0.690$$

$$\pi_{B_+} = \frac{258}{303} = 0.851$$

$$\pi_{A_+B_+C_+} = \frac{30}{303} = 0.099 \neq \pi_{A_+} \cdot \pi_{B_+} \cdot \pi_{C_+}$$

$$\pi_{C_+} = \frac{45}{303} = 0.149$$

$$\pi_{A_+B_+C_+} = \frac{30}{303} = 0.099 \simeq \pi_{A_+} \cdot \pi_{B_+C_+}$$

$$\pi_{B_+C_+} = \frac{43}{303} = 0.142$$

Tablas de 3 dim. Total fijado. Independencia por bloques (Ej)

- A es independiente de B sin considerar C:

$$OR_{AB} = \frac{14 \cdot 178}{31 \cdot 80} \approx 1$$

- A es independiente de C sin considerar B:

$$OR_{AB} = \frac{80 \cdot 31}{178 \cdot 14} \approx 1$$

- A es independiente de B y C conjuntamente

	B- C-	B+ C-	B- C+	B+ C+	*Total*
A-	13	67	1	13	94
A+	30	148	1	30	209
Total	43	215	2	43	303

Expected table under independence

```
##          [,1]      [,2]     [,3]     [,4]
## [1,] 13.33993  66.69967 0.620462 13.33993
## [2,] 29.66007 148.30033 1.379538 29.66007
```

- Modelo minimal

```
##
## Call:
## glm(formula = Y ~ A + B * C, family = poisson, data = df0)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
## -0.09347   0.06230   0.03675  -0.02467   0.44216  -0.33997  -0.09347
##        8
##  0.06230
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.5908     0.1749  14.812  < 2e-16 ***
## AYes          0.7990     0.1242   6.434 1.24e-10 ***
## BYes          1.6094     0.1671   9.634  < 2e-16 ***
## CYes         -3.0681     0.7234  -4.241 2.22e-05 ***
## BYes:CYes     1.4586     0.7424   1.965   0.0494 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 381.74419  on 7  degrees of freedom
## Residual deviance:   0.33828  on 3  degrees of freedom
## AIC: 46.54
##
## Number of Fisher Scoring iterations: 4
```

Tablas de 3 dim. Total fijado. Independencia por bloques (Ej. R)

- Modelo maximal

```
##
## Call:
## glm(formula = Y ~ A * B * C, family = poisson, data = df0)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.56495    0.27735   9.248  < 2e-16 ***
## AYes           0.83625    0.33205   2.518   0.0118 *
## BYes           1.63974    0.30307   5.411 6.28e-08 ***
## CYes          -2.56495    1.03774  -2.472   0.0134 *
## AYes:BYes     -0.04373    0.36323  -0.120   0.9042
## AYes:CYes     -0.83625    1.45267  -0.576   0.5648
## BYes:CYes      0.92521    1.08109   0.856   0.3921
## AYes:BYes:CYes 0.87998    1.49739   0.588   0.5568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3.8174e+02  on 7  degrees of freedom
## Residual deviance: 5.0626e-14  on 0  degrees of freedom
## AIC: 52.202
##
## Number of Fisher Scoring iterations: 3
```

Tablas de 3 dim. Total fijado. Independencia parcial o condicional

- En tablas de dimensión 3, la hipótesis de **independencia parcial entre 2 factores**, por ejemplo A (filas) y B (columnas) se verificaría mediante el contraste de la interacción de A con B y la interacción de orden 2.

Valor fijado	**Total de observaciones**: μ	
H_0 (ind. parcial)	Expresión	$\pi_{ij\cdot} = \pi_{i\cdot\|k} \cdot \pi_{\cdot j\|k} \quad \forall i,j,k$ $E[Y_{ijk}] = Y_{+++} \cdot \pi_{i\cdot\|k} \cdot \pi_{\cdot j\|k}$
	Modelo	$log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk}$
	Parámetros	$(I + J - 1) \cdot K$
H_1 (dependencia)	Expresión	$\pi_{ij\cdot} \neq \pi_{i\cdot\|k} \cdot \pi_{\cdot j\|k} \quad \exists i,j,k$ $E[Y_{ijk}] \neq Y_{+++} \cdot \pi_{i\cdot\|k} \cdot \pi_{\cdot j\|k}$
	Modelo	Cualquiera con interacción distinta a AC y BC
Modelo saturado	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$
	Parámetros	$I \cdot J \cdot K$

Tablas de 3 dim. Total fijado. Independencia parcial o condicional

- Conceptualmente, implica que la relación entre 2 (A y B) de las 3 variables viene explicada por una tercera (C):
- Ejemplos:

$$A \leftarrow \mathbf{C} \rightarrow \mathbf{B}$$

$$A \rightarrow \mathbf{C} \leftarrow \mathbf{B}$$

$$A \rightarrow \mathbf{C} \rightarrow \mathbf{B}$$

$$A \leftarrow \mathbf{C} \leftarrow \mathbf{B}$$

- Condicionado a C (ya sea C+ o C-) A y B son independientes.
- La relación entre llevar habitualmente una cajetilla de tabaco encima (A) y tener cáncer de pulmón (B) viene explicado por fumar (C). Condicionado a fumar (C), llevar una cajetilla de tabaco (A) y tener cáncer de pulmón (B) son independientes.

Tablas de 3 dim. Total fijado. Independencia parcial o condicional.

- Condicionado a C, A y B son independientes

$C-$				$C+$				$Total$			
	B-	B+	Total		B-	B+	Total		B-	B+	Total
A-	30	6	36	A-	3	20	23	A-	33	26	59
A+	40	8	48	A+	22	148	170	A+	62	150	212
Total	70	24	94	Total	25	168	193	Total	95	176	271

- Es equivalente mirar las probabilidades que los ORs:

$$OR_{AB|C+} = \frac{30 \cdot 8}{6 \cdot 40} = 1$$

$$OR_{AB|C-} = \frac{3 \cdot 148}{20 \cdot 22} \approx 1$$

$$OR_{AB} = \frac{33 \cdot 150}{62 \cdot 26} = 3.07 \neq 1$$

$$OR_{AC} = \frac{36 \cdot 48}{23 \cdot 170} = 5.54$$

$$OR_{BC} = \frac{70 \cdot 168}{94 \cdot 25} = 5.00$$

Tablas de 3 dim. Total fijado. Asociación uniforme

- En tablas de dimensión 3, la hipótesis de **asociación uniforme para cualquiera 2 factores condicionado a un tercero** se verificaría mediante el contraste de la interacción de orden 2 (es una generalización de la independencia parcial)

Valor fijado	**Total de observaciones**: μ	
H_0 (asoc. unif.)	Expresión	$\pi_{ij\cdot} = \pi_{i\cdot\vert k} \cdot \pi_{\cdot j\vert k} \quad \forall i,j,k$ $\pi_{i\cdot k} = \pi_{i\cdot\vert j} \cdot \pi_{\cdot k\vert j} \quad \forall i,j,k$ $\pi_{jk\cdot} = \pi_{j\cdot\vert i} \cdot \pi_{\cdot k\vert i} \quad \forall i,j,k$
	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$
	Parámetros	$IJK - (I-1)\cdot(J-1)\cdot(K-1)$
H_1 (dependencia)	Expresión	$\pi_{ij\cdot} \neq \pi_{i\cdot\vert k} \cdot \pi_{\cdot j\vert k} \quad \exists i,j,k$ $E[Y_{ijk}] \neq Y_{+++} \cdot \pi_{i\cdot\vert k} \cdot \pi_{\cdot j\vert k}$
	Modelo	Modelo saturado
Modelo saturado	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$
	Parámetros	$I \cdot J \cdot K$

Tablas de 3 dim. Total fijado. Asociación uniforme

- Condicionado a C, A y B son independientes
- Condicionado a B, A y C son independientes
- Condicionado a A, B y C son independientes
- A, B y C no son independientes

Tablas de 3 dim. Total bivariado fijado. Homogeneidad por fila comunes a las subtablas

- En tablas de dimensión 3, la hipótesis de homogeneidad o probabilidades idénticas por fila comunes a todas las subtablas (probabilidad marginal univariante igual a probabilidad condicional) se verifica mediante el análisis de las interacciones simples

Valor fijado	**Total bivariado**: μ	
H_0 (homogeneidad)	Expresión	$\pi_{j\|ik} = \pi_{.j.} \quad \forall i,j,k$ $E[Y_{ijk}] = Y_{i+k} \cdot \pi_{.j.}$
	Modelo	$log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik}$
	Parámetros	$IK + J - 1$
H_1 (no homogeneidad)	Expresión	$\pi_{j\|ik} \neq \pi_{.j.} \quad \exists i,j,k$ $E[Y_{ijk}] \neq Y_{i+k} \cdot \pi_{.j.}$
	Modelo	Cualquiera con interacción distinta a AC
Modelo saturado	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$
	Parámetros	$I \cdot J \cdot K$

Tablas de 3 dim. Total bivariado fijado. Homogeneidad de cada fila dentro de cada subtabla

- En tablas de dimensión 3, la hipótesis de homogeneidad, probabilidades idénticas por filas dentro de cada subtabla, donde la variable de respuesta es la columna, factor B, las variables explicativas son los factores A y C con totales bivariantes según A y C fijados (la función de probabilidad conjunta es por tanto, producto de probabilidades multinomiales)

Valor fijado	**Total bivariado**: μ	
H_0 (homogeneidad)	Expresión	$\pi_{ijk} = \pi_{i\cdot k} \cdot \pi_{\cdot jk} \qquad \forall i,j,k$ $E[Y_{ijk}] = Y_{i+k} \cdot \pi_{i\cdot k} \cdot \pi_{\cdot jk}$
	Modelo	$log(\mu_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk}$
	Parámetros	$K \cdot (I + J - 1)$
H_1 (no homogeneidad)	Expresión	$\pi_{ijk} \neq \pi_{i\cdot k \cdot \pi_{\cdot jk}} \qquad \exists i,j,k$ $E[Y_{ijk}] \neq Y_{i+k} \cdot \pi_{i\cdot k} \cdot \pi_{\cdot jk}$
	Modelo	Cualquiera con interacción distinta a AC y BC
Modelo saturado	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$
	Parámetros	$I \cdot J \cdot K$

- La hipótesis de homogeneidad, asociación entre el factor C y B es la misma para todos los niveles de A-B (probabilidad marginal bivariante de C y B idéntica, para cada grupo de A-B)

Valor fijado	**Total de observaciones**: μ	
H_0 (Homogeneidad)	Expresión	$\pi_{ijk} = \pi_{ij\cdot} \cdot \pi_{i\cdot k} \cdot \pi_{\cdot jk} \qquad \forall i,j,k$ $E[Y_{ijk}] = Y_{i+k} \cdot \pi_{ij\cdot} \cdot \pi_{i\cdot k} \cdot \pi_{\cdot jk}$
	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$
	Parámetros	$(I-1)\cdot(J-1)\cdot(K-1)$
H_1 (Heterogeneidad)	Expresión	$\pi_{ijk} \neq \pi_{ij\cdot} \cdot \pi_{i\cdot k} \cdot \pi_{\cdot jk} \qquad \forall i,j,k$ $E[Y_{ijk}] \neq Y_{i+k} \cdot \pi_{ij\cdot} \cdot \pi_{i\cdot k} \cdot \pi_{\cdot jk}$
	Modelo	Modelo saturado
Modelo saturado	Modelo	$log(\mu_{ijk}) = \eta_{ijk} =$ $\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$
	Parámetros	$I \cdot J \cdot K$

Tablas de 3 dim. Relación logística vs Modelos Log-lineales

- Supongase que el factor **B** es la **respuesta dicotómica** y los Factores **A** y **C** las **variables explicativas** (totales bivariantes A y C fijados)

Modelos log-lineales	*Regresión logística*
$AC + B$	$1 \quad (Minimal)$
$AC + AB$	A
$AC + BC$	C
$AC + AB + BC$	$A + C$
ABC	$AC \quad (Maximal)$

- La relación viene dada por:

$$log(\mu_{ij}) = \mu + \theta_i + \alpha_j + x_i^T \beta_j$$

$$log(\mu_{iJ}) = \mu + \theta_i + \alpha_J + x_i^T \beta_J$$

$$log(\mu_{ij}) - log(\mu_{iJ}) = log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = (\alpha_j - \alpha_J) + x_i^T (\beta_j - \beta_J)$$

Validación. Diagnosis

- Estadístico de devianza. Si el modelo es correcto, para muestras grandes tiene una distribución χ^2 con grados de libertad calculados como la diferencia entre el número de celdas no nulas menos el número de parámetros independientes del modelo:

$$D = 2 \sum y_i \cdot log\left(\frac{y_i}{\hat{\mu}_i}\right) \sim \chi^2$$

- Los residuos estandarizados de Pearson extremos tendrán valores superiores a 2 o 3 desviaciones típicas.

- Un grupo de 4991 estudiantes de secundaria de Wisconsin se clasifican en la siguiente tabla de contingencia según su ESTATUS socio-económico (A, con 4 niveles), la MOTIVACIÓN recibida de los padres en sus estudios (C, 2 niveles BAJO-ALTO) y sus PLANES de continuación en la Universidad (B, 2 niveles SI-NO). Se consideran las 3 variables como respuesta. Datos de Fienberg (1977)

	FACTOR C- Motivación (E)					
	FACTOR B - Universidad?			FACTOR B Universidad?		
FACTOR A	C_1 - Bajo			$C_{K=2}$ Alto		
Estatus Social	B_1 No	$B_{J=2}$ Si	TOTAL	B_1 No	$B_{J=2}$ Si	TOTAL
A_1 Bajo	749	35	784	233	133	366
A_2 Medio- Bajo	627	38	665	330	303	633
A_2 Medio- Alto	420	37	457	374	467	841
$A_{I=4}$ Alto	153	26	179	266	800	1066
TOTAL	1949	136	2085	1203	1703	2906

Ejemplo. Datos (II)

```
##              A  B    C Freq
## 1         Bajo No Baja  749
## 2         Bajo No Alta  233
## 3         Bajo Si Baja   35
## 4         Bajo Si Alta  133
## 5   Medio-Bajo No Baja  627
## 6   Medio-Bajo No Alta  330
## 7   Medio-Bajo Si Baja   38
## 8   Medio-Bajo Si Alta  303
## 9   Medio-Alto No Baja  420
## 10  Medio-Alto No Alta  374
## 11  Medio-Alto Si Baja   37
## 12  Medio-Alto Si Alta  467
## 13        Alto No Baja  153
## 14        Alto No Alta  266
## 15        Alto Si Baja   26
## 16        Alto Si Alta  800
```

Ejemplo. Comparación de modelos

- La siguiente tabla contiene los modelos ajustados, su devianza y su interpretación
  - A: **ESTATUS** socio-económico
  - B: Planes para la **UNIVERSIDAD**
  - C: **MOTIVACIÓN** de los padres
- ¿Qué modelo se escogería?

Modelo	Devianza	GL	Intrepretacion
A + B + C	2714	10	Motivación, Universidad y Estatus social independientes
A + B * C	1092	9	Estatus social es independiente de la Motivación y Universidad
B + A * C	1877	7	Asistencia a Universidad es independiente de Motivación y Estatus
C + A * B	1920	7	Motivación de los padres es independiente de Estatus y Universidad
A * B + A * C	1084	4	Condicionado al Estatus, Motivación y Universidad son independientes
A * B + B * C	298	6	Condicionado a Universidad, Estatus y Motivación son independientes
A * C + B * C	255	6	Condicionado a Motivación, Estatus y Universidad son independientes
A * B + A * C + B * C	2	3	Las 3 previas juntas
A * B * C	0	0	Nada es independiente

Ejemplo. Modelo seleccionado: AB + AC + BC

```
##
## Call:
## glm(formula = get(paste0("form", i)), family = "poisson", data = d)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
## -0.15119   0.27320   0.73044  -0.35578   0.04135  -0.05691  -0.16639
##        8         9        10        11        12        13        14
##  0.05952  -0.04446   0.04719   0.15116  -0.04217   0.32807  -0.24539
##       15        16
## -0.75147   0.14245
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        6.62426    0.03602 183.918  < 2e-16 ***
## AMedio-Bajo       -0.18496    0.05304  -3.487 0.000489 ***
## AMedio-Alto       -0.58183    0.05931  -9.810  < 2e-16 ***
## AAlto             -1.62046    0.08450 -19.178  < 2e-16 ***
## BSi               -3.19497    0.11850 -26.962  < 2e-16 ***
## CAlta             -1.19117    0.07166 -16.622  < 2e-16 ***
## AMedio-Bajo:BSi    0.42013    0.11768   3.570 0.000357 ***
## AMedio-Alto:BSi    0.73851    0.11382   6.488 8.69e-11 ***
## AAlto:BSi          1.59311    0.11527  13.820  < 2e-16 ***
## AMedio-Bajo:CAlta  0.55410    0.09469   5.852 4.87e-09 ***
## AMedio-Alto:CAlta  1.07056    0.09649  11.095  < 2e-16 ***
## AAlto:CAlta        1.78588    0.11444  15.606  < 2e-16 ***
## BSi:CAlta          2.68292    0.09867  27.191  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

Ejemplo. Modelo maximal. Interpretación (Ejercicio)

```
##
## Call:
## glm(formula = get(paste0("form", i)), family = "poisson", data = d)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            6.61874    0.03654 181.141  < 2e-16 ***
## AMedio-Bajo           -0.17779    0.05413  -3.285  0.00102 **
## AMedio-Alto           -0.57848    0.06096  -9.490  < 2e-16 ***
## AAlto                 -1.58830    0.08872 -17.903  < 2e-16 ***
## BSi                   -3.06339    0.17294 -17.714  < 2e-16 ***
## CAlta                 -1.16770    0.07501 -15.567  < 2e-16 ***
## AMedio-Bajo:BSi        0.26003    0.24045   1.081  0.27951
## AMedio-Alto:BSi        0.63405    0.24355   2.603  0.00923 **
## AAlto:BSi              1.29105    0.27369   4.717 2.39e-06 ***
## AMedio-Bajo:CAlta      0.52585    0.10125   5.193 2.06e-07 ***
## AMedio-Alto:CAlta      1.05170    0.10335  10.176  < 2e-16 ***
## AAlto:CAlta            1.72076    0.12618  13.637  < 2e-16 ***
## BSi:CAlta              2.50270    0.20425  12.253  < 2e-16 ***
## AMedio-Bajo:BSi:CAlta  0.21530    0.27561   0.781  0.43469
## AMedio-Alto:BSi:CAlta  0.14871    0.27557   0.540  0.58945
## AAlto:BSi:CAlta        0.37076    0.30286   1.224  0.22088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3.211e+03  on 15  degrees of freedom
## Residual deviance: 5.107e-14  on  0  degrees of freedom
```

Basandote en el modelo maximal:

1. Interpreta la intercept
2. Interpreta el coeficiente: *AAlto*
3. Interpreta el coeficiente: *AAlto:CAlta*
4. Interpreta el coeficiente: *AAlto:BSi:CAlta*

Ejemplo. Modelo maximal. Interpretación (Solución)

1. La exponencial de la ***intercept*** representa el número de individuos en el nivel de referencia: $exp(-0.17779) = 753$
2. La exponencial de ***AAlto*** representa el cociente entre el número de individuos en este nivel respecto al nivel de referencia dentro de los alumnos situados en las categorías de referencia de B (No) y C (Bajo):

$$exp(-1.58830) = \frac{153}{749} = 0.20427$$

3. La exponencial de ***AAlto:CAlta*** representa, para el nivel de referencia de B, cuanto mayor es el ratio de los que tienen factor *Alta* en el nivel C respecto a la referencia del nivel C comparando el mismo ratio entre los niveles *Alto* y *Bajo* del nivel A:

$$exp(1.72076) = \frac{266/153}{233/749} = 5.5887$$

4. La exponencial de ***AAlto:BSi:CAlta*** representa cuanto mayor es el anterior ratio en el nivel Alto de A respecto al de referencia:

$$exp(0.37076) = \frac{\frac{800/26}{133/35}}{\frac{266/153}{233/749}} = 1.4488$$

Ejemplo. Estupefacientes. Datos

- Datos
  - ***cigarette***: yes/no
  - ***marijuana***: yes/no
  - ***alcohol***: yes/no
- Tabla en diferentes formatos:

```
## , , alcohol = yes
##
##           marijuana
## cigarette yes  no
##       yes 911 538
##       no   44 456
##
## , , alcohol = no
##
##           marijuana
## cigarette yes  no
##       yes   3  43
##       no    2 279
```

```
##                   marijuana yes  no
## alcohol cigarette
## yes     yes                 911 538
##         no                   44 456
## no      yes                   3  43
##         no                    2 279
```

Ejemplo. Estupefacientes. Inspeccionar datos

- El hecho de fumar y consumir marihuana está relacionado tanto en los que beben alcohol como en los que no. Vemos que los estudiantes que probaron cigarrillos y alcohol, el 62% también probó marihuana. Del mismo modo, de aquellos estudiantes que no probaron cigarrillos ni alcohol, el 99% también no probó marihuana. Definitivamente parece que hay alguna relación.

```
## , , alcohol = yes
##
##          marijuana
## cigarette       yes        no
##       yes 0.6287095 0.3712905
##       no  0.0880000 0.9120000
##
## , , alcohol = no
##
##          marijuana
## cigarette         yes        no
##       yes 0.065217391 0.9347826
##       no  0.007117438 0.9928826
```

Ejemplo. Estupefacientes. Hipotesi: Independencia total

- $H_0 : pi_{ijk} = pi_i \cdot pi_j \cdot pi_k \leftrightarrow$ Independencia Total

```
##                 Estimate Std. Error   z value      Pr(>|z|)
## (Intercept)    4.1725378 0.06495836 64.234043  0.000000e+00
## cigaretteyes   0.6493063 0.04415087 14.706534  5.852911e-49
## marijuanayes  -0.3154188 0.04244454 -7.431316  1.075222e-13
## alcoholyes     1.7851115 0.05975887 29.871911 4.559915e-196
```

```r
1-pchisq(mod0$deviance,mod0$df.residual) # p-valor
```

```
## [1] 0
```

- Mirando el *summary* parece que este es un gran modelo ya que hay coeficientes muy significativos con p-valores cercanos a 0, pero tenemos que mirar la desvianza residual. Se rechaza H_0 de independencia total ya que obtenemos un p-valor para contrastar la validez del modelo cercano a cero y una deviança residual (1286) muy por encima del punto crítico (9.5). Debemos probar la independencia por bloques

Ejemplo. Estupefacientes. Valores predichos vs. observados

- Se confirma que no es un gran ajuste

```
cbind(mod0$data, fitted(mod0))
```

```
##   cigarette marijuana alcohol Freq fitted(mod0)
## 1       yes       yes     yes  911    539.98258
## 2        no       yes     yes   44    282.09123
## 3       yes        no     yes  538    740.22612
## 4        no        no     yes  456    386.70007
## 5       yes       yes      no    3     90.59739
## 6        no       yes      no    2     47.32880
## 7       yes        no      no   43    124.19392
## 8        no        no      no  279     64.87990
```

Ejemplo. Estupefacientes. Interpretación

- El odd de haber consumido marihuana coincide con la odd manualmente calculada

```r
exp(coef(mod0)['marijuanayes'])
```

```
## marijuanayes
##    0.7294833
```

```r
pt <- with(seniors.df,tapply(Freq,marijuana,sum))
pt[2]/pt[1]
```

```
##       yes
## 0.7294833
```

Ejemplo. Estupefacientes. Hipotesis: Independencia por bloques

- $H_0 : pi_{ijk} = pi_j \cdot pi_{ik} \leftrightarrow$ Independència por bloques

```
##                          Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)             5.0905320 0.06228346  81.731679 0.000000e+00
## cigaretteyes           -1.8097133 0.15905298 -11.378054 5.378681e-30
## alcoholyes              0.5762534 0.07455681   7.729051 1.083512e-14
## marijuanayes           -0.3154188 0.04244461  -7.431304 1.075320e-13
## cigaretteyes:alcoholyes  2.8737341 0.16729609  17.177534 3.911936e-66
```

```
1-pchisq(mod1a$deviance,mod1a$df.residual)
```

```
## [1] 0
```

- Rechazo H_0 de independencia por bloques de marihuana

Ejemplo. Estupefacientes. Hipotesis: Independencia por bloques

- $H_0 : pi_{ijk} = pi_j \cdot pi_{ik} \leftrightarrow$ Independència por bloques

```
##                          Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)             4.7049519 0.06282180 74.893617 0.000000e+00
## cigaretteyes            0.6493063 0.04415071 14.706589 5.848207e-49
## marijuanayes           -4.1651136 0.45067171 -9.242013 2.419038e-20
## alcoholyes              1.1271857 0.06412166 17.578860 3.576938e-69
## marijuanayes:alcoholyes 4.1250878 0.45294386  9.107283 8.447108e-20
```

```
1-pchisq(mod1b$deviance,mod1b$df.residual)
```

```
## [1] 0
```

- Rechazo H_0 de independencia por bloques de cigarros

Ejemplo. Estupefacientes. Hipotesis: Independencia por bloques

- $H_0 : pi_{ijk} = pi_j \cdot pi_{ik} \leftrightarrow$ Independencia por bloques

```
##                          Estimate Std. Error    z value      Pr(>|z|)
## (Intercept)             4.6596559 0.06308154  73.867189  0.000000e+00
## cigaretteyes           -0.2351197 0.05551319  -4.235385  2.281605e-05
## marijuanayes           -2.7712291 0.15198577 -18.233477  2.799415e-74
## alcoholyes              1.7851115 0.05975941  29.871639 4.597227e-196
## cigaretteyes:marijuanayes  3.2243089 0.16098117  20.029106  3.071201e-89
```

```
1-pchisq(mod1c$deviance,mod1c$df.residual)
```

```
## [1] 0
```

- Rechazo H_0 de independencia por bloques de alcohol

Ejemplo. Estupefacientes. Hipótesis: Independencia Parcial

- $H_0 : pi_{ijk} = pi_{ik} \cdot pi_{jk} \leftrightarrow$ Independencia parcial

```
##                          Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)            5.62294604 0.06005168  93.635111 0.000000e+00
## marijuanayes          -4.16511363 0.45066572  -9.242136 2.416258e-20
## cigaretteyes          -1.80971327 0.15905294 -11.378056 5.378524e-30
## alcoholyes            -0.08167244 0.07809686  -1.045784 2.956608e-01
## marijuanayes:alcoholyes 4.12508777 0.45293789   9.107403 8.437777e-20
## cigaretteyes:alcoholyes 2.87373412 0.16729594  17.177548 3.910951e-66
```

```
1-pchisq(mod2a$deviance,mod2a$df.residual)
```

```
## [1] 0
```

- Rechazo H_0 de independencia parcial mediada por alcohol

Ejemplo. Estupefacientes. Hipótesis: Independencia Parcial

- $H_0 : pi_{ijk} = pi_{ij} \cdot pi_{ik} \leftrightarrow$ Independencia parcial

```
##                            Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)               5.5776500 0.06032291  92.463214 0.000000e+00
## marijuanayes             -2.7712291 0.15198577 -18.233477 2.799425e-74
## alcoholyes                0.5762534 0.07455682   7.729051 1.083513e-14
## cigaretteyes             -2.6941394 0.16257385 -16.571788 1.114624e-61
## marijuanayes:cigaretteyes 3.2243089 0.16098117  20.029106 3.071215e-89
## alcoholyes:cigaretteyes   2.8737341 0.16729609  17.177534 3.911936e-66
```

```r
1-pchisq(mod2b$deviance,mod2b$df.residual)
```

```
## [1] 0
```

- Rechazo H_0 de independencia parcial mediada por tabaco

Ejemplo. Estupefacientes. Hipótesis: Independencia Parcial

- $H_0 : pi_{ijk} = pi_{ij} \cdot pi_{jk} \leftrightarrow$ Independencia parcial

```
##                             Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)                5.1920699 0.06087902  85.285042 0.000000e+00
## cigaretteyes              -0.2351197 0.05551319  -4.235385 2.281604e-05
## alcoholyes                 1.1271857 0.06412196  17.578777 3.582175e-69
## marijuanayes              -6.6209239 0.47370455 -13.976906 2.156623e-44
## cigaretteyes:marijuanayes  3.2243089 0.16098115  20.029108 3.071103e-89
## alcoholyes:marijuanayes    4.1250878 0.45293602   9.107440 8.434840e-20
```

```r
1-pchisq(mod2c$deviance,mod2c$df.residual)
```

```
## [1] 0
```

- Rechazo H_0 de independencia parcial mediada por marihuana

Ejemplo. Estupefacientes. Hipótesis: Asociación uniforme

```
##                               Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)                   5.633420 0.05970084  94.360822 0.000000e+00
## cigaretteyes                 -1.886669 0.16269698 -11.596213 4.307172e-31
## marijuanayes                 -5.309042 0.47519695 -11.172299 5.571964e-29
## alcoholyes                    0.487719 0.07576720   6.437073 1.217997e-10
## cigaretteyes:marijuanayes     2.847889 0.16383940  17.382200 1.125516e-67
## cigaretteyes:alcoholyes       2.054534 0.17406432  11.803304 3.752817e-32
## marijuanayes:alcoholyes       2.986014 0.46467793   6.425987 1.310164e-10
```

```
1-pchisq(mod3$deviance,mod3$df.residual)
```

```
## [1] 0.5408396
```

- Acepto $H_0 \rightarrow$ la tabla de contingencia es consistente con la asociación uniforme.

Ejemplo. Estupefacientes. Valors predichos vs. observados

```
##   cigarette marijuana alcohol Freq fitted(mod3)
## 1       yes       yes     yes  911    910.38317
## 2        no       yes     yes   44     44.61683
## 3       yes        no     yes  538    538.61683
## 4        no        no     yes  456    455.38317
## 5       yes       yes      no    3      3.61683
## 6        no       yes      no    2      1.38317
## 7       yes        no      no   43     42.38317
## 8        no        no      no  279    279.61683
```

Ejemplo. Estupefacientes. Interpretación

- Los estudiantes que probaron la marihuana tienen odds estimadas de haber probado el alcohol que son 19 veces superiores a las odds de los estudiantes que no probaron la marihuana. Y el hecho de que no aparezca la interacción triple nos dice que este hecho es independiente de si han probado el tabaco o no.

```
exp(coef(mod3)["marijuanayes:alcoholyes"])
```

```
## marijuanayes:alcoholyes
##                19.80658
```