Clase 3: Independencia Condicional:

(**AC**, **BC**), (AB, AC), (AB, BC)

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

Sólo un par de variables son condicionalmente independientes.

- Dado C, A y B son condicionalmente independientes
- La tabla se puede colapsar en A y B.
- Asociación marginal A–C = Asociación parcial A–C
- Asociación marginal B–C = Asociación parcial B–C
- Asociación parcial A–B es nula. Asociación marginal = ?????.

Este es un modelo en el que la asociación marginal A–B puede ser espúrea si uno ignora la variable C.

Grados de Libertad

$$\begin{aligned} df &= IJK - [1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) \\ &\quad + (J-1)(K-1)] \\ &= K(I-1)(J-1) \end{aligned}$$

Ejemplo

Volvemos al ejemplo del estudio longitudinal. Examinaremos el modelo en el que Nivel de Colesterol y Presión Diastólica son independientes dada la personalidad. Esto corresponde al modelo

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC}$$

Testear la validez de este modelo es testear simultáneamente de independencia entre las tablas:

n_{1kj}	Personalidad I	
	Presión Diastólica	
Colesterol	Normal	Alto
Normal	716	79
Alto	207	25

n_{2kj}	Personalidad II	
	Presión Diastólica	
Colesterol	Normal	Alto
Normal	819	67
Alto	186	22

donde cada tabla tiene $(2-1)(2-1) = 1$ grados de libertad.

Los estadísticos de bondad de ajuste obtenidos son:

$$X^2 = 2.188 \qquad G^2 = 2.062$$

Los grados de libertad de los estadísticos de bondad de ajuste serán $df = 2(2-1)(2-1) = 2$ y por lo tanto obtenemos un ajuste razonable ya que el p–valor de cada uno es 0.3348743 y 0.3566501, respectivamente.

Veamos la salida de LEM para este ejemplo.

```
LEM: log-linear and event history analysis with missing data.
Developed by Jeroen Vermunt (c), Tilburg University, The Netherlands.
Version 1.0 (September 18, 1997).

*** INPUT ***
  * A = Personalidad
  * B = Colesterol
  * C = Presión Diastólica

  man 3
  dim 2 2 2
  lab A B C
  mod {AB,AC}
  dat [ 716 79
        207 25
        819 67
        186 22]

*** STATISTICS ***

  Number of iterations = 2
  Converge criterion   = 0.0000000000
  X-squared            = 2.1876 (0.3349)
```

```
L-squared            = 2.0626 (0.3565)
Cressie-Read         = 2.1439 (0.3423)
Dissimilarity index  = 0.0062
Degrees of freedom   = 2
Log-likelihood       = -3195.22085
Number of parameters = 5 (+1)
Sample size          = 2121.0
BIC(L-squared)       = -13.2567
AIC(L-squared)       = -1.9374
BIC(log-likelihood)  = 6428.7399
AIC(log-likelihood)  = 6400.4417

*** FREQUENCIES ***

A B C     observed  estimated  std. res.
1 1 1     716.000    714.494      0.056
1 1 2      79.000     80.506     -0.168
1 2 1     207.000    208.506     -0.104
1 2 2      25.000     23.494      0.311
2 1 1     819.000    813.921      0.178
2 1 2      67.000     72.079     -0.598
2 2 1     186.000    191.079     -0.367
2 2 2      22.000     16.921      1.235
```

```
*** LOG-LINEAR PARAMETERS ***
* TABLE ABC [or P(ABC)] *
  effect         beta     std err  z-value   exp(beta)   Wald  df  prob
  main           4.8147                     123.3082
  A
   1             0.0495   0.0410    1.207      1.0507
   2            -0.0495                      0.9517     1.46   1 0.228
  B
   1             0.6702   0.0268   24.991      1.9546
   2            -0.6702                      0.5116   624.56   1 0.000
  C
   1             1.1518   0.0379   30.423      3.1640
   2            -1.1518                      0.3161   925.54   1 0.000
  AB
   1 1          -0.0544   0.0268   -2.029      0.9471
   1 2           0.0544                      1.0559
   2 1           0.0544                      1.0559
   2 2          -0.0544                      0.9471     4.11   1 0.043
  AC
   1 1          -0.0602   0.0379   -1.591      0.9416
   1 2           0.0602                      1.0621
   2 1           0.0602                      1.0621
   2 2          -0.0602                      0.9416     2.53   1 0.112
```

Clase 4: Asociación Homogénea:

(**AB**, **AC**, **BC**)

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

Todos los pares de variables son condicionalmente dependientes dada la tercera. La asociación parcial entre dos variables no varía con los distintos niveles de la tercera.

- ¿Podemos colapsar en alguna dirección?

Grados de Libertad

$$\begin{aligned} df &= IJK - [1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) \\ &\quad + (I-1)(K-1) + (J-1)(K-1)] \\ &= (I-1)(J-1)(K-1) \end{aligned}$$

Ejemplo: Fienberg(1980)

Los siguientes observaciones corresponden a datos sobre accidentes de autos pequeños. Las variables consideradas son: severidad de las heridas del conductor, tipo de accidente y si el conductor fue eyectado o no. La tabla de valores observados y esperados estimados usando el modelo de Asociación Homogénea son:

n_{ijk}		Tipo de Accidente			
		Colisión		Vuelco	
Herido		No Severo	Severo	No Severo	Severo
Conductor Eyectado	No	350	150	60	112
	Sí	26	23	19	80

Cuadro 13: Valores Observados

$\widehat{\mu}_{ijk}$		Tipo de Accidente			
		Colisión		Vuelco	
Herido		No Severo	Severo	No Severo	Severo
Conductor Eyectado	No	350.5	149.5	59.51	112.5
	Sí	25.51	23.49	19.49	79.51

Cuadro 14: Valores Estimados

Los estadísticos de bondad de ajuste resultan

$$X^2 = 0.04323 \qquad G^2 = 0.04334 \quad df = (2-1)(2-1)(2-1) = 1$$

con un p–valor 0.8352928 y 0.8350864, respectivamente. Por lo tanto, el modelo de igualdad de los odds ratios ajusta muy bien.

Sugiero comprobar que los modelos de independencia e independencia condicional no dan un buen ajuste.

Nos falta considerar una cosa importante:

- ¿Cómo elegir un modelo log–lineal adecuado para un conjunto de datos?

¿Cómo elegir un modelo log–lineal adecuado para un conjunto de datos?

Consideremos nuevamente el ejemplo de víctima – defendido – veredicto. Los datos se muestran en el Cuadro 15. Recordemos que con este ejemplo ilustramos la paradoja de Simpson.

Víctima	Defendido	Pena de Muerte	
		Si	No
Blanca	Blanco	53	414
	Negro	11	37
Negra	Blanco	0	16
	Negro	4	139
Total	Blanco	53	430
	Negro	15	176

Cuadro 15: Víctima – Defendido – Veredicto

En la siguiente tabla mostramos los resultados correspondientes a tests de

bondad de ajuste de diferentes modelos loglineales.

Modelo	G^2	df	p–valor
(D,V,P)	137.93	4	0.000
(VP,D)	131.68	3	0.000
(DP,V)	137.71	3	0.000
(DV,P)	8.63	3	0.043
(DP,VP)	131.46	2	0.000
(DP,DV)	7.91	2	0.019
(VP,DV)	1.88	2	0.390
(DP,VP,DV)	0.70	1	0.402
(DVP)	0	0	–

Cuadro 16: Tests de bondad de ajuste

Como sabemos, dado df fijo un G^2 mayor refleja un peor ajuste. La tabla anterior muestra que los modelos (D, V, P), (VP, D), (DP, V) y (DP, VP) dan un ajuste muy pobre. La característica común de todos ellos es que la asociación $D - V$ no está presente. Esto sugiere una asociación importante entre estas dos variables.

De los restanes 4 modelos no saturados (VP, DV) y (DP, VP, DV) parecen dar un ajuste satisfactorio. En el modelo (DP, VP, DV) todas las variables son condicionalmente dependientes, mientras que en el (VP, DV) la raza del defendido y el veredicto son independientes dada la raza de la víctima.

Inferencia sobre asociación condicional

Los tests respecto a las asociaciones condicionales comparan modelos loglineales.

Supongamos que tenemos dos modelos M_0 y M_1, donde M_0 está anidado en M_1, teniendo un término menos que M_1.

El test de cociente de verosimilitud $-2(\ell_0 - \ell_1)$ es, como ya vimos, idéntico a la diferencia de las deviances $G^2(M_0) - G^2(M_1)$. Esto tiene sentido cuando M_1 da un buen ajuste.

Así por ejemplo, supongamos que para el modelo (AB, AC, BC) consideramos la hipótesis de que A–B son condicionalmente independientes. Esto es

$$H_0 : \lambda_{ij}^{AB} = 0$$

para las $(I-1)*(J-1)$ parámetros de asociación AB.

El estadístico del test es

$$G^2((AC, BC)) - G^2((AB, AC, BC))$$

con $(I-1)*(J-1)$ grados de libertad.

En el caso de la situación anterior, por ejemplo, el test para independencia condicional entre veredicto y raza del defendido compara el modelo (VP, DV) con el (DP, VP, DV).

El estadístico correspondiente es

$$G^2(VP, DV) - G^2(DP, VP, DV) = 1.88 - 0.70 = 1.18$$

con $df = 1$ y p–valor$= 0.277356$.

Finalmente, podríamos considerar una secuencia de modelos anidados tales como

$$(D, V, P) \quad (P, DV) \quad (VP, DV) \quad (DP, VP, DV) \quad (DVP)$$

Para asegurar un nivel global en nuestra decisión que no exceda a 0.10 realizamos cada test con un nivel $1-(0.90)^{0.25}$ =0.026. Para un grado de libertad el valor crítico de una χ^2 es 4.96. En la siguiente tabla se muestran los resultados de las comparaciones

De la tabla concluimos que primero aceptamos el modelo de asociación homogénea (DP, VP, DV) y luego aceptamos (VP, DV) dado (DP, VP, DV). En tercera instancia el modelo (P, DV) es rechazado.

Por otro lado, tal como vimos el modelo (VP, DV) da un buen ajuste global (en un test no condicional), por lo que parece razonable elegirlo para representar los datos.

Modelo	G^2	diferencia	df
(D,V,P)	137.93		4
		129.80	1
(DV,P)	8.63		3
		6.25	1
(VP,DV)	1.88		2
		1.18	1
(DP,VP,DV)	0.70		1
		0.70	1
(DVP)	0		0

Cuadro 17: Tests de bondad de ajuste

Por supuesto, la selección en una secuencia de modelos anidados, puede depender de la secuencia elegida. Así, por ejemplo, si la secuencia es (D, V, P), (P, DV), (DP, DV), (DP, VP, DV) el modelo resultante será (DP, VP, DV).