Modelos loglineales

Robert J. Flowers *
Universidad Juárez Autónoma de Tabasco, DACB
Carr. Cunduacán-Jalpa Km 1, Cunduacán Tabasco, México
A.P. 24 C.P. 86690. Tel.(+52)914 336-0928
Eleazar Toledo Cruz
Universidad Juárez Autónoma de Tabasco, DACB
Carr. Cunduacán-Jalpa Km 1, Cunduacán Tabasco, México
A.P. 24 C.P. 86690. Tel.(+52)914 336-0928

Mediante el uso de los modelos loglineales se pueden realizar tablas de contingencias que son estudiadas con dos modelos; el modelo loglineal general y el modelo de logit. Los algoritmos que se utilizan para estimar los parámetros de los modelos loglineales son: 1) el de máxima verosimilitud (MV). 2) el de cuadrados mínimos ponderados (CMP). En este trabajo se define el modelo loglineal, así como el algoritmo para calcular los estimadores usando el método de máxima verosimilitud, proporcionando ejemplos donde se utiliza dicho método en diferentes modelos como de asociación, de logit, de productos cruzados, entre otros y que tiene la forma $C\ln(m) = XB$, además se proporciona un programa hecho en lenguaje `MatLab`, con el fin de poder realizar los cálculos pertinentes.

A class of loglineal models is presented which permits the definition of the general loglinear model and the logit model as well as other loglinear models of interest. Currently, there exist algorithms for obtaining estimates for these models using maximum likelihood estimation or weighted least squares estimates. In this paper a maximum likelihood algorithm is presented which permits one to obtain estimates for models of the form $C\ln(m) = XB$. A program written in `MatLab` is provided.

Palabras clave: Estimadores de máxima verosimilitud, Modelos de asociación, Modelo de logit, Razón de productos cruzados, Modelo loglineal general.
Keywords: Maximum likelihood estimates, Association models, Logit models, Cross product ratios, General loglinear model.

1. Introducción

Dos de los modelos más populares para analizar tablas de contingencia son el modelo loglineal general y el modelo de logit. Estos dos modelos son casos particulares de una clase de modelos loglineales de la forma $C\ln(m) = X\beta$ donde C es una matriz de transformación y m es un vector de valores esperados para las casillas de la tabla de contingencia.

Las dos metodologías principales usados hoy en día para estimar los parámetros de los modelos loglineales son: el de máxima verosimilitud (MV) y el de cuadrados mínimos ponderados (CMP). Estos dos tipos de estimadores tienen la propiedad de ser óptimos asintóticamente normales (OAN). Las razones para considerar más de un tipo de estimadores según Neyman [10] y Ferguson [1] son que (1) se pueden calcular algunos estimadores OAN más fácilmente que otros, y (2) algunos estimadores OAN pueden tener mejores propiedades para muestras pequeñas.

*robert.flowers@basicas.ujat.mx

El problema que tenemos que enfrentar debido a la simplicidad del método CMP es que a veces resulta necesario sumar un valor positivo a cada casilla de la tabla de contingencia que tiene el valor cero para poder usar el procedimiento. El problema con hacer esto es que los coeficientes del modelo de regresión pueden ser sensibles al valor elegido. Otro problema es que, excepto para algunos modelos lineales y loglineales, no se pueden obtener estimadores para los valores esperados mi. Estos estimadores pueden ser útiles para determinar la validez del modelo considerado. En algunos de los modelos de CMP es posible obtener estimadores para las mi que son negativos. Esto no es aceptable, ya que las observaciones son cuentas y deben ser positivas. Esto explica en parte la preferencia que muchos estadísticos tienen por el método de MV. No es necesario modificar las observaciones para poder obtener los estimadores de MV, y se pueden obtener los estimadores de mi para cualquier modelo de interés.

Grizzle, Starmer, y Koch definieron un procedimiento de CMP basado en la distribución multinomial para la clase de modelos loglineales mencionada arriba [6]. Estimadores de MV para modelos loglineales que usan transformaciones de la forma de los modelos de Grizzle, Starmer, y Koch han sido obtenidos por Flowers [3] suponiendo una distribución de Poisson. El algoritmo de Flowers es una extensión del algoritmo de Haberman [7] para obtener estimadores de MV para una gran clase de modelos loglineales. Al final de este artículo se presenta en el apéndice un programa en `MatLab` para calcular los estimadores de este modelo.

2. Metodología

En esta sección, definiremos una clase de modelos de la forma $C\ln(m) = X\beta$. Esta clase de modelos incluye el modelo loglineal general, el modelo de logit y varios otros de interés.

Para la mayoría de los modelos de interés se obtienen los mismos estimadores bajo los supuestos de una distribución multinomial, una distribución producto-multinomial, o una distribución de Poisson. Aquí, supondremos que la variable Y_{ij} sigue una distribución de Poisson con media m_{ij}.

Definimos

$$\begin{aligned} y_i' &= (y_{i1}, y_{i2}, \ldots, y_{iJ}), \\ m_i' &= (m_{i1}, m_{i2}, \ldots, m_{iJ}), \\ y' &= (y_1, y_2, \ldots, y_I), \\ m' &= (m_1, m_2, \ldots, m_I) \end{aligned}$$

Obtenemos los estimadores para el modelo loglineal al maximizar el logaritmo de la función de verosimilitud sujeto a las restricciones de que

$$C\ln(m) = X\beta.$$

Se define la matriz C como

$$C = \begin{bmatrix} \varsigma_1 & 0 & \ldots & 0 \\ 0 & \varsigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \varsigma_I \end{bmatrix}$$

donde ς_i es una matriz de transformación $H \times J$.

Ahora, se puede obtener los estimadores de máxima verosimilitud por diferenciar el Lagrangiano

$$y' \ln(m) - \iota' m - \tau'[C \ln(m) - X\beta]$$

con respeto a m, donde ι es un vector $IJ \times 1$ de unidades y τ es un vector de multiplicadores de Lagrange. Igualando la derivada a cero se obtiene

$$D_m^{-1}(y - m) - D_m^{-1} C' \tau = 0$$

donde D_m es una matriz diagonal compuesto de los elementos del vector m. Si multiplicamos la expresión para la derivada del Lagrangiano por D_m obtenemos

$$y - m = C'\tau.$$

Se puede hacer un desarrollo en serie de Taylor para el logaritmo de y para obtener

$$\ln(y) = \ln(m) + D_m^{-1}(y - m) + \gamma.$$

donde γ representa el término de error en la serie de Taylor. Se puede combinar estas últimas dos ecuaciones para obtener

$$\ln(y) - \gamma - \ln(m) = D_m^{-1} C' \tau.$$

Ahora, si se multiplica esta ecuación por la matriz C se obtiene

$$C \ln(y) - C\gamma - C \ln(m) = C D_m^{-1} C' \tau.$$

Se puede definir $u = \ln(y) - \gamma$ y aplicar la restricción $C \ln(m) = X\beta$ para obtener

$$Cu - X\beta = C D_m^{-1} C' \tau.$$

La matriz definida por

$$V = C D_m^{-1} C'$$

representa la varianza asintótica de u. Sustituyendo en la ecuación anterior se obtiene

$$Cu - X\beta = V\tau.$$

Multiplicando por la matriz V^{-1} da

$$V^{-1}(Cu - X\beta) = \tau.$$

Ahora, si derivamos el lagrangiano con respecto a β obtenemos $X'\tau = 0$. Entonces se puede multiplicar la ecuación anterior por X' para obtener

$$X'V^{-1}(Cu - X\beta) = 0.$$

Despejando β da

$$\beta = (X'V^{-1}X)^{-1}X'V^{-1}Cu.$$

Entonces se puede dar el algoritmo como:

$$\begin{aligned}
\beta^{(s+1)} &= [X'[V^{(s)}]^{-1}X]^{-1}X'[V^{(s)}]^{-1}Cu^{(s)} \\
m^{s+1} &= y - C'[V^{(s)}]^{-1}[Cu^{(s)} - X\beta^{(s+1)}] \\
u^{(s)} &= \ln[m^{(s)}] + [D_m^{(s)}]^{-1}(y - m^{(s)})
\end{aligned}$$

Para valores iniciales se puede usar $m^{(0)} = y + \frac{1}{2}\iota$ y $u^{(0)} = \ln m^{(0)}$.

3. El modelo de Bradley-Terry

Se puede representar la probabilidad de que el equipo T_i gane contra el equipo T_j por la relación siguiente:

$$Pr[T_i > T_j] = \frac{\pi_i}{\pi_i + \pi_j}, i \neq j.$$

Este modelo se llama el modelo de Bradley-Terry. Bajo este modelo, el número total de juegos entre equipos y el número total de victorias por equipo tienen valores fijos. Hay varias maneras de estimar los parámetros del modelo de Bradley-Terry. Se puede usar un modelo de la cuasi-independencia, un modelo de la cuasi-simetría, o un modelo de logit.

En esta sección, se definirá un modelo de logit en la misma forma que hizo Koch, Freeman, y Tolley [9] con la excepción que aquí se usará estimadores de MV.

Si se define

$$p_{ij} = Pr[T_i > T_j] = \frac{\pi_i}{\pi_i + \pi_j},$$

entonces

$$\ln(p_{ij}/p_{ji}) = \ln(\pi_i) - \ln(\pi_j).$$

Ahora se puede definir un modelo de logit con una definición apropiada de la matriz del modelo.

Ejemplo: Los datos de la tabla 1 son de la División Oeste de la Liga Americana para el año 2002. Los equipos representados son Anaheim (Ana.), Oakland (Oak.), Texas (Tex.) y Seattle (Sea.). Los valores presentados en la tabla son el número de victorias por equipo en cada serie, entre equipos de la división del oeste. Por ejemplo, en la primera pareja Anaheim ganó 8 juegos contra Oakland y perdió 11 veces.

El vector y se define tomando los valores de la tabla 1 fila por fila. Para definir los subíndices para los parámetros se considera a Anaheim como el equipo 1, Oakland el equipo 2, Texas el equipo 3, y Seattle el equipo 4. Se puede definir los coeficientes β

Pareja	Victorias	
	Primer equipo	Segundo equipo
Ana-Oak	8	11
Ana-Tex	12	7
Ana-Sea	9	10
Oak-Tex	13	6
Oak-Sea	8	11
Tex-Sea	7	13

Tabla 1. Número de victorias entre equipos de la División Oeste de la Liga Americana

en términos de los parámetros π en la siguiente manera:

$$\begin{aligned}
\ln(p_{12}/p_{21}) &= \ln(\pi_1) - \ln(\pi_2) = \beta_1 \\
\ln(p_{13}/p_{31}) &= \ln(\pi_1) - \ln(\pi_3) = \beta_2 \\
\ln(p_{14}/p_{41}) &= \ln(\pi_1) - \ln(\pi_4) = \beta_3 \\
\ln(p_{23}/p_{32}) &= \ln(\pi_2) - \ln(\pi_3) = \beta_2 - \beta_1 \\
\ln(p_{24}/p_{42}) &= \ln(\pi_2) - \ln(\pi_4) = \beta_3 - \beta_1 \\
\ln(p_{34}/p_{43}) &= \ln(\pi_3) - \ln(\pi_4) = \beta_3 - \beta_2
\end{aligned}$$

La matriz de modelo se define por

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

Ahora se puede definir el modelo de logit como $C\ln(m) = X\beta$ donde $C = I_6 \otimes [1 \ -1]$ y I_d es la matriz identidad $d \times d$.

Se muestran los estimadores de máxima verosimilitud para este modelo en la tabla 2. El modelo tiene $G^2 = 0.593$ con 3 grados de libertad. Esto representa un muy buen ajuste a los datos y por lo tanto se concluye que el modelo de Bradley-Terry es aceptable.

Pareja	Victorias	
	Primer equipo	Segundo equipo
Ana-Oak	8.7337	10.2663
Ana-Tex	11.8705	7.1295
Ana-Sea	8.3957	10.6043
Oak-Tex	12.5749	6.4251
Oak-Sea	9.1588	9.8412
Tex-Sea	6.4454	13.5546

Tabla 2. Valores esperados para el modelo de Bradley y Terry División Oeste de la Liga Americana

En la tabla 3, se muestran las estimaciones para los coeficientes de regresión.

Variable	Coeficiente	Error Estándar	Estadística Z
1	-0.1617	0.3288	-0.4918
2	0.5098	0.3342	1.5254
3	-0.2335	0.3287	-0.7105

Tabla 3. Coeficientes de regresión

Para una explicación más completa del modelo de Bradley-Terry, se recomienda que el lector vea Fienberg y Larntz [2].

4. Un modelo para medir el efecto de luces de alto

Aquí, consideramos un modelo definido por Flowers y Cruz [4] para medir el efecto de luces de alto que tienen algunos modelos de autos en el parabrisas trasero al centro, las cuales encienden cuando el auto se frena. Los datos de la tabla 4 son de un estudio hecho por Kahane [8]. Kahane noto que el modelo usado debe incluir un efecto por la edad del auto. Si definimos un modelo con la variable dependiente definido en términos de la razón de productos cruzados de los años adyacentes, entonces si hay un efecto de edad las razones de productos cruzados no deben ser iguales a la unidad. A partir del 1 de Septiembre de 1985 se exigieron luces de alto en todos los nuevos automóviles en los Estados Unidos. Debido a esta ley, si la ley es efectiva la razón de producto de cruzados que compara los accidentes del año 1985 con los de 1986 debe bajar relativo a las razones de producto cruzados anteriores.

Año del modelo	Sin daño por impacto trasero	Con daño por impacto trasero
1980	16467	5408
1981	15344	5435
1982	14171	5265
1983	16100	6212
1984	24644	10216
1985	27131	11828
1986	29946	12695
1987	19983	8620

Tabla 4. Accidentes en el estado de Michigan con y sin daño por impacto trasero

En la tabla 5, se muestran las razones de producto cruzados observadas para los datos del estado de Michigan. Se puede ver que para todas las comparaciones que las razones de productos cruzados están ligeramente arriba de 1 con la excepción de la comparación de los años 1985 y 1986. Se espera los valores arriba de 1 por el efecto de la edad del auto y el valor bajo de 1 por el efecto de la ley.

Si se define un modelo con la razón de producto cruzados como variable indepen-

Comparación	Razón de productos cruzados
80-81	1.07854622
81-82	1.04890684
82-83	1.03850285
83-84	1.07439533
84-85	1.05166115
85-86	0.97240766
86-87	1.01754282

Tabla 5. Razones de productos cruzados para años adyacentes.

diente, entonces se puede definir la matriz del modelo como:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$$

Para definir un modelo de la forma $C\ln(m) = X\beta$ donde la variable dependiente es un vector de las razones de producto cruzado definimos C como

$$C = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{bmatrix}$$

donde el vector y define tomando los elementos de la tabla 4 fila por fila. Los parámetros β_1 y β_2 del vector β representan los efectos de la antigüedad y de las luces de alto respectivamente. Bajo este modelo se supone que el efecto de la antigüedad es constante. La prueba de bondad de ajuste indica que este modelo hace un buen ajuste a los datos ya que $G^2 = 6.746$ con 5 grados de libertad. Para estos datos, $\hat{\beta}_2 = -0.0946$. Haciendo la prueba z se obtiene $z = -6.454$ la cual es altamente significativa. Entonces se puede concluir que las luces de alto si bajaron el número de accidentes traseras.

Se muestran los estimadores de máxima verosimilitud para este modelo en la tabla 6.

5. Modelos de asociación

Hay muchas tablas de contingencia en dos dimensiones donde existe una dependencia entre dos factores A y B. En este caso, es probable que el modelo de la independencia no hace un buen ajuste a los datos. En la tabla 7, se tiene datos para jugadores de béisbol con por lo menos 30 veces al bate en los años 2000 y 2001. Esta tabla fue creada usando datos de la revista Major League Baseball Yearbook 2002. Los jugadores están clasificados según sus promedios de bateo en los años 2000 y 2001. Aquí existe una dependencia entre los resultados de los dos años. Hay una alta probabilidad que el jugador mantenga el mismo nivel

Año del modelo	Sin daño por impacto trasero	Con daño por impacto trasero
1980	16417	5458
1981	15382	5397
1982	14184	5252
1983	16044	6268
1984	24683	10177
1985	27147	11812
1986	30077	12564
1987	19852	8751

Tabla 6. Valores esperados del modelo para medir el efecto de luces de alto. Accidentes en el estado de Michigan con y sin daño por impacto trasero

en los dos años. Para tablas de este tipo es natural considerar modelos de simetría y homogeneidad marginal pero aquí solamente se considera los modelos de asociación desarrollados por Goodman (1979).

2000	2001				Total
	$\leq .260$	$(.260, .280]$	$(.280, .300]$	$> .300$	
$\leq .260$	35	15	9	7	66
$(.260, .280]$	13	22	11	4	50
$(.280, .300]$	16	19	12	14	61
$> .300$	13	16	7	31	67
Total	77	72	39	56	244

Tabla 7. Clasificación de los bateadores según sus promedios de bateo

En esta sección se usan modelos de la forma $C\ln(m) = X\beta$ donde C es la matriz identidad. Los primeros dos modelos de interés son el modelo de la independencia y el modelo de la asociación uniforme. Para el modelo de la asociación uniforme, la matriz del modelo se define como sigue:

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 2 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 4 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 4 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 6 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 8 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 3 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 6 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 9 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 12 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 4 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 8 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 12 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

Se puede obtener la matriz del modelo para el modelo de la independencia excluyendo la última columna de esta matriz. En la tabla 8 se dan los valores esperados para el modelo de la independencia. Este modelo tiene $G^2_{Ind} = 45.1944$ con 9 grados de libertad y no hace

un buen ajuste en el nivel $\alpha = 0.05$. En la tabla 9 se muestran los valores esperados para el modelo de la asociación uniforme. Este modelo tiene $G^6_{as} = 16.0669$ con 8 grados de libertad. Este modelo tiene un ajuste mucho mejor que el anterior, pero tampoco hace un buen ajuste aunque falla por poco.

2000	2001				Total
	$\leq .260$	$(.260, .280]$	$(.280, .300]$	$> .300$	
$\leq .260$	20.8279	19.4754	10.5492	15.1475	66
$(.260, .280]$	15.7787	14.7541	7.9918	11.4754	50
$(.280, .300]$	19.2500	18.0000	9.7500	14.0000	61
$> .300$	21.1434	19.7705	10.7090	15.3770	67
Total	77	72	39	56	244

Tabla 8. Valores esperados para el modelo de la independencia

2000	2001				Total
	$\leq .260$	$(.260, .280]$	$(.280, .300]$	$> .300$	
$\leq .260$	31.7856	20.9503	7.2378	6.0263	66
$(.260, .280]$	18.4545	16.0912	7.3541	8.1002	50
$(.280, .300]$	15.7744	18.1954	11.0008	16.0294	61
$> .300$	10.9856	16.7631	13.4073	25.8440	67
Total	77	72	39	56	244

Tabla 9. Valores esperados para el modelo de la asociación uniforme

La matriz del modelo para el modelo de la asociación por efecto de fila y columna se define como sigue:

$$X = \begin{bmatrix}
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 3 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 2 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 2 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 3 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 3 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 3 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 3 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 4 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

En la tabla 10 se dan los valores esperados para el modelo de la asociación por efecto de fila y columna. Este modelo tiene $G^2_{EFC} = 10.9168$ con 4 grados de libertad y no hace un buen ajuste.

Si se excluyen las últimas dos columnas de la matriz, se obtiene la matriz del modelo para el modelo de asociación por efecto de fila. Se muestran los valores esperados para este modelo en la tabla 11. Este modelo tiene $G^2_{EF} = 15.8687$ con 6 grados de libertad y no hace un buen ajuste a los datos.

2000	2001				Total
	$\leq .260$	$(.260, .280]$	$(.280, .300]$	$> .300$	
$\leq .260$	31.5727	19.1347	11.0125	4.2801	66
$(.260, .280]$	17.8391	15.7231	9.0366	7.4012	50
$(.280, .300]$	16.6039	19.1496	9.8891	15.3574	61
$> .300$	10.9844	17.9926	9.0617	28.9613	67
Total	77	72	39	56	244

Tabla 10. Valores esperados para el modelo de la asociación por efecto de fila y columna

2000	2001				Total
	$\leq .260$	$(.260, .280]$	$(.280, .300]$	$> .300$	
$\leq .260$	31.5774	20.9766	7.3146	6.1314	66
$(.260, .280]$	18.1393	16.0594	7.4634	8.3379	50
$(.280, .300]$	16.7416	18.5139	10.7472	14.9972	61
$> .300$	10.5417	16.4501	13.4748	26.5334	67
Total	77	72	39	56	244

Tabla 11. Valores esperados para el modelo de la asociación por efecto de fila

Para obtener el modelo de la asociación por efecto de columna, hay que definir la matriz de modelo como sigue:

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 3 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 3 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 4 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 4 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 4 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

En la tabla 12 se muestran los valores esperados para el modelo de asociación por efecto de columna. Este modelo tiene $G^2_{EC} = 11.0334$ con 6 grados de libertad y si hace un buen ajuste a los datos.

6. Sumario

Se presento un modelo, el cual puede usarse para definir un gran número de modelos para el análisis de tablas de contingencias. Usando esta metodología se pudo definir modelos de logit, modelos loglineales generales, modelos de asociación y otros. De donde la matriz de

2000	2001				Total
	$\leq .260$	$(.260, .280]$	$(.280, .300]$	$> .300$	
$\leq .260$	31.2559	19.1567	11.1652	4.4223	66
$(.260, .280]$	18.5577	15.7701	8.7440	6.9282	50
$(.280, .300]$	16.1170	18.9897	10.0166	15.8767	61
$> .300$	11.0694	18.0835	9.0743	28.7728	67
Total	77	72	39	56	244

Tabla 12. Valores esperados para el modelo de asociación por efecto de columna

transformación tenga que ser bien definida.

Para hacer los análisis se incluyo un programa en `MatLab`, proporcionando una herramienta básica que puede ser accesible.

Se espera que con el uso de este programa y el de Flowers, López Segovia y Alavez Ramirez [5], se pueda cubrir una gran parte de los modelos que se enseñan tradicionalmente en un curso de Análisis Multivariado Discreto.

Apéndice

```
%Este program calcula los valores de beta(B),estimadores(M),ji-cuadrada
%de la razon de verosimilitud(g),matriz varianza covarianza(VB),prueba
% Z(Z), y grados de libertad(gl) de un modelo de la forma Cln(m)=XB con
% relacion a la materia de analisis multivariado discreto. Se da la
% matriz X, Y, y C desde la ventana de comando de MatLab y ahi se llama
% o se corre el programa.
tol=0.000001;
Xt=X';
z=length(Y);
L=ones(z,1);
M0=Y+(0.5*L);  %Vector inicial que se da cuando alguna observacion es
Dm=diag(M0);   %igual a 0
Vo=(C*inv(Dm))*C';
u=log(M0);
n=size(X,1);
norma=tol+1;
while norma>tol;
    A=Xt*inv(Vo);
    Cz=A*X;
    E=inv(Cz);
    F=inv(Vo)*(C*u);
    G=E*Xt;
    B=G*F;     %vector beta
    M=Y-((C'*inv(Vo))*((C*u)-(X*B)));  %vector de estimadores
    norma=norm(M-M0);
    M0=M;
    Dm=diag(M0);
    u=log(M)+((inv(Dm))*(Y-M0));
    V=(C*inv(Dm))*C';
    Vo=V;
end;
```

```
for i=1:z
    if Y(i)~=0
        g=g-2*Y(i)*log(M(i)/Y(i));
    else g=g;
    end;
end;
VB=inv(Xt*inv(V)*X);
RVB=sqrt(diag(VB));
Z=B./RVB;
m=size(X,2)
gl=n-m;
fprintf(1,'La matriz B es: ',B);  %Indica la matriz beta%
B
fprintf(1,'La matriz M de los estimadores es: ',M);  %la matriz de los
M                                                    %estimadores%
fprintf(1,'El valor de G-cuad es: ',g);   %Da el valor de ji-cuadrada de
g                                         %verosimilitud%
fprintf(1,'La matriz de varianza-covarianza es: ',VB); %la matriz de
VB                                                   %covarianza%
fprintf(1,'El vector z es: ',Z); %los valores de la prueba z%
Z
fprintf(1,'Los grados de libertad son: ',gl);  %da el valor de grados de
gl                                             %libertad%
end
```

Referencias

[1] Ferguson, T.S., 1958. A meted of generating best asymptotically normal estimates with application to the estimation of bacterial densities. Annals of Mathematical Statistics, 29: 1046-1062.

[2] Fienberg y Larntz (1976). Loglinear representation for paired and múltiple comparison models. Biométrica; 63: 245-254.

[3] Flowers, R.J., 1984. Discrete multivariate analysis using loglinear models. Universidad y Ciencia, 1(1), 45-56.

[4] Flowers, R.J. y H. D. Cruz-Suarez, S., 1994. Un procedimiento de máxima verosimilitud para medir el efecto de las luces de alto. Revista Unidad Chontalpa, No. 4: 1-10.

[5] Flowers, R.J., L.López-Segovia, J.Alavez-Ramirez, 2004. Pruebas estadísticas para homogeneidad marginal o simetría. Revista de Ciencias Básicas, 3 (1), 29-44.

[6] Grizzle, J.E., C.F. Starmer, y G.G. Koch, 1969. Análisis of categorical data by linear models. Biometrics, 25:489-504.

[7] Haberman, S.J., 1974. Log-linear models for frequency tables with ordered classifications. Biometrics, 30:589-600.

[8] Kahane, C.J., 1989. An evaluation of center high mounted stop lamps on 1987 data. Report No. DOT HS 807 442 Washington, D.C.: National Highway Traffic Safety Administration.

[9] Koch, G.G, D.H. Freeman, y H.D. Tolley, (1975). The asymptotic covariance structure of estimated parameters from contingency table log-linear models (Institute of Statistics Mimeo Series No. 1046). Chapel Hill: University of North Carolina.

[10] Neyman, J. 1949. Contributions to the theory of the X2 test. Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability, (edited by J. Neyman). Berkeley: University of California Press, 230-273.