Revista Investigación Educativa - Vol. 14, n.º 1, págs. 97-114

ANÁLISIS SECUENCIAL DE DATOS OBSERVACIONALES EN INVESTIGACIÓN EDUCATIVA (Y II): PERSPECTIVA MULTIVARIANTE CON MODELOS LOG-LINEALES Y *LOGIT*

por
Juan Carlos Tójar y José Serrano
Métodos de Investigación y Diagnóstico en Educación
Universidad de Málaga

RESUMEN

Los modelos log-lineales y logit proporcionan una alternativa multivariante a las técnicas clásicas de análisis de datos categóricos secuenciales procedentes de la observación sistemática.

En este trabajo se muestra la utilidad y la adecuación de estos modelos matemáticos para representar fenómenos de interacción observados en sus contextos de origen. Se muestra el procedimiento completo incluyendo la organización de los datos registrados en tablas de contingencia multidimensionales, la construcción, la evaluación y la interpretación de los modelos así como la estimación de sus correspondientes parámetros. Se apuntan además otras posibilidades diferentes de análisis secuencial de datos observacionales mediante modelos log-lineales a desarrollar en futuras investigaciones (modelos de cuasi-independencia, simetría y cuasi-simetría, modelos con datos ordinales y relaciones con los modelos causales). Por último se concreta una aplicación de los modelos logit en la evaluación de la calidad de los datos registrados (concordancia secuencial).

Descriptores: *Análisis secuencial de datos; Modelos logarítmico lineales; Modelos* logit; *Observación; Metodología de investigación.*

ABSTRACT

Log-linear and logit models provide a multivariant choice to the traditional techniques of analysis of categorial and sequential data derived from systematic observation.

Throughout this report it is shown the usefulness and fitting of these mathematical models to represent phenomenons of interaction observed in their original contexts. The whole process is shown including the organization of data recorded in tables of multidimensional contingencies, the construction, assessment and interpretation of the models, as well as the estimation of its corresponding parameters. Other different possibilities from the sequential analysis of observational data are pointed out by means of log-linear models to be developed in later researches (e. g. models of quasi-independence, symmetry and quasi-symmetry, models with ordinal data and relations with causal models). Finally, it is established an application of logit models to assess at the recorded data quality (sequential concordance).

Key words: *Sequential data analysis; Log-linear models; Logit models; Observation; Research methodology.*

1. INTRODUCCIÓN

La necesidad del empleo de diseños multivariantes de investigación se traduce en la incorporación progresiva de técnicas y procedimientos más sofisticados de análisis de datos para responder a problemas en contextos sociales y educativos. Los modelos log-lineales y *logit* tienen en este empeño un importante papel, ya que ofrecen una alternativa multivariante a las técnicas clásicas de análisis secuencial de datos observacionales (a las que fue dedicada la primera parte de esta serie de dos artículos), capaces sólo de enfocar los problemas desde una perspectiva bivariante.

La forma más común de organizar los datos observacionales, desde un punto de vista secuencial, es en una tabla de contingencia (Bakeman, 1991). Cuando la tabla tiene dos dimensiones *axb* (si contemplan por ejemplo 6 categorías de un sistema en la secuencia de t a $t + 1$ se obtendría una tabla 6x2), una forma razonable de analizarla es mediante el estadístico χ^2. El problema surge cuando se incluye una tercera dimensión (por ejemplo la variable «observador» con las categorías «observador 1», «observador 2»,... «observador m»), o n dimensiones. En esta tesitura, con una tabla de contingencia multidimensional $a \times b \times c \times ... \times n$ la mejor vía de análisis pasa por los modelos log-lineales (Goodman, 1972a, 1972b; Haberman, 1978, 1979; Agresti, 1984; Hagenaars, 1990).

Como señalaron Bakeman, Adamson y Strisik (1989) el hecho de que hayan proliferado recientemente paquetes estadísticos (v. g. SPSS) y trabajos específicos para investigadores sociales (v. g. Kennedy, 1983) y para no estadísticos (v. g. Fienberg, 1980) sobre los modelos log-lineales, facilita el hecho de que dichos modelos se vayan insertando en el uso común de investigadores del comportamiento.

Como muestra de la importancia que el análisis log-lineal está alcanzando actualmente en la metodología observacional se puede considerar el hecho de que uno de

los más recientes manuales de análisis secuencial, como es el de Gottman y Roy (1990), dedique más de tres capítulos a este tipo de modelos.

En las páginas siguientes se expone el proceso que permite aplicar los modelos log-lineales para realizar análisis secuenciales. La aplicación en una investigación en el contexto del aula de lo expuesto en los próximos apartados puede consultarse en los trabajos de Tójar (1993; en prensa).

2. DE LAS TABLAS DE CONTINGENCIA A LA CONSTRUCCIÓN DE LOS MODELOS

Para empezar supóngase una tabla de contingencia de doble entrada de a x b niveles (v tabla 1). Por ejemplo, la variable A podría hacer referencia a las frecuencias de un registro de un sistema de codificación en el tiempo t y B, a la misma variable, pero en el tiempo $t + 1$.

TABLA 1
TABLA DE CONTINGENCIA DE A X B

		B			
A	B_1	B_2	...	B_b	
A_1	F_{11}	F_{12}	...	F_{1b}	$F_{1.}$
A_2	F_{21}	F_{22}	...	F_{2b}	$F_{2.}$
...	...	...	...	...	...
A_a	F_{a1}	F_{a2}	...	F_{ab}	$F_{a.}$
	$F_{.1}$	$F_{.2}$	...	$F_{.b}$	$F_{..}$

A partir de la tabla 1, es posible construir un modelo que sea capaz de generar frecuencias esperadas en las celdas en función de parámetros que representen las características relevantes de las variables categóricas y las relaciones de interacción de unas con otras.

Existen fundamentalmente dos aproximaciones log(arítmico)-lineales para modelizar las frecuencias en una tabla de contingencia. Por un lado, los *modelos log-lineales*, en los que no existe diferencia entre variables dependientes e independientes (*estudios simétricos* en los que todas serían variables respuesta) y, por otro, los *modelos logit*, en los que una variable es considerada como dependiente (*estudios asimétricos*).

Tanto si son log-lineales o logit, los modelos se clasifican en *saturados* e *insaturados*. Los modelos saturados estiman las frecuencias observadas a partir de todas las variables y todas las relaciones posibles entre ellas. Por ejemplo, para estimar las frecuencias esperadas en una tabla de doble entrada como la anterior se escribiría:

$$\widehat{F}_{ab} = \eta \, \tau_a^A \, \tau_b^B \, \tau_{ab}^{AB} \qquad [1]$$

donde η es la media geométrica de las F_{ij}.
τ_a^A representa los efectos (uno por cada nivel de *A*) que se presentan si la distribución de la variable *A* en las categorías de *B* son en promedio desiguales.
τ_b^B representa los efectos (uno por cada nivel de *B*) que se presentan si la distribución de la variable *B* en las categorías de *A* son en promedio desiguales.
τ_{ab}^{AB} representa los efectos (uno por cada casilla) y están presentes si las variables no son independientes, esto es, están relacionadas.

En general, si $\tau = 1$, no habría efecto. Si $\tau > 1$ ó $\tau < 1$, sí habría efecto, y significaría que $\hat{F} < F$ ó $F > \hat{F}$, respectivamente.

En una tabla de 2 x 2 serían posibles los cuatro efectos siguientes:

TABLA 2
FRECUENCIAS ESPERADAS GENERADAS POR EL MODELO DE LA EXPRESIÓN [1]

		B
A	B_1	B_2
A_1	$\hat{F}_{11} = \eta\, \tau_1^A\, \tau_1^B\, \tau_{11}^{AB}$	$\hat{F}_{12} = \eta\, \tau_1^A\, \tau_2^B\, \tau_{12}^{AB}$
A_2	$\hat{F}_{21} = \eta\, \tau_2^A\, \tau_1^B\, \tau_{21}^{AB}$	$\hat{F}_{22} = \eta\, \tau_2^A\, \tau_2^B\, \tau_{22}^{AB}$

Goodman (1972b) introdujo una notación diferente, utilizada con mayor asiduidad en la actualidad por su similitud en la forma con la usada en la regresión. Tomando logaritmos en la expresión [1] se tiene:

$$Ln\left(\hat{F}_{ab}\right) = Ln\left(\eta\, \tau_a^A\, \tau_b^B\, \tau_{ab}^{AB}\right) = Ln\left(\eta\right) + Ln\left(\tau_a^A\right) + Ln\left(\tau_b^B\right) + Ln\left(\tau_{ab}^{AB}\right) \qquad [2]$$

Y denominando λs a los logaritmos de las τs, θ al logaritmo de η y G_{ab} al logaritmo de $\hat{F}_{ab}$ se tiene:

$$G_{ab} = \theta + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB} \qquad [3]$$

Para esta nueva notación, si $\lambda = 0$, no habría efecto. Si $\lambda > 0$ ó $\lambda < 0$, sí habría efecto, y significaría que $\hat{F} > F$ ó $\hat{F} > F$, respectivamente.

La significación estadística de los parámetros se determina fácilmente mediante el error estándar de las λs (Goodman, 1972b).

$$\hat{S}_\lambda = \sqrt{\frac{\sum_a \sum_b \left(\frac{1}{F_{ab}}\right)}{C^2}} \qquad [4]$$

donde C^2 es el número total de celdas.

Con esta nueva notación, si añadiésemos una nueva variable (v. g. el registro de las frecuencias de un sistema de categorías en el tiempo $t + k$, o el registro realizado en el

tiempo t por otro codificador), el modelo log-lineal saturado con tres variables (A, B y C) con a x b x c niveles sería el siguiente:

$$G_{abc} = \theta + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{cb}^{CB} + \lambda_{abc}^{ABC} \quad [5]$$

Y para n variables con j x k x l x ... x v niveles, el modelo saturado se construiría así:

$$G_{abcd....n} = \theta + \lambda_a^A + \lambda_b^B + \lambda_c^C + ... + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{ad}^{AD} + \\ + ... + \lambda_{abc}^{ABC} + ... + \lambda_{abc...n}^{ABCD...N} \quad [6]$$

Los modelos no saturados o *insaturados*, son preferibles por ser más parsimoniosos. En general, como se comentará más adelante a propósito de la selección de los modelos, se trata de conseguir un modelo que con el mínimo número de efectos posibles proporcione una buena estimación de las frecuencias (F_{ij}) de las celdas.

Un ejemplo de modelo insaturado es el *modelo de independencia* que consiste en suprimir todos los efectos de las interacciones, que en el caso de tres variables tendría la forma:

$$G_{abc} = \theta + \lambda_a^A + \lambda_b^B + \lambda_c^C \quad [7]$$

En el caso general, un modelo log-lineal de independencia de n variables con a x b x c x ... x n niveles se construiría así:

$$G_{abcd....n} = \theta + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_d^D + ... + \lambda_n^N \quad [8]$$

El principal tipo de modelos log-lineales insaturados lo constituyen los *modelos jerárquicos*. Este tipo de modelos se construyen a partir del modelo saturado eliminando ciertos términos siguiendo la condición siguiente: si se incluye un término que representa la interacción de un conjunto de variables (*clase generadora*), obligatoriamente han de ser incluidos todos los términos de orden inferior que representen las combinaciones posibles de ese conjunto de variables.

Por ejemplo, si en un modelo con 4 variables se quiere incluir el término que representa el efecto λ_{abc}^{ABC} (la clase generadora sería la ABC), obligatoriamente se han de incluir los efectos: λ_a^A, λ_b^B, λ_c^C, λ_{ab}^{AB}, λ_{ac}^{AC}, λ_{cb}^{CB}, además de λ_d^D como efecto principal y θ como el logaritmo de la media geométrica (η).

Cuando Goodman presentó su trabajo sobre los modelos lineales, no incluyó en ellos la posibilidad de utilizar *modelos no jerárquicos*. Aunque, matemáticamente son posibles, no siempre tiene sentido sustantivo el plantearlos (v. Haberman, 1978, 1979; Rindskopf, 1990).

Para construir modelos logit, por su parte, se puede tener en cuenta la misma clasificación hecha en los log-lineales (saturados o insaturados, jerárquicos o no jerárquicos). La diferencia está en la concepción de partida, el hecho de que una variable

sea tomada como dependiente, y esto condiciona la forma del modelo. Los modelos logit son para las variables categoriales lo que los modelos de regresión lineal para las variables dependientes continuas (Knoke y Burke, 1980: 24).

Estos modelos toman el nombre de los logaritmos neperianos de las razones entre frecuencias (*odds*), denominados comúnmente *logit*. Un *logit* se define usualmente como 1/2 de los logaritmos neperianos de estas razones, no obstante Goodman (1972a: 35) adoptó la convención de analizar directamente los logaritmos de las razones.

Si se parte de un modelo log-lineal de tres variables (*A*, *B* y *C*):

$$\widehat{F}_{abc} = \eta\, \tau_a^A\, \tau_b^B\, \tau_c^C\, \tau_{ab}^{AB}\, \tau_{bc}^{BC}\, \tau_{ac}^{AC}\, \tau_{abc}^{ABC} \qquad [9]$$

la *odds* esperada de la variable *A* (supuestamente dicotómica) valdrá:

$$\frac{\widehat{F}_{1bc}}{\widehat{F}_{2bc}} = \frac{\eta\, \tau_1^A\, \tau_b^B\, \tau_c^C\, \tau_{1b}^{AB}\, \tau_{bc}^{BC}\, \tau_{1b}^{AC}\, \tau_{1bc}^{ABC}}{\eta\, \tau_2^A\, \tau_b^B\, \tau_c^C\, \tau_{2b}^{AB}\, \tau_{bc}^{BC}\, \tau_{2c}^{AC}\, \tau_{2bc}^{ABC}} =$$

$$= \frac{\tau_1^A\, \tau_{1b}^{AB}\, \tau_{1c}^{AC}\, \tau_{1bc}^{ABC}}{\tau_2^A\, \tau_{2b}^{AB}\, \tau_{2c}^{AC}\, \tau_{2bc}^{ABC}} = (\tau^A)^2\, (\tau^{AB})^2\, (\tau^{AC})^2\, (\tau^{ABC})^2$$

y expresándolo en forma logarítmica:

$$Ln \frac{\widehat{F}_{1bc}}{\widehat{F}_{2bc}} = 2Ln\, \tau^A + 2Ln\, \tau_b^{AB} + 2Ln\, \tau_c^{AC} + 2Ln\, \tau_{bc}^{ABC} \qquad [10]$$

Siguiendo la notación de Goodman (1972b) expuesta aquí por primera vez en la expresión [3] se llegaría a:

$$Ln \frac{\widehat{F}_{1bc}}{\widehat{F}_{2bc}} = 2\lambda^A + 2\lambda_b^{AB} + 2\lambda_c^{AC} + 2\lambda_{bc}^{ABC} \qquad [11]$$

El propio Goodman (1972b) propone denominar Φ_{jk}^A al logaritmo de las *odds* condicionales de *A*, y β a 2λ, ó bien $\beta = 2$ Ln τ.[Aplicando esta notación a [11] (o a [10]) se obtiene:

$$\Phi_{bc}^A = \beta^A + \beta_b^{AB} + \beta_c^{AC} + \beta_{bc}^{ABC} \qquad [12]$$

De acuerdo con la nueva notación, las suma de las *betas* de cada nivel de una variable independiente relacionada con la dependiente (*A*) ha de ser cero. Así, si *C* tiene por ejemplo tres niveles:

$$\beta_1^{AC} + \beta_2^{AC} + \beta_3^{AC} = 0 \qquad [13]$$

Con el uso de los modelos logit, en lugar de los log-lineales, no sólo se ahorra espacio a la hora de formular el modelo en cuestión sino que además el número de parámetros a estimar es menor y el modelo se hace aún más comprensible al adoptar la *apariencia* de la regresión lineal. Esto se puede comprobar comparando las expresiones [9] y [12], o la expresión [6] que representa el modelo log-lineal saturado para n variables con la siguiente en la que se formula un modelo logit saturado también para n variables, tomando como dependiente a la A e independientes al resto:

$$\Phi^{A}_{bcd\ldots n} = \beta^{A} + \beta^{AB}_{b} + \beta^{AC}_{c} + \beta^{ABC}_{bc} + \ldots + \beta^{ABC\ldots N}_{bc\ldots n} \qquad [14]$$

En la expresión [9] se tomaba la variable dependiente (A) como dicotómica, sin embargo los modelos son también perfectamente posibles con variables politómicas (Hagenaars, 1990: 75). El hecho de considerar más de dos niveles en la variable dependiente supone un considerable aumento de la formulación debido a la multiplicación de parámetros en el proceso desarrollado desde la ecuación [9] a la [12], aunque el resultado final es bastante similar al obtenido en la ecuación [12].

Para indicar que la variable dependiente A tiene más de dos niveles se suele introducir el elemento subíndice a/a', con lo que la expresión [12] quedaría de la manera siguiente:

$$\Phi^{A}_{a/a'\,bc} = \beta^{A}_{a/a'} + \beta^{AB}_{a/a'\,b} + \beta^{AC}_{a/a'\,c} + \beta^{ABC}_{a/a'\,bc} \qquad [15]$$

donde $a \neq a'$

Y el modelo general de la expresión [14], considerando la variable dependiente como politómica, toma la forma:

$$\Phi^{A}_{a/a'\,bcd\ldots n} = \beta^{A}_{a/a'} + \beta^{AB}_{a/a'\,b} + \beta^{AC}_{a/a'\,c} + \beta^{ABC}_{a/a'\,bc} + \ldots + \beta^{ABC\ldots N}_{a/a'\,bc\ldots n} \qquad [16]$$

Como mostraron Bakeman, Adamson y Strisik (1989) la perspectiva asimétrica de los modelos logit es perfectamente compatible con el diseño secuencial. Simplemente basta con situar una variable consecuente en el tiempo $t + k$ como variable dependiente y el resto (v. g. antecedentes en los tiempos $t + k$ -1, $t + k$ - 2, $t + k$ - 3,..., t, registros realizados por diferentes codificadores, o en diferentes momentos, o con otros instrumentos...), como independientes.

3. SELECCIÓN E INTERPRETACIÓN DE LOS MODELOS

Tras la construcción de los modelos el proceso del análisis log-lineal debe continuar con la *evaluación* de los mismos, midiendo la bondad del ajuste a los datos, la *selección* del modelo más adecuado, la *estimación* de los efectos (parámetros) y la *interpretación* de los resultados.

Para la evaluación de los modelos se pueden utilizar dos técnicas: la prueba χ^2 de Pearson y la razón de verosimilitud ($LR\chi^2$). Los grados de libertad vienen determina-

dos por el número de τs iguales a 1.00 (Davis, 1974). Cuando el valor del estadístico hallado es alto, el modelo hipotético no se ajusta a los datos y debe ser rechazado como una inadecuada representación de las relaciones entre variables. Lo deseable, para aceptar el modelo, es que el estadístico sea bajo en relación a los grados de libertad y que por tanto la probabilidad exacta sea grande. Knoke y Burke (1980) proponen aceptar un modelo como adecuado si la probabilidad exacta se encuentra entre 0.10 y 0.35, puesto que una probabilidad más alta indicaría un *ajuste demasiado elevado* que traería consigo la inclusión en el modelo de parámetros redundantes.

Cuando se utiliza una muestra grande y el modelo es rechazado (v. g. $p_{(\alpha)} < 0.05$), las dos pruebas, $LR\chi^2$ y χ^2, se espera que tengan el mismo valor. Cuando el tamaño de la muestra es pequeño o los datos están escasamente repartidos en la tabla, cada una de ellas tiene una distribución muestral diferente (incluso de la distribución de probabilidad χ^2).

Si un tamaño de la muestra pequeño coincide con un gran número de celdas y la hipótesis nula es verdad, ambos estadísticos se aproximan a la distribución teórica de χ^2. Haberman (1978) y Fienberg (1980) concluyeron que la aproximación a la distribución de probabilidad χ^2 del estadístico χ^2 de Pearson era bastante buena (incluso con casillas con una frecuencia esperada igual a 1). En este caso $LR\chi^2$ subestima la probabilidad del error tipo I.

Cuando se utiliza una cantidad ingente de datos el problema está en que sólo es aceptable el modelo saturado. Este efecto se produce porque un gran número de datos magnifica leves diferencias entre las frecuencias esperadas y observadas con lo que los modelos son rechazados con mucha facilidad. Knoke y Burke (1980: 41) proponen evitar este problema utilizando el coeficiente de determinación R^2 para determinar el tanto por cien que es explicado de un modelo situado como línea de base por uno alternativo.

$$R^2 = \frac{(LRX^2 \text{ del modelo línea base}) - (LRX^2 \text{ del modelo alternativo})}{(LRX^2 \text{ del modelo línea base})} \qquad [17]$$

Si el modelo alternativo explica el 90% o más del modelo línea base, el alternativo puede considerarse como un buen ajuste a los datos incluso cuando el contraste sea significativo.

Según Knoke y Burke (1980: 30), $LR\chi^2$ es preferible a χ^2 ya que las frecuencias de las casillas son estimadas mediante la técnica de máxima verosimilitud y además, puede ser *separado en partes* para pruebas con mayor potencia de independencia condicional en tablas multivariantes.

Para seleccionar qué modelo es el más adecuado existen varios procedimientos. Uno de los más utilizados es el *paso a paso* (*stepwise*), que consiste en ir añadiendo, o suprimiendo, términos a un modelo inicial. Cuando se parte del modelo saturado y se van eliminando términos evaluando cada uno de los modelos que se obtiene el procedimiento se denomina *backward* o hacia atrás (Benedetti y Brown, 1978 y Upton, 1978). Knoke y Burke (1980) por su parte proponen el procedimiento opuesto *foward*,

o hacia adelante, basándose en que se utilizan modelos más parsimoniosos de partida.

Otra técnica posible es la de criba (*screening*), que consiste en estudiar la asociación parcial y marginal entre variables para localizar los términos que producen efectos (Brown, 1976).

También es posible examinar todos los parámetros en el modelo saturado y eliminar todos los términos cuyos parámetros no aportan nada al modelo (λs iguales a cero o τs iguales a uno, según si se haga o no la transformación logarítmica).

En igualdad de condiciones (significación estadística), lo mejor es guiarse por seleccionar el modelo parsimonioso y que posea una interpretación sustantiva (Bisquerra, 1989: 590).

Una vez que se ha seleccionado un modelo parsimonioso, que se ajusta a los datos y que es sustantivamente interpretable, conviene estimar los parámetros (λs) para determinar la contribución de cada uno de ellos en el modelo.

En general, la interpretación del análisis es bastante simple (Kennedy, 1983), ya que existe una gran coincidencia conceptual con la interpretación del análisis de la variancia (sobre todo en los modelos *logit*). Especial atención merecen las interacciones de primer orden y de órdenes superiores que, teniendo en cuenta cada uno de los niveles de las variables que interactúan, pueden tener una interpretación algo más compleja (Iacobucci y Wasserman, 1988; Elliott, 1988 y DeMaris, 1991).

Otra ventaja importante es la *gran potencia estadística* (frente al control complejo de los errores tipo I y tipo II en múltiples pruebas bivariantes), que no impone limitaciones en cuanto a la forma de la distribución ni en cuanto a la homoscedasticidad (Kennedy, 1983).

4. OTRAS POSIBILIDADES EN EL ANÁLISIS LOG-LINEAL

Las limitaciones que sugieren los modelos log-lineales no son demasiado importantes. Por ejemplo, las frecuencias esperadas bajas o muy bajas reducen la potencia estadística, todas las frecuencias esperadas han de ser como mínimo igual a 1 y el 80% debería ser igual a 5 o mayor (Bakeman, Adamson y Strisik, 1989). Cuando aparecen ceros en algunas celdas de las tablas no se puede esperar que las pruebas estadísticas $LR\chi^2$ y χ^2 se aproximen a la distribución teórica de χ^2 (Hagenaars, 1990).

Para solucionar estos problemas Zelterman (1987) desarrolló pruebas de χ^2 alternativas. No obstante, con ellas no se resuelve el problema de la potencia estadística, no están implementadas en los paquetes estadísticos estándares y la ventaja sobre las pruebas tradicionales no está clara (Hagenaars, 1990).

Existen una serie de soluciones más clásicas al problema de los ceros en las tablas.

Los ceros producidos en las frecuencias esperadas por pequeñas muestras y/o muchas categorías por variable (lo que conlleva lógicamente bajas frecuencias observadas de algunas categorías), son conocidos como *ceros muestrales* o *aleatorios*. Este problema puede intentar paliarse:

1) aumentando el tamaño de la muestra, cuando sea posible,
2) *colapsando*[1] tablas sobre variables o categorías cuando sea posible,
3) sumando una cierta y pequeña cantidad fija a todas las casillas (v. g. 0.5), como sugería Goodman (1970), que resulta ser un procedimiento restrictivo pues subestima los parámetros y la significación en la bondad del ajuste, y
4) definiendo arbitrariamente que el cociente 0/0 es igual a cero (Fienberg, 1980).

Si los ceros se producen como resultado de una tabla incompleta, porque no sean posibles ciertos cruces entre categorías de variables (ceros lógicos o *estructurales*), es preciso construir un *modelo de cuasi-independencia* (v. Knoke y Burke, 1980; Hagenaars, 1990). La cuasi-independencia *es una forma de independencia o no asociación entre variables cuando se considera sólo aquella porción de tabla que no contiene ningún cero* (Knoke y Burke, 1980: 64).

Otros modelos posibles en análisis log-lineal son los de simetría y cuasi-simetría. Los primeros consisten en dividir en dos triángulos con idéntica forma (simétrica) la matriz de la tabla de frecuencia eliminando la diagonal principal y se formula considerando las frecuencias del triángulo superior iguales a las del inferior (Bishop, Fienberg y Holland, 1975):

$$F_{ab}^{AB} = F_{ba}^{AB} \qquad \forall\, a \neq b \qquad [18]$$

En la expresión [1], el modelo de simetría quedaría:

$$\widehat{F}_{ab} = \eta\, \tau_a^A\, \tau_b^B\, \tau_{ab}^{AB} \qquad [19]$$

con las restricciones:

$$\tau_a^A = \tau_b^B \qquad \forall\, a = b \qquad [20]$$

$$\tau_{ab}^{AB} = \tau_{ba}^{AB} \qquad \forall\, a, b \qquad [21]$$

Los efectos de los parámetros que determinan el valor de F_{ab}^{AB} son iguales a los efectos de F_{ba}^{AB}, debido a ambas restricciones. Y las frecuencias esperadas serán (Bishop, Fienberg y Holland, 1975: 283):

$$\widehat{F}_{ab}^{AB} = \frac{1}{2}\left(F_{ab}^{AB} + F_{ba}^{AB} \right) \qquad \text{cuando } a \neq b \qquad [22]$$

---

1 El teorema de la *colapsabilidad*, formulado por Bishop, Fienberg y Holland (1975), aclara bajo qué circunstancias los efectos de modelos jerárquicos cambian cuando se introducen variables adicionales y bajo qué circunstancias el dejar a un lado ciertas variables y analizar tablas marginales particulares conduce a conclusiones diferentes. De una forma muy resumida, un grupo de variables es susceptible de ser colapsado con respecto a los t términos que incluye un segundo grupo, y no con respecto a un tercero, si y sólo si los dos primeros grupos son independientes uno de otro.

$$\hat{F}_{ab}^{AB} = F_{ab}^{AB} \qquad \text{cuando } a \neq b \qquad [23]$$

El modelo de cuasi-simetría es muy similar al anterior. La formulación del modelo puede ser la misma que en la ecuación [1], con la única restricción de la expresión [21]. Cuando existe homogeneidad marginal (las distribuciones marginales son iguales) los resultados coinciden con los producidos por un modelo de simetría (Bishop, Fienberg y Holland, 1975).

El análisis secuencial mediante los modelos log-lineales es incluso posible cuando se considera la *intensidad* de las categorías. En este caso es necesario utilizar los *modelos log-lineales con datos ordinales* (ver p. ej. Haberman, 1978, 1979). El uso de datos ordinales produce modelos más parsimoniosos (Agresti, 1984). Como ejemplo se puede consultar Feick y Novak (1985), que presentan un análisis secuencial de la interacción diádica (padre-hijo) en torno a la preferencia en asistencia a convenciones.

Por último, para cerrar este breve repaso a los modelos log-lineales, se va a recoger un interesante paralelismo que se produce entre los modelos logit y los *modelos de paso o causales* El análisis causal hace referencia al uso de una serie de estrategias capaces de elaborar modelos que permitan explicar fenómenos (relaciones causa-efecto) mediante diagramas de paso y *ecuaciones estructurales* (v. p. e. Heise, 1975 y Visauta, 1986). En el terreno de la psicopedagogía por ejemplo, Gómez (1986) aplicó los modelos causales a cuestiones de validez de constructo.

Goodman (1972a, 1979) presentó el paralelismo entre el análisis causal con variables continuas y los modelos logit con variables categóricas. Las analogías se pueden cifrar más en la forma (fases y esquematización) que en el fondo (diferente formalización matemática, específica estimación de parámetros,...).

Las fases en ambas técnicas pueden ser muy similares, con la diferencia de que en el análisis causal la estimación de los parámetros se realiza con anterioridad a la evaluación del modelo. El *diagrama de paso* (propio de los modelos causales), por su parte, puede ser un esquema muy eficaz para representar los análisis propuestos mediante los modelos logit. En dichos diagramas incluso se pueden incluir las estimaciones de los parámetros de los efectos aunque hayan sido realizados por un procedimiento diferente.

Las diferencias entre ambos enfoques se podrían resumir en las siguientes:

— En los modelos de paso los coeficientes (parámetros estructurales estimados) se pueden descomponer en efectos causales *totales* (tanto *directos* como *indirectos*), en efectos conjuntos y en relaciones *espúreas* (Saris y Stronkhorst, 1984). Dicha descomposición no es posible en los parámetros estimados en los modelos log-lineales.

— En el análisis causal los modelos suelen clasificarse en *no recursivos*, cuando presentan efectos bidireccionales, y *recursivos*, si no presentan efectos causales recíprocos (Heise, 1975). Los efectos bidireccionales de los modelos no recursivos no son comparables con el enfoque logit.

— El parámetro estimado entre dos variables (coeficiente) en un diagrama de paso de un modelo causal representa un único valor. Cuando se utilizan variables politómicas en los modelos logit esto no es posible, ya que en general, en el análisis log-lineal se estima un parámetro por cada nivel de cada variable (sin contar los parámetros de la interacción).

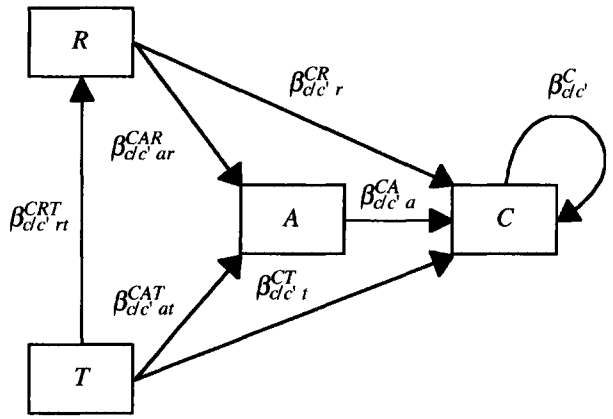

Figura 1. *Diagrama de paso del modelo logit de la ecuación [24].*

Supóngase el siguiente modelo con cuatro variables *T*, *R*, *A* e *C*:

$$\Phi^{C}_{c/c'\,art} = \beta^{C}_{c/c'} + \beta^{CA}_{c/c'\,a} + \beta^{CR}_{c/c'\,r} + \beta^{CT}_{c/c'\,t} + \beta^{CAR}_{c/c'\,ar} + \beta^{CAT}_{c/c'\,at} + \beta^{CRT}_{c/c'\,rt} \qquad [24]$$

donde *A* es un sistema de dos categorías (eventos o estados) que ocurren en el tiempo *t*,
R es la variable registro, tal que r_1 y r_2 son dos codificadores independientes,
T representa dos escenarios diferentes, y
C un sistema de dos categorías (eventos o estados) que ocurren en el tiempo $t + 1$.

El modelo de la ecuación [24] es un modelo logit jerárquico insaturado. Si este modelo representara adecuadamente los datos indicaría que todos los efectos excepto el de orden máximo ($\beta^{CART}_{c/c'\,art}$) son necesarios para explicar la variable dependiente *C* (las categorías que ocurren en el tiempo $t + 1$) y generar un patrón de datos como el observado. Este modelo se puede representar en un diagrama de paso como el de la figura 1.

5. CONCORDANCIA Y ANÁLISIS SECUENCIAL

En otro lugar (Tójar, 1994) se mostró como los modelos log-lineales, y especialmente los logit, pueden ser utilizados para abordar la concordancia secuencial. De esta forma, los modelos log-lineales no sólo pueden utilizarse para realizar el análisis secuencial, sino que pueden ser utilizados como una técnica de control de la calidad de los datos que se van a analizar.

Muy resumidamente, si *A* es una variable formada por las *a* categorías de un sistema tomadas como antecedentes (ocurrencia de las categorías en el tiempo *t*), *C* representa las *c* categorías de un sistema tomadas como consecuentes (y por tanto ocurrencia de las categorías en el tiempo $t + 1$), y *R* la variable registro (con las *r* modalidades relativas por ejemplo a diferentes observadores), los modelos logit para evaluar la concordancia se pueden clasificar en tres tipos:

a) *Modelos tipo I*: las categorías consecuentes sólo dependen de ellas mismas. No hay secuencialidad, esto es, no es posible determinar patrones puesto que ninguna categoría consecuente es debida a una antecedente. Utilizando la notación propuesta por Goodman (1972b) este modelo tendría la forma:

$$\Phi^{C}_{c/c'\,ar} = \beta^{C}_{c/c'} \qquad [25]$$

b) *Modelos tipo II*: los consecuentes dependen de los antecedentes (secuencialidad), pero no del registro (concordancia). Este modelo denominado de concordancia secuencial tendría la forma:

$$\Phi^{C}_{c/c'\,ar} = \beta^{C}_{c/c'} + \beta^{CA}_{c/c'\,a} \qquad [26]$$

Este tipo de modelos sugiere que las *categorías consecuentes* (tomadas como variable dependiente) son explicadas, además de por sí mismas, por las *categorías antecedentes*. Toda vez que los datos proceden de diversos registros y la variable *registro* no aparece en el modelo como independiente, puesto que sus efectos son irrelevantes para el ajuste, se puede concluir que las opiniones de los diferentes codificadores concuerdan.

Los modelos tipo II pueden además representarse gráficamente en un diagrama de paso, obteniéndose la siguiente figura 2.

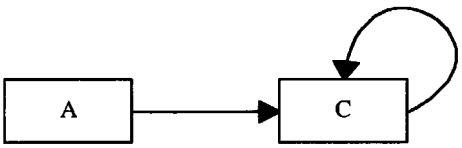

Figura 2. *Diagrama de paso de los modelos de concordancia secuencial.*

c) *Modelos tipo III*: los consecuentes dependen tanto de la interacción con los antecedentes (secuencialidad), como de la producida con los diferentes registros (discordancia). El modelo de *discordancia secuencial* tiene la forma:

$$\Phi^{C}_{c/c'\,ar} = \beta^{C}_{c/c'} + \beta^{CA}_{c/c'\,a} + \beta^{CR}_{c/c'\,r} \qquad [27]$$

Esta otra clase de modelos supone la inclusión del registro como variable explicativa, de manera que, según las diferentes modalidades del mismo, el resultado de las conexiones entre variables antecedentes y consecuentes difieren.

Si se representa este tipo de modelos mediante un diagrama de paso se obtiene la siguiente figura 3 La selección de este modelo (tipo III) sugiere que los registros no son intercambiables.

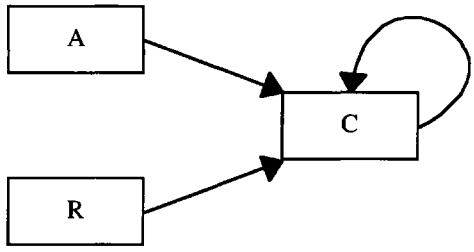

Figura 3. *Diagrama de paso de los modelos de discordancia secuencial.*

d) *Modelos tipo IV* (modelo saturado): los consecuentes se hayan influidos por ellos mismos, por los antecedentes, por los registros y por la interacción de estos dos últimos.

$$\Phi^{C}_{c/c'\,ar} = \beta^{C}_{c/c'} + \beta^{CA}_{c/c'\,a} + \beta^{CR}_{c/c'\,r} + \beta^{CAR}_{c/c'\,ar} \qquad [28]$$

Además de la bondad del ajuste, en cada modelo logit es interesante considerar un análisis de la dispersión así como la cuantificación de la magnitud de asociación entre las variables (Haberman, 1982). Las dos medidas de asociación comúnmente utilizadas para medir la magnitud de la asociación entre las variables que forman parte de un modelo logit son la *concentración* y la *entropía*.

La concentración es un estadístico similar a la τ_Y de Goodman y Kruskal (1954). τ_Y es una medida de *Reducción Proporcional del Error*. Las medidas de asociación *RPE* se basan en dos simples reglas que pueden formularse respectivamente en dos probabilidades: *P(A)*, o la probabilidad de cometer error al predecir los valores de una variable sin tener en cuenta, ningún conocimiento, la otra, y *P(B)*, o la probabilidad de error cuando se tiene conocimiento de una variable a la hora de realizar las predicciones.

Lógicamente, la variable sobre la que se realizan las predicciones se considera dependiente y la explicativa como independiente. Las observaciones individuales seleccionadas al azar son asignadas con o sin conocimiento de los valores de una variable (la independiente), teniendo en cuenta que estas asignaciones se realizan respetando las distribuciones originales, esto es, las distribuciones de observaciones estimadas coinciden con las originales.

En su expresión general un coeficiente *RPE* toma la forma:

$$RPE = \frac{P(A) - P(B)}{P(A)} \qquad [29]$$

τ_Y se calcula así (sólo una variable independiente):

$$\tau_Y = \frac{\sum_i F_{i.}\left(\frac{\sum_{i'} F_{i'.}}{F_{..}}\right) - \sum_j \left[\sum_i F_{ij}\left(\frac{\sum_{i'} F_{i'j}}{F_{.j}}\right)\right]}{\sum_i F_{i.}\left(\frac{\sum_{i'} F_{i'.}}{F_{..}}\right)} \qquad [30]$$

donde $i' \neq i$

$\sum_i$ significa la suma de todas las filas excepto la i-ésima

$\sum_i F_{i.} \left(\frac{\sum_{i'} F_{i'.}}{F_{..}} \right)$ es $P(A)$, o la probabilidad de cometer error al predecir los valores de una variable (dependiente) sin tener ningún conocimiento de la independiente y

$\sum_j \left[\sum_i F_{ij} \left(\frac{\sum_{i'} F_{i'j}}{F_{.j}} \right) \right]$ es $P(B)$, o la probabilidad de error cuando se tiene conocimiento de la variable independiente a la hora de realizar las predicciones en la dependiente.

La entropía es un estadístico definido por Theil (1970) como coeficiente de incertidumbre. En el caso de dos variables el coeficiente de incertidumbre vale:

$$U = - \frac{\sum_i \sum_j p_{ij} \log \left(\frac{p_{ij}}{p_{i.} p_{.j}} \right)}{\sum_j p_{.j} \log p_{.j}} \qquad [31]$$

Tanto τ_Y como U son medidas de reducción proporcional del error. Si $\tau_Y = U = 0$ existe independencia entre las dos variables. Si en el otro extremo $\tau_Y = U = 1$, al conocer la variable independiente no existirá incertidumbre en la información sobre la dependiente. Sin embargo, con ambas medidas es difícil determinar con un valor intermedio cuando se produce una fuerte asociación. A medida que la variable independiente aumenta el número de categorías, ambas medidas tienden a ofrecer valores próximos a 0 (Agresti, 1990).

En el caso de un modelo logit de más de dos variables (o más de una variable independiente), la entropía se puede obtener del cociente entre la dispersión de la entropía debida al modelo y la dispersión total (suma de la dispersión debida al modelo y la dispersión residual). La concentración se obtiene con el cociente entre la dispersión de la concentración debida al modelo y la dispersión total (suma de la dispersión debida al modelo y la dispersión residual) (Haberman, 1982).

Un aspecto de especial relevancia consiste en concretar la estructura del desacuerdo. Esto se puede realizar a partir de las *secuencias discordantes significativas* que se ponen de manifiesto estudiando los residuales entre las frecuencias observadas y esperadas. Una medida simple de discrepancia son los residuales estandarizados:

$$z = \frac{F - \hat{F}}{\sqrt{\hat{F}}} \qquad [32]$$

Sin embargo, esta medida tiende a subestimar el valor real (Bakeman, Adamson y Strisik, 1989). Una mejor aproximación a z la ofrece el *residual ajustado* (Haberman,

1978: 77-79). El residual ajustado es además conceptualmente equivalente a la medida sugerida por Gottman (1980) y Allison y Liker (1982) para realizar el análisis secuencial (z*).

Los resultados del análisis de la concordancia secuencial se apoyan en la construcción y selección de modelos (logit o log-lineales), en base a la capacidad de ajuste (significación estadística) y al análisis de los residuales, que permite concretar las secuencias concordantes y discordantes.

Cuando se ha determinado la estructura de la concordancia de los registros observados es importante dirigir la mirada sobre la propia *estructura secuencial* de los datos observacionales. Esto es posible gracias a un estudio de los *parámetros estimados* por el modelo seleccionado.

De esta forma, una vez seleccionado un modelo parsimonioso, ajustado y que, de la manera que ha sido formulado, sea susceptible de una interpretación acorde a la teoría y a las hipótesis propuestas, cabe fijar la atención en estimar los parámetros (*l* o *b* según la notación), para determinar qué peso tiene cada uno en la contribución del modelo.

Un medio de concretar la importancia relativa que los parámetros tienen sobre el ajuste del modelo consiste en realizar un simple contraste dividiendo cada uno de ellos por su error estándar y comparando el resultado con un valor crítico de la distribución normal estandarizada (Kennedy, 1983: 149):

$$z' = \frac{\lambda_i}{EE_{\lambda_i}} \qquad [33]$$

donde λ_i son los parámetros estimados a partir del modelo seleccionado, y

$EE_{\lambda i}$ es la desviación estándar asintótica estimada, que para un modelo log-lineal saturado con tres variables *AxBxC* vale (Haberman, 1978: 232):

$$EE_{\lambda_{ijk}} = \frac{1}{abc}\sqrt{\sum_i \sum_j \sum_k \frac{1}{F_{ijk}}} \qquad [34]$$

donde *abc* son el número total de celdas.

Los parámetros que significativamente destaquen por encima del resto mostrarán qué efectos (ya sean simples o interacciones entre ellos), son los principales protagonistas del ajuste del modelo seleccionado, esto es, se determinarán patrones estocásticos y se evaluará el efecto del resto de variables en la estructura secuencial, objetivos del análisis secuencial (Gottman y Roy, 1989).

REFERENCIAS

AGRESTI, A. (1984). *Analysis of ordinal Categorial data.* New York: John Wiley & Sons.

AGRESTI, A. (1990). *Categorial data analysis.* New York: John Wiley & Sons.

ALLISON, P. D. y LIKER, J. K. (1982). Analyzing Sequential Categorial data on Dyadic Interaction: A Comment on Gottman. *Psychological Bulletin, 91* (2), 393-403.

BAKEMAN, R. (1991). *A Brief Introduction to Sequential Analysis and Loglinear Analysis.* Conferencia pronunciada en la Facultad de Psicologia de la Universidad Central de Barcelona (13/3/91).

BAKEMAN, R., ADAMSON, L. B. y STRISIK, P. (1989). Lags and Logs: Statistical Aproaches to Interaction. En M. H. BORNSTEIN y J. BRUNER (Eds.). *Interaction in Human Development* (pp. 241-260). Hillsdale, N. J.: Erlbaum.

BENEDETTI, J. K. y BROWN, M. B. (1978). Strategies for the selection of log-linear models. *Biometrics, 34,* 680-686.

BISHOP, Y. M. M., FIENBERG, S. E. y HOLLAND, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambrige: MIT Press.

BISQUERRA, R. (1989). *Introducción conceptual al análisis Multivariante. Un enfoque informático con los paquetes SPSS-X, BMDP, LISREL, y SPAD* (Vol. II). Barcelona: PPU.

BROWN, M. B. (1976). Sreening effects in multidimensional contingency tables. *Aplications of Statistics, 25,* 37-46.

DAVIS, J. A. (1974). Hierarchical models for significance tests in multivariate contingency tables: exegesis of Goodman's recent papers. En H. L. COSTNER (Ed.). *Sociological Methodology 1973-1974* (pp. 189-231). San Francisco: Jossey-Bass.

DEMARIS, A. (1991). A framework for the interpretation of first-order interaction in logit modeling. *Psychological Bulletin, 110* (3), 557-570.

ELLIOTT, G. C. (1988). Interpreting higher order interactions in log-linear analysis. *Psychological Bulletin, 103* (1), 121-130.

FEICK, L. F. y NOVAK, J. A. (1985). Analyzing Sequential categorial data on dyadic interaction: Log-linear models exploiting the orden in variables. *Psychological Bulletin, 98* (3), 600-611.

FIENBERG, S. E. (1980). *The analysis of cross-classified categorial data* (2nd Ed.), Cambridge: MIT Press.

GOODMAN, L. A. (1972a). A modified multiple regression approach to the analysis of dichotomous variables. *American Sociological Review, 37,* 28-46.

GOODMAN, L. A. (1972b). A general model for the analysis of surveys. *American Journal of Sociology, 77,* 1.035-1.086.

GOODMAN, L. A. (1979). A brief guide to the causal analysis of data from surbeys. *American Journal of Sociology, 84,* 1.078-1.095.

GOODMAN, L. A. y KRUSKAL, W. H. (1954). Measures of association for cross-clasifications, Part I. *Journal of the American Statistical Association, 49,* 732-764.

GÓMEZ, J. (1986). *Los modelos causales como metolología de validez de constructo.* Barcelona: Alamex.

GOTTMAN, J. M. (1980). On analyzing for sequential connection and assessing interobsever reliability for the sequential analysis of observational data. *Behavioral Assessment, 2,* 361-368.

GOTTMAN, J. M. y ROY, A. K. (1990). *Sequential analysis. Aguide for behavioral researchers.* Cambridge: Cambridge University Press.

HABERMAN, S. J. (1978). *Analysis of cualitative data. Volume 1. Introductory topics.* New York: Academic Press.

HABERMAN, S. J. (1979). *Analysis of cualitative data. Volume 2. New Developments.* New York: Academic Press.

HABERMAN, S. J. (1982). Análysis of dispersion of multinomial responses. *Journal of the American Statistical Association, 77,* 568-580.

HAGENAARS, J. A. (1990). *Categorial longitudinal data.* Newbury Park, CA: Sage.

HEISE, D. R. (1975). *Causal analysis.* Nueva York: John Wiley.

IACOBUCCI, D. y WASSERMAN, S. (1988). A General Framework for the Statistical Analysis of Sequential Dyadic Interactional Data. *Psychological Bulletin, 103* (3), 379-390.

KENNEDY, J. J. (1983). *Analyzing qualitative data: Introductory log-linear analysis for behavioral research.* New York: Praeger.

KNOKE, D. y BURKE, P. J. (1980). *Log-linear Models.* Beverly Hills, CA: Sage.

RINDSKOPF, D. (1990). Nonsatndard log-linear models. *Psychological Bulletin, 108* (1), 150-162.

SARIS, W. E. y STRONKHORST, L. H. (1984). *Introduction to causal models in non-experimental research.* Amsterdam: Sociometrics Research Foundation.

SPSS Inc (1988). *SPSS-X. User's Guide.* Chicago: McGraw-Hill.

SPSS Inc. (1990). *SPSS for the Macintosh v. 4.0.* Chicago: SPSS Inc.

THEIL, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology, 76,* 103-154.

TÓJAR, J. C. (1993). *Concordancia del registro observacional en datos secuenciales. Investigación aplicada en el contexto del aula.* Málaga: Secretariado de Publicaciones de la Universidad de Málaga.

TÓJAR, J. C. (1994). *Concordancia en los registros de observación. Calidad de la investigación educativa en Metodología Observacional.* Barcelona: PPU.

TÓJAR, J. C. (En prensa). Classroom interaction assessment throught sequential analysis of observational data. *European Journal of Psychologycal Assessment.*

UPTON, G. (1978). *The analysis of cross-tabulated data.* New York: John Wiley.

VISAUTA, B. (1986). *Técnicas de investigación social. Modelos causales.* Barcelona: Hispano Europea.

ZELTERMAN, D. (1987). Goodness-of-fit tests for large sparse multinomial distributions. *Journal of the American Statistical Association, 82,* 624-629.