TRABAJOS METODOLÓGICOS

Revista Investigación Educativa - Vol. 9 - nº 18 - 1991 (P. 81-96)

PROBLEMAS FUNDAMENTALES DEL ANÁLISIS LOGARÍTMICO LINEAL (I): EL COLAPSAMIENTO DE VARIABLES Y SU INFLUENCIA EN EL AJUSTE DE MODELOS

por
Ana Delia Correa Piñero
Dpto. Didáctico e Investigación Educativa
Universidad de la Laguna

RESUMEN

En este trabajo analizamos algunos de los problemas que pueden surgir en el ajuste de modelos logarítmicos lineales cuando se toman decisiones de colapsamiento (recombinar categorías, eliminar categorías o variables, categorizar datos continuos,...). Ésta es una práctica frecuente entre los investigadores cuando la tabla de contingencia presenta un exceso de casillas con valores esperados muy bajos. La influencia del colapsamiento puede manifestarse de diversas maneras. En unos casos, la modificación de las categorías es tan drástica que la variable subyacente, en cierta medida, cambia. En otros, pueden verse afectados los valores de los parámetros y la calidad de ajuste de los modelos. Finalmente, se presentan una serie de recursos mediante los cuales minimizar el efecto distorsionador del colapsamiento.

ABSTRACT

This paper analyze some problems that can raise in the application of log-linear models when the researcher make decisions of collapsing (recombining categories, deleting categories or variables, treating continuous data as if they were discrete,...). This is a common practice among researchers when they obtain too many cells with small expected values. The effects of collapsing can be observed in different ways. In some cases, a drastic modification of categories can lead to changing the underlying variable. In other cases, the value of parameters and models goodness of fit can be affected. Finally, it is presented a serie of recourses to reduce the distorting effect of collapsing.

1. INTRODUCCIÓN

Durante mucho tiempo el análisis estadístico de variables nominales en la investigación educativa se limitó a técnicas sumamente sencillas, donde —bajo un enfoque univariado— se describían las diversas categorías de la variable X en términos de su frecuencia absoluta o bien su porcentaje de ocurrencias empleando las conocidas tablas de distribución de frecuencias.

Dando un paso más —bajo un enfoque bivariado— se establecía el porcentaje de ocurrencia conjunta de las categorías pertenecientes a dos variables distintas (X e Y), a partir de tablas de contingencia bidimensionales. Bajo este último enfoque también se aplican pruebas de independencia (como ji^2) y diversos índices de asociación entre X e Y (como la Q de Yule, el coeficiente de contigencia, phi, etc.) elegidos en función de diversos criterios restrictivos (número de categorías, tablas cuadradas o no cuadradas, tamaño de la muestra, sesgo de los marginales, etc.)

Ocasionalmente, se incluye una tercera variable (Z) en el análisis, empleando medidas de asociación parcial que describan el grado de independencia condicional o de independencia mutua, mediante el control de las categorías de Z mientras se analiza la relación entre X e Y para cada categoría de la variable control Z separadamente, o mediante la unificación de dos de las tres variables en una única de categorías combinadas (XZ o YZ).

Parece obvio que la selección de las técnicas de análisis estadístico apropiadas en cada investigación no debería depender de un criterio superficial que oponga sofisticación frente a simplicidad, o antigüedad frente a novedad. Es decir, que esos procedimientos, simples y antiguos, continúan siendo válidos siempre que los propósitos de la investigación no vayan más allá de las soluciones que esas viejas técnicas nos proporcionan. Pero no siempre se da el caso de que sean suficientes para ciertos problemas de investigación. En concreto, el análisis de datos nominales más complejos (y, entre otras cosas, entendemos «más complejos» como «con más variables»), que se organizan en tablas de contingencia multidimensionales —donde el número y tipo de problemas a plantearse crece rápidamente—, no puede ser resuelto satisfactoriamente por esos antiguos procedimientos. Son necesarios otros métodos estadísticos para analizar las interacciones complejas que pueden subyacer en tablas de frecuencias multivariadas.

Los investigadores en Ciencias Sociales muestran un interés creciente por un procedimiento relativamente nuevo (nuevo en cuanto a su conocimiento y utilización por los investigadores, aunque no tanto en cuanto a sus orígenes y desarrollo) conocido como análisis logarítmico lineal o modelos logarítmico lineales (en adelante MLL). Brevemente descrita, es una técnica de análisis estadístico multivariado aplicada a variables nominales. Su gran utilidad en la investigación educativa se basa, fundamentalmente, en dos cosas: una, la abundancia de variables de tipo cualitativo (categorial) en la investigación pedagógica, bien por la propia naturaleza de las variables o bien porque la carencia de instrumentos de medida más precisos para ciertos constructos no permiten obtener datos continuos; y otra, su carácter de

técnica multivariada, que permite analizar problemas de investigación más complejos.

En muchos aspectos, los MLL siguen un esquema de trabajo similar al de otros procedimientos estadísticos ya bien establecidos, como el ANOVA o el análisis de regresión (BAKER, 1981), con lo que el análisis estadístico de variables categoriales alcanza un nivel de sofisticación sólo disponible hasta ahora para variables continuas.

Su finalidad es obtener un modelo que describa adecuadamente las relaciones e interacciones entre las variables de una tabla de contingencia multidimensional. Es decir, un modelo que se ajuste a las frecuencias observadas. Las relaciones e interacciones entre variables pueden ordenarse por su importancia (por su contribución al ajuste del modelo), averiguar sus parámetros y determinar su significación. A diferencia de la mayoría de los métodos, en los MLL la unidad de análisis no son las puntuaciones individuales, sino conjuntos de sujetos que comparten ciertas características específicas por las categorías de las variables. Y lo que usualmente se entiende por variable dependiente también debe ser reconceptualizado, ya que no es una variable en el usual sentido de la palabra, sino una probabilidad de casilla: la probabilidad de que un individuo seleccionado al azar pertenezca a determinadas categorías de interés en lugar de a otras. Es decir, la probabilidad de que tenga una determinada combinación de características (KNOKE y BURKA, 1982). Por ejemplo, en una tabla IxJxK, P_{ijk} es la probabilidad de que un sujeto pertenezca a la categoría i de la primera variable, a la categoría j de la segunda y a la categoría k de la tercera. Este conjunto de probabilidades, o alguna función derivada, es lo que sirve como «variable dependiente».

Sin embargo, en muchos otros sentidos, el análisis MLL es semejante a otros procedimientos de construcción de modelos. Su objetivo es obtener un modelo o ecuación que explique las variaciones en las probabilidades de las casillas, postulando una serie de relaciones de diverso tipo entre las variables.

No es propósito de este trabajo hacer una descripción general del procedimiento (aunque lo revisaremos brevemente), cosa que ya ha sido hecha por muchos autores en diversos trabajos: a partir de BIRCH (1963), a quien se atribuye el origen indirecto de los MLL, a raíz de un trabajo donde plantea problemas de asociación entre tres variables, el estudio de la relación entre variables en tablas de contingencia multidimensionales ha venido siendo desarrollado por una serie de autores bajo una diversidad de enfoques, entre los que se encuentra el de MLL. Entre ellos, destacaremos a L.A. Goodman, un matemático que ha extendido el modelo, su interpretación y sus aplicaciones de forma considerable, construyendo a lo largo de diversas obras un compacto conjunto de procedimientos de estimación, pruebas de hipótesis, estimación de parámetros, etc. (GOODMAN 1968, 1970, 1971, 1972a, 1972b, 1973a, 1973b, 1984). También BAKER (1981), BENEDETTI y BROWN (1978a, 1978b), BISHOP, FIENBERG y HOLLAND (1975), BROWN (1976), FIENBERG (1977), HABERMAN (1974, 1978, 1979), KNOKE y BURKE (1982), PAYNE (1977) y UPTON (1978, 1981), entre otros, han contribuido a su perfeccionamiento y/o popularización. En nuestro país, los trabajos de SÁNCHEZ CARRIÓN (1984),

TEJEDOR (1985), TEJEDOR y CARIDE (1988), TEJEDOR y GODÁS (1988), JORNET y SUÁREZ (1988), CORREA (1989, 1991) entre otros, han venido a cubrir la carencia de trabajos en castellano, tanto con propósitos de explicación y divulgación de los MLL como técnica estadística, como haciendo uso de la misma en diversos campos de investigación.

Nuestro objetivo se centra en profundizar en uno de sus aspectos: las alteraciones que pueden provocar diversas formas de colapsamiento (explícito o implícito) de las categorías de las variables en el ajuste de MLL y en el valor de los parámetros del modelo. El colapsamiento (junto con el problema relativo al tamaño de la muestra y el problema de las casillas extremas o análisis de residuales, que trataremos en otros trabajos) es, a nuestro juicio, uno de los más relevantes y no suficientemente tratado en la literatura al respecto. No es, por tanto, todo lo conocido que sería deseable entre los investigadores que aplican este procedimiento en sus estudios. El objetivo primordial de este trabajo, por consiguiente, es de difusión o divulgación de esta problemática entre los investigadores en Ciencias de la Educación (y, por supuesto, otras Ciencias Sociales), sin especiales conocimientos avanzados de estadística, que deseen aplicar este procedimiento de análisis a sus datos. También con ese propósito, optamos por una exposición que es, con frecuencia, más intuitiva que técnica.

2. BREVE REVISIÓN DEL ANÁLISIS LL

A título de ubicación mental y para refrescar nociones, recordemos los pasos fundamentales en el análisis logarítmico-lineal. En otro número de esta misma revista (SANS I MARTÍN, 1987) hay una introducción a la formulación de ecuaciones de modelos l-l mediante la transformación logarítmica del conocido índice de asociación «razón de razones» (o razón interproducto). Por nuestra parte, nos centraremos en recordar las etapas que nos llevan a encontrar un modelo satisfactorio que explique los datos observados: 1) En primer lugar, el investigador propone un modelo para explicar los datos observados (frecuencias empíricas obtenidas en cada casilla de la tabla multidimensional). Un modelo es una hipótesis acerca de las relaciones entre las variables. Si, por ejemplo, el investigador hipotetiza que las variables son mutuamente independientes, la formulación de su modelo debe reflejar tal independencia y no contener elementos de interacción. Es decir, si trabaja dos variables A y B, su modelo sería:

$$(A, B) \quad \text{(según la notación simplificada) o}$$
$$\mu+\mu^{A}+\mu^{B} \quad \text{(en forma de ecuación)}$$

De la misma forma, puede postular modelos con una interacción de dos variables:

$$(AB) \text{ o } \mu+\mu^{A}+\mu^{B}+\mu^{AB},$$

con dos interacciones de dos variables:

$$(AB, AC) \text{ o } \mu+\mu^{A}+\mu^{B}+\mu^{C}+\mu^{AB}+\mu^{AC},$$

con una interacción de tres variables:

$$(ABC) \text{ o } \mu+\mu^{A}+\mu^{B}+\mu^{C}+\mu^{AB}+\mu^{AC}+\mu^{BC}+\mu^{ABC},$$

etc., etc. Normalmente, los modelos surgen de la teoría, del conocimiento previo. Pero hay otros procedimientos —más exploratorios o inductivos— para formular MLL: por ejemplo, las pruebas de asociación parcial y marginal, las pruebas simultáneas, la adición o eliminación de efectos a partir de un modelo base, etc. (GOODMAN, 1971; BROWN, 1976; BENEDETTI y BROWN, 1978a, 1978b). Estos procedimientos son sumamente útiles en estudios de talante exploratorio, en los que el investigador no cuente con un sustrato teórico suficiente que le permita conjeturar a priori las relaciones que cabe esperar entre las variables.

2) A continuación, el investigador deriva un conjunto de expectativas bajo la suposición de que el modelo propuesto es cierto. Es como si se preguntara: si mi modelo (hipótesis) fuera cierto ¿cómo tendrían que ser los datos? Y procede a estimar estas expectativas, derivadas de su modelo, a partir de los datos que ha obtenido en la muestra. Si, por ejemplo, adoptase el modelo que asevera la independencia mutua, procedería a estimar «cómo tendría que ser» una muestra de tamaño dado si perteneciera a una población donde las variables fueran mutuamente independientes. En la práctica, este se traduce en calcular las frecuencias esperadas bajo el modelo, bien utilizando la rutina del ajuste proporcional iterativo de BARTTLET (que es la que emplean programas estadísticos computarizados como el P4F de BMDP o el programa ECTA de FAY y GOODMAN) o bien mediante el algoritmo de NEWTON-RAPHSON (que sería el caso del programa MULTIQUAL de BOCK o la sentencia LOGLINEAR del programa SUPSSX) (BISHOP et al, 1975; HABER y BROWN, 1986).

3) El siguiente paso del investigador consiste en comparar las expectativas derivadas del modelo (frecuencias esperadas o teóricas) con los datos obtenidos en la muestra (frecuencias observadas o empíricas) y decidir si el ajuste del modelo es o no aceptable. Si la muestra pertenece realmente a una población donde las variables analizadas son independientes, cualquier posible discrepancia entre los datos observados y las expectativas tendría que deberse al azar. Es decir, tiene que someter a prueba su modelo. La comprobación del ajuste del modelo se realiza en base a algún estadístico, como ji^2 o L^2 (razón de verosimilitud), que comparan frecuencias esperadas y observadas. La elección de L^2 permitiría, además, la comprobación de subhipótesis de interacción específicas, aislando ciertos componentes del modelo, debido a su propiedad de partición exacta, propiedad que no tiene ji^2 (SHAFFER, 1973a, 1973b; HALPERIN, NEHRKE, HULICKA y MORGANTI, 1976; FIENBERG, 1977). A la hora de analizar las discrepancias con el modelo, la pregunta fundamental es: ¿las discrepancias entre los valores esperados y los observados pueden ser atribuidas de forma razonable al azar o son tan considerables que el modelo mismo resulta equivocado?

4) Si las discrepancias son pequeñas, el investigador mantiene el modelo y da el siguiente paso. De lo contrario, vuelve al paso 1 y propone otro modelo que sea un «refinamiento» del modelo previo para llevar a cabo el análisis de la forma más

eficiente posible. Quizá el rechazo del modelo le obligue al replanteamiento de los fundamentos teóricos, o de las hipótesis. O posiblemente, si se piensa que las hipótesis estaban bien fundamentadas, a una revisión de los instrumentos de obtención de datos, del procedimiento para obtener la muestra, etc.

5) La última etapa del análisis, como en otros procedimientos de construcción de modelos, es estimar sus parámetros, sus errores típicos, intervalos de confianza, etc. Estos parámetros pueden ser traducidos a términos sustantivos y sirven de base para proceder a explicaciones y hacer predicciones. Se trata, pues, de la fase inferencial del análisis logarítmico lineal, tras la cual se procede, habitualmente, a exponer las conclusiones de la investigación.

Es probable que estas fases generales del análisis LL resulten familiares a la mayoría de los investigadores, ya que la prueba ji^2 ampliamente conocida y utilizada en el análisis de una tabla bivariada IxJ, contiene en sí misma los pasos 1 al 4. Como se sabe, esta prueba de independencia consiste en proporner unas frecuencias esperadas (bajo el modelo de independencia estadística) y compararlas con los valores observados. Si la prueba es significativa, se rechaza el modelo y se concluye que las variables están relacionadas; de lo contrario, se acepta la independencia como una adecuada descripción de la población.

Sin embargo, a la hora de efectuar un análisis LL, el investigador puede tomar, explícita o implícitamente, una serie de decisiones ajenas al LL en sí, pero que influyen notablemente sobre los resultados. Una de estas decisiones, con lo que ya entramos en lo que es el eje de este trabajo, puede ser la relativa a colapsar variables, o categorías de algunas variables, o categorizar lo que en principio era una variable continua. Todas estas decisiones pueden afectar, de uno o otro modo, los resultados que obtenga, de manera que es preciso que el investigador sea consciente de los efectos de su decisión.

3. FORMAS DE COLAPSAMIENTO EN VARIABLES CATEGORIALES

3.1. Colapsamiento intencional o explícito

En el análisis de tablas de contingencia multidimensionales el investigador quizá proceda, de forma intencional, a colapsar las categorías de una o más variables. Esto puede hacerse con dos propósitos: simplificar la interpretación de los resultados o evitar los ceros muestrales.

Obtenemos un cero muestral en una tabla de contingencia cuando alguna combinación de categorías, es decir, alguna casilla, no tiene referente en la muestra utilizada (su frecuencia es 0), aunque sí puede darse esa casilla en la población. Es distinto de un cero estructural, que resulta de combinaciones «imposibles» de categorías, sin referente en la población. Cuando existen estos ceros estructurales, se

trabaja con tablas truncadas o incompletas, para las cuales existen las llamadas pruebas de cuasi-independencia si la tabla en cuestión es bivariada. Para tablas mayores el problema se complica y algunos autores sugieren lo que ellos llaman modelos cuasi-logarítmico lineales (FIENBERG, 1977).

Aún podría citarse un tercer propósito para el colapsamiento: evitar la aparición de frecuencias esperadas muy pequeñas, problema que pondría en entredicho la pertinencia de un análisis LL (profundizaremos este tema en un próximo trabajo referido a problemas del tamaño de la muestra).

Colapsar categorías consiste en combinarlas, de forma que se reduce el tamaño de la tabla y aumentan las frecuencias por casilla, ya que se suman las frecuencias de las casillas colapsadas. En una tabla 3x2, por ejemplo, se pueden combinar dos de las tres categorías de la primera variable para formar una tabla 3x2. Si la tabla original 3x2 era, por ejemplo:

		X		
		a	b	TOT.
Y	a	12	22	34
	b	5	20	25
	c	3	15	18
	TOT.	20	57	

Combinando las categorías b y c de Y se obtiene la tabla 2x2:

		X		
		a	b	TOT.
Y	a	12	22	34
	b	8	35	43
	TOT.	20	57	

Si el colapsamiento de categorías se realiza ad hoc puede suponer problemas de interpretación. Por ejemplo, las rutinas de análisis estadístico computarizado disponibles para ello en el programa P4F de BMDP colapsan automáticamente categorías adyacentes, vecinas (BROWN, 1983), sean las que sean. Si se trata de categorías de una variable ordinal, o de una variable continua agrupada en intervalos, esto no supone demasiados problemas, porque el orden se mantiene. Pero en variables estrictamente nominales esto puede dar lugar a una categoría sin sentido. Por ejemplo, pensemos en una tabla donde la variable fila fuese «Carrera» con las

categorías «ciencias naturales», «humanísticas» y «técnicas», colocadas en ese orden. La combinación de categorías adyacentes produciría aquí una categoría combinada «humanas-técnicas», o bien «naturales-humanas», de difícil explicación, tanto una como otra, desde un punto de vista sustantivo, teórico. Puede ser aceptable, en cambio, el colapsamiento automático de categorías si las nuevas que se obtienen siguen teniendo sentido, coherencia, aunque siempre cambia en cierta medida la variable subyacente que estamos midiendo, y en la interpretación debemos tener en cuenta ese cambio o replanteamiento. Si esas categorías no van juntas en la tabla, debe recurrirse a otras rutinas, distintas del colapsamiento automático, para generar las combinaciones deseadas (por ejemplo, con el párrafo CATEGORY, de BMDP).

Entre otras condiciones o requisitos que se establecen para el análisis de variables nominales, está la de que sus categorías, además de mutuamente excluyentes, sean exhaustivas: dado un sujeto cualquiera de la muestra, y para toda variable considerada, debe haber una categoría donde pueda ser ubicado; en el peor de los casos, hasta la categoría «inclasificable» es una categoría, aunque también es cierto que deben evitarse las categorías tipo «cajón de sastre», porque otra condición es que las categorías sean homogéneas con respecto a la propiedad subyacente.

Por ejemplo, la variable «estado civil», exhaustiva y un tanto libremente considerada, podría tener las siguientes categorías: casado, soltero, separado, divorciado y viudo. Agrupar categorías adyacentes —suponiendo que su orden en la tabla sea el mismo en que las hemos detallado aquí— produciría algunas combinaciones insostenibles, como casado-soltero. Otra posible combinación adyacente sería separado-divorciado, que sí podría ser aceptable considerando que es problable que los separados estén tramitanto el divorcio. Y en cuanto a la categoría viuda, su combinación con el resto de los no casados sería difícil, en tanto que su estado es (supuestamente) forzoso y no voluntario. En definitiva, si seguimos ideando presuntas combinaciones puede que lleguemos a la conclusión de que la variable subyacente se ha «transformado»: ya no medimos estado civil, sino otra cosa que puede ser, por ejemplo, tipo de convivencia, en la cual las categorías se reorganizarían de distinta forma, incluso podrían aparecer otras inicialmente no consideradas («en pareja», «solo», «en comuna», etc., etc.).

En suma, puede efectuarse, en principio, una reorganización de categorías siempre y cuando las nuevas que se obtengan sigan teniendo coherencia sustantiva. Y ser conscientes de que una reorganización demasiado profunda de las categorías puede llevar a un cambio tan drástico en las variables que el problema inicial de investigación deba, prácticamente, replantearse. La agrupación de categorías en ocasiones puede llevar a simplificar las interpretaciones, pero en otras puede llevar a confusiones, todo depende de qué y cómo se agrupe y hasta qué punto es permisible el agrupamiento que se proponga. En muchos casos, no existe una intención de agrupamiento desde el comienzo de la investigación, sino que surgen a posteriori, porque se han obtenido en ciertas casillas valores tan pequeños que no queda otro recurso, a menos que se vuelva atrás, al proceso de obtención de datos y se añadan casos a la muestra, lo que no es siempre fácil ni justificable desde otros puntos de vista, por ejemplo,

porque el paso del tiempo entre una recogida de datos y otra podría introducir elementos de distorsión importantes. En definitiva, es cada investigador quien debe decidir en qué medida y cómo puede utilizar este recurso en el análisis de sus datos y qué consecuencias tiene para la interpretación de los resultados.

Aún cuando podamos colapsar categorías de una variable sin caer en la irrelevancia o el disparate desde un punto de vista teórico, desde un punto de vista estadístico puede traer problemas, ya que el análisis LL trata las variables en relación unas con otras y el resto puede verse afectado.

Imaginemos una tabla de 3 variables XYZ (5x5x5) en la cual deseamos colapsar las categorías de la variable Z, por ejemplo, de forma que se obtenga una tabla 5x5x2. Pues bien, si los parámetros pertenecientes a X e Y permanecen inalterados después del colapsamiento, Z se considera «colapsable», de lo contrario no lo será.

Llevando el colapsamiento de las categorías de una variable a sus últimos extremos (combinando en una sola todas sus categorías) se llega al total colapsamiento de esa variable. En el ejemplo, nos quedaríamos con una tabla 5x5. De nuevo, si las relaciones entre X e Y siguen igual que antes del colapsamiento, éste podría mantenerse. De lo contrario, se pueden «crear» nuevas relaciones de forma artificial o alterarse las previamente existentes (BISHOP, 1971).

3.1.1. *Teorema para la «colapsabilidad» permisible*

De forma general se ha propuesto un teorema para determinar la colapsabilidad de una variable. Es el siguiente:

> «En una tabla tridimensional, la interacción entre dos variables puede medirse a partir de la tabla obtenida colapsando una tercera variable si la tercera variable es independiente de *al menos una* de esas dos variables» (FIENBERG, 1977:49).

Lo mismo se aplica a tablas mayores. En la práctica, esto se traduce en dividir las variables en tres grupos: el primero contiene las variables a ser colapsadas; el segundo, las que son independientes de las anteriores; y el tercero, las que no son independientes de las colapsadas. Se entiende aquí independencia en el sentido de que los términos μ uniendo dos variables son cero. La regla para la colapsabilidad es que el primer grupo de variables es colapsable con respecto a los términos μ que contengan variables del segundo grupo, pero no con respecto a los términos μ que contengan sólo variables del tercer grupo.

Por ejemplo, consideremos el modelo LL (AB, AC). En este modelo los términos μ^{BC} y $\mu^{ABC} = 0$. Es decir, las variables B y C se consideran independientes. Así, podremos colapsar C sin cambiar μ^{AB}. Los parámetros μ^{A}, sin embargo, serían distintos en la tabla original y en la tabla colapsada, ya que A se relaciona tanto con B como con C.

Supongamos ahora que colapsamos A, ya que A no es independiente ni de B ni

de C (es decir, el segundo grupo de variables está vacío) todos los términos (μ^B, μ^C y μ^{BC}) resultarían afectados.

Hasta ahora nos hemos limitado a exponer las consecuenicas que puede tener la colapsabilidad en el valor de los parámetros de un modelo LL. Pero estos cambios pueden afectar al nivel de significación de un modelo, medido por ji^2 o por L^2. Puede afectar, en definitiva, la aceptación o el rechazo de un modelo en el análisis LL. Así, se llegaría a tomar una decisión de aceptación o de rechazo debido a lo que se suele llamar un «artefacto» del análisis.

3.1.2. *Influencia de colapsamiento en el ajuste de modelos*

Reducir el número de categorías simplifica, qué duda cabe, los análisis. Muchas veces se hace por esa razón, no sólo en las investigaciones, sino en los libros o artículos sobre análisis estadístico. Muchos trabajos de GOODMAN, por ejemplo, sólo incluyen variables dicotómicas, algunas de las cuales eran en principio politómicas. Podría ser justificable en el contexto de un trabajo de metodología de análisis estadístico, dado que esto aclara la descripción de los datos, los procesos de cálculo y la interpretación de resultados. En definitiva, puede comprenderse mejor la técnica. El problema es que los investigadores, de forma bastante imprudente, a partir de estos trabajos, pueden concluir que el número de categorías de una variable es un asunto sin importancia, que pueden manipular los datos de la forma que quieran sin perder ni cambiar la información. Pero el número de categorías afecta las conclusiones que se alcancen. Por ejemplo, KNOKE (1975) utilizó para un análisis datos acerca de 4 variables:

(1) actitud hacia la legalización del aborto —2 categorías—
(2) religión —3 categorías—
(3) nivel educativo —4 categorías— y
(4) frecuencia de asistencia a la iglesia —5 categorías—

Como se ve, hay variables categoriales propiamente dichas y variables continuas categorizadas. Con estas variables se hicieron una serie de análisis encaminados a:

a) ajustar una serie de modelos dejando el número de categorías inicial (variables originales)
b) ajustar los mismos modelos después de dejar sólo dos categorías en todas las variables (variables dicotomizadas).

A continuación se resumen los resultados, representando como N a los ajustes no significativos —el modelo no ajusta bien— y como S a los ajustes significativos —el modelo sí ajusta bien—. Las variables se representan con los números (1) al (4) utilizados arriba.

Ajuste de modelos con categorías
a) colapsadas y b) sin colapsar

MODELO	V. DICOTÓMICAS	V. ORIGINALES
a 12, 13, 14, 23, 24, 34	N	S
b 123, 14, 24, 34	N	S
c 134, 12, 23, 24	N	S
d 124, 13, 23	N	N
e 123, 134	N	N
f 123, 134, 24	S	S
g 123, 134, 124	S	S

En el caso de que se hayan dicotomizado las variables, sólo son aceptables los modelos más complejos: 'f' y 'g'. El modelo 'f' contiene todas las interacciones de dos factores y dos interacciones de tres factores; el modelo 'g' contiene todas las interacciones de dos factores y todas las interacciones de tres factores que implican la variable 1 (opinión hacia la legalización del aborto). Pero cuando no han sido colapsadas las variables, también se obtiene buen ajuste con modelos mucho más simples —es decir, parsimoniosos—, por ejemplo, con el modelo 'a' que sólo tiene interacciones de dos variables. Se entiende aquí que un modelo es más simple que otro cuando contiene interacciones de menor orden. Por ejemplo, de 2 variables, como el modelo 'a', y no de 3, como el modelo 'g'. No tiene nada que ver el número aparente de interacciones: debido al principio jerárquico que subyace a los modelos LL, el modelo 'g' contiene en sí mismo todas las interacciones de orden inferior expresadas en el modelo 'a'.

Esta y otras comparaciones, muestran que, por lo general, las variables no colapsadas (es decir, con una categorización más exhaustiva) llevan a un esquema o imagen mucho más sencilla, parsimoniosa, de la estructura de los datos. El colapsamiento, por el contrario, hace precisa la adopción de modelos complejos, donde existen múltiples interacciones entre las variables. La intención, pues, de colapsar con la intención de facilitar la interpretación de los resultados resulta, en definitiva, contraproducente en cierta medida, pues es obvio que cuanto más complejo sea el modelo, más difícil será su interpretación, pese a que el número de categorías haya disminuido.

Naturalmente, si la base teórica de la que parte el investigador y las hipótesis que (consecuentemente con esa base) formula son de índole complejo, la aparición de interacciones múltiples no debe tomarse como un mal resultado, sino todo lo contrario. Llegar a la verificación de las hipótesis cuando se ha llevado a cabo honestamente el análisis de datos es un indicio de la potencia del marco teórico y de que se va por buen camino, y eso no desagrada a ningún investigador. Pero lo que se entiende por un resultado «parsimonioso» no puede ser establecido de forma uniforme para todas

las investigaciones. Lo que se hace preciso es distinguir entre obtener interacciones complejas de forma natural y obtenerlas de forma artifactual mediante el colapsamiento.

A partir de esto, podría darse, en algunos contextos, una cierta picaresca, que recurre al artefacto mencionado para obtener artificialmente interacciones complejas y no por razones más confesables (como la evitación de ceros muestrales, frecuencias esperadas pequeñas, etc.). Pero es obvio que no nos referimos a ese tipo de investigación ni nos dirigimos a esa clase de investigadores.

3.2. Colapsamiento involuntario o implícito

Existen formas camufladas de colapsamiento, que inicialmente pueden no ser reconocidas como tales porque no responden a una intención manifiesta en ese sentido. Veamos algunas.

3.2.1. *Categorización de datos continuos*

Muchos fenómenos se dan en niveles discretos, categoriales. Pero otras muchas variables de interés para la educación son, indudablemente, continuas. A veces estas variables se tratan como si fueran categoriales. En ocasiones se hace, sencillamente, por simplificar; en otras, porque no se tienen los medios (instrumentos) para medirlas con más precisión. Tratarlas como si fueran discretas es una forma de colapsamiento; a veces incluso se dicotomizan, lo cual representa un colapsamiento extremo. La categorización de variables continuas, en muchos casos, significa no captar bien la «riqueza» de las mismas y eso tiene consecuencias.

REYNOLDS (1977a) cita un trabajo de simulación por ordenador que ilustra el problema. Mediante un programa se generaban diversos modelos de 3 variables representando correlaciones «espúreas»: A 'causaba' tanto a B como a C. Es decir, B y C eran condicionalmente independientes. Los datos inicialmente estaban en una escala de intervalo, por tanto se probaba cada modelo, mediante el criterio de que r_{BC*A} sería siempre o excepto por error de muestreo. Los datos siempre satisfacieron este criterio. Posteriormente, se agruparon las variables en categorías ordenadas; se hicieron distintos tipos de agrupación: desde 2 categorías hasta 10, obteniéndose diversas tablas posibles. Finalmente, se probó en cada una de ellas el modelo LL de independencia condicional. Los resultados indican que el mismo conjunto de datos, categorizados de formas diferentes, llevan a distintos resultados. Cuando todas las variables tenían al menos 5 categorías, se mantenía el modelo de independenica condicional —que era lo que se esperaba dada la estructura subyacente a los datos—. Si algunas variables, especialmente A, tenía menos de 5 categorías, el modelo se desbarataba. Por el contrario, cuando la variable A —que en este caso actúa como variable 'control'— tiene al menos 5 categorías, el modelo ajusta bien, aun cuando las otras dos variables tengan sólo dos categorías.

Aún manteniendo la importancia del teorema que vimos en el apartado anterior, es un alivio pensar que manteniendo 5 o más categorías en las variables apropiadas se pueden evitar inferenicas erróneas cuando se categorizan datos continuos, ya que con mucha frecuencia las investigaciones recaban datos de tipo cuantitativo, junto con los estrictamente nominales, e interesa incluirlos todos en el mismo análisis.

La conclusión más obvia a extraer es la precaución sobre el nivel de medida de las variables. No basta con categorizar variables para aplicarles un análisis LL. La posible continuidad subyacente debe analizarse muy bien para evitar distorsiones. En todo caso, se aconseja medir esas variables con tanta precisión como sea posible. Al fin y al cabo, esta técnica fue diseñada para variables categoriales y no puede pretenderse que funcione igual si los datos no lo son.

3.2.2. *Variables ausentes y otras formas de colapsamiento implícito*

Se refiere a variables que se han dejado fuera de un análisis determinado. Bien porque éste se haga por partes (siendo las variables —algunas o todas— distintas en cada parte) o bien porque no las hayamos medido.

Las variables omitidas son, en último término, variables colapsadas. Naturalmente, decidir no medir una variable o no incluirla en un análisis específico no es un asunto estadístico. Es una decisión que toma el investigador, por lo general basándose en su marco teórico, sus hipótesis, etc. El investigador que trabaja una tabla AxBxC debe preguntarse si realmente no estará trabajando con una tabla colapsada, si no se le habrá escapado alguna variable fundamental. Si esta duda existe, es aconsejable aclararla. La inclusión, en plan ensayo, de una cuarta variable D dispararía el número total de posibles modelos a probar hasta una cifra casi inmanejable. En este punto, las pruebas de asociación parcial y marginal, o la adición vs. eliminación de efectos a partir de un modelo base, son una ayuda inestimable para limitar el número de modelos cuyo ajuste sería conveniente someter a prueba. Para todas esas pruebas previas y para el ajuste de los modelos definitivos existen rutinas computarizadas en el programa P4F de BMDP (BROWN, 1983).

Existirían aun otros casos, en los cuales se trabajaría —aun sin pretenderlo específicamente— con variables colapsadas. Por ejemplo, algunos procedimientos especiales del LL que describe GOODMAN (1973c y 1973d) para llevar a cabo análisis causales sobre los datos, procedimientos análogos a los usuales análisis de modelos recursivos con variables cuantitativas. No vamos a extendernos en ese procedimiento aquí, ya que se inscribiría más propiamente en los llamados modelos 'logit', y no en los logarítmico lineales ('log'). Arriesgándonos a los peligros que conlleva la simplificación, podríamos resumir que la diferencia entre el enfoque «log» y el enfoque «logit» es que en el primero no se consideran variables dependientes e independientes ni se entra en relaciones de causalidad y en el segundo sí se hace esta distinción y se analizan ese tipo de relaciones (BAKER, 1981). Describirlo en más detalle exigiría hablar de otros muchos temas (modelos jerárquicos vs.

no jerárquicos, distribución multinomial vs. distribución producto multinomial, etc.), lo cual excedería el propósito de este trabajo. Simplemente, observar que los procedimientos de análisis causal de GOODMAN mencionados suponen en sí, por sus características de procedimiento, una forma de colapsamiento de variables. Aunque las razones para colapsar sean distintas, se aplican en este caso las mismas consideraciones que hemos expuesto en este trabajo.

En definitiva, y a modo de conclusión-resumen, el investigador debe ser consciente de que los resultados que obtenga a partir de la aplicación del análisis LL a sus datos dependerán hasta cierto punto de la forma en que categorice las variables, ya que el número de categorías afecta tanto al valor de los parámetros como a la calidad del ajuste del modelo. En general, las variables exhaustivamente categorizadas favorecen el ajuste de modelos mucho más parsimoniosos; la reducción del número de categorías, por el contrario, puede forzar la adopción de modelos muy complejos. En cuanto a la modificación del valor de los parámetros, puede evitarse este efecto mediante el teorema de la colapsabilidad permisible, estableciendo claramente tres grupos de variables: las colapsadas, las que son dependientes de las colapsadas y las que no lo son. La combinación de categorías in extremis puede llevar incluso a replanteamientos de tipo sustantivo o teórico, sobre todo si se trata de variables estrictamente nominales, sin un orden subyacente entre sus categorías. Por lo que respecta a la categorización de datos continuos, finalmente, el número de intervalos establecido también puede introducir diferencias en el ajuste, recomendándose en este caso dejar al menos 5 intervalos, sobre todo si la variable continua categorizada ejerce en el modelo un papel de variable control.

REFERENCIAS

BAKER, F. B. (1981): Log-linear, logit-linear models: A didactic. *Journal of Educational Statistics*, 6 (1), 75-102.

BENEDETTI, J. K. & BROWN, M. B. (1978a): Alternate Methods of building log-linear models. *Proceedings of the 9th international biometric conference*, 2, 209-227.

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

BIRCH, M. W. (1963): Maximum likelyhood in three way contingency tables. *J. Royal Statistical Soc.*, 25 B, 220-233.

BISHOP, Y. M. (1971): Effects of collapsing multidimensional contingency tables. *Biometrics*, 24, 545-562.

BISHOP, Y. M.; FIENBERG, S.E. & HOLLAND, P. W. (1975): *Discrete Multivariate Analysis: Theory and Practice*. MIT press, Cambridge, Mass.

BROWN, M. B. (1976): Screening effects in multidimensional contingency tables. *Journal of Applied Statistics*, 25, 37-46.

— (1983): P4F. Two-way and Multi-way Frequency Tables-Measures of Association and the Log-Linear Model (Complete and Incomplete Tables). En W. J. DIXON (Ed.): *BMDP Statistical Software*. Univ. California Press, 143-206.

CORREA, A. D. (1989): *Análisis y aplicación del modelo Log-linear a la investigación educativa.* Tesis Doctoral. Depto. Didáctica e Investigación Educativa. Universidad de La Laguna. Canarias.

— (1991): Estudios multivariados con datos nominales: aportaciones del análisis logarítmico-lineal. *Qurriculum*, n.º 3, 35-52.

FIEMBERG, S. E. (1977): *The analysis of cross-classified categorical data.* MIT Press, Cambridge, Mass.

GOODMAN, L. A. (1968): The analysis of cross-classified data. *J. Amer. Statist. Assoc.*, 63, 1.091-1.131.

— (1970): The multivariate analysis of qualitative data: interactions among multiple classifications, *Journal American Statistica*, 65, 226-256.

— (1971): The analysis of multidimensional contingency tables: stepwise procedures and direct estimations methods for building models for multiple classifications, *Technometrics*, 13(1), 33-61.

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

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

— (1973a): The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, 60, 178-192.

— (1973b): Causal analysis of data panel studies and other kinds of surveys. *American Journal of Sociology*, 78, 1.135-1.191.

— (1984): *The analysis of cross-classiffied data having ordered categories.* Harvard University Press.

HABER, M. y BROWN, M. B. (1986): Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *Journal of the American Statistical Association*, 81(394), 477-482.

HABERMAN, S. J. (1974): *The analysis of frequency data.* University of Chicago Press.

— (1978): *Analysis of qualitative data* (Vol. I). New York, Academic Press. Vol. II, 1979.

HALPERIN, S.; NEHRKE, M.; HULICKA, I. & MORGANTI, J. (1976): Partitioning chi-square: the analysis of contingency tables with repeated measurements. *Experimental Aging Research*, 2(2), 105-118.

JORNET, J. M. y SUÁREZ, J. M. (1988): Aplicación de los modelos log-lineales para el análisis de elementos en pruebas de referencia criterial (TRC). En I. DENDALUCE (Coord.): *Aspectos metodológicos de la investigación educativa* (189-193). Madrid, Narcea.

KNOKE, D. (1975): A comparison of log-linear and regression models for systems of dichotomous variables. *Sociological Methods and Research*, 3, 416-434.

KNOKE, D. & BURKE, P. J. (1982): *Log-Linear Models.* Sage Pub., California.

KOTZE, T. J. (1982): The log-linear model and its applications to multi-way contingency tables. En D. M. HAWKINS (Ed.): *Topics in applied multivariate analysis* (142-182). Cambridge Univ. Press.

PAYNE, C. (1977): The log-linear model for contingency tables. En O'MUIRCHEARTAIGH & C. PAYNE (Eds.): *The Analysis of Survey Data.* Vol. II, 105-145. Wiley, London.

REYNOLDS, H. T. (1977a). Some comments on the causal analysis of surveys with log-linear models. *American Journal of Sociology*, 83(1), 127-143.

— (1977b): *Analysis of nominal data.* Sage Pub., California.

SÁNCHEZ CARRIÓN, J. J. (1984): Análisis de tablas de contingencia: modelos lineales logarítmicos. En J. SÁNCHEZ CARRIÓN (Ed.): *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales* (267-294). Madrid, C.I.S.

SANS I MARTÍN, A. (1987): Análisis de tablas de contingencia multidimensionales. *Revista de Investigación Educativa*, Vol. 5(10), 141-147.

SHAFFER, J. P. (1973a): Defining and testing hypotheses in multidimensional contingency tables. *Psychological Bulletin*, 79, 127-141.

— (1973b): Testing especific hypotheses in contingency tables: Chi-square partitioning and other methods. *Psychological Reports*, 33(2), 343-348.

TEJEDOR TEJEDOR, F. J. (1985): Análisis de tablas de contingencia multidimensionales. En: A. DE LA ORDEN (Coord.): Investigación Educativa, 32-36. Col. *Diccionario de Ciencias de la Educación*. Madrid, Anaya.

TEJEDOR, F. J. y GODAS, A. (1988): *Relación entre las variables motivación, expectativas académicas y sexo en base a la aplicación de los modelos log-linear*. Actas del Congreso Nacional de Psicología Social. Alicante, PPU.

TEJEDOR, F. J. y CARIDE, J. A. (1988): Influencia de las variables contextuales en el rendimiento académico. *Revista de Educación*, 287, 113-146.

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

— (1981): Log-linear models, screening and regional industrial surveys. *Reg. Studies*, 15, 33-45.