

**UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS**

FACULTAD DE CIENCIAS MATEMÁTICAS

E.A.P. DE ESTADÍSTICA

**La complementariedad del modelo Log-lineal y del análisis de correspondencia en el estudio de los factores médicos, sociales y de sexualidad en adolescentes con presencia de infección vaginal**

**Capítulo 4. Modelo log-lineal multidimensional**

**MONOGRAFÍA**

Para optar el Título de Licenciado en Estadística

AUTOR

**Raymundo Ismael Pillaca Ortega**

LIMA – PERÚ 2003

CAPITULO IV

MODELO LOG- LINEAL MULTIDIMENSIONAL

* 1. INTRODUCCIÓN

En la mayoría de los análisis estadísticos se hace una distinción entre dos clases de variables: variables de respuesta (dependientes) y variables de explicación (independientes). Sin embargo cuando se realiza un análisis log- lineal no se hace esta distinción, todas las variables bajo estudio son consideradas como independientes y deben ser categóricas, en tanto que la variable dependiente es considerada como él número de frecuencias en las celdas correspondientes al cruce entre categorias. Para ello él número de frecuencias de cada slda se expresa en valores de logaritmo natural de las mismas.

De manera resumida, se toma el logaritmo natural de las frecuencias de ocurrencias en cada celda debida a los efectos de varias variables categóricas como una suma de parámetros (modelo lineal). Estos parámetros nos dan una medida de la dirección y magnitud de las relaciones en cada nivel de las variabI‹sé categóricas. Para saber si existe o no relación entre las variables, se establece una estructura mediante la supresión de efectos o parámetros del modelo. Se verifica la validez estadística de la hipótesis planteada por el modelo. Para encontrar la relación entre dos variables se utiliza una medida de asociación que se presenta para cada tabla bivariada. En un análisis multivariado se estudian las relaciones existentes entre cada par de variables cuando se toma en cuenta que el resto de las variables está presente.

* 1. OBJETIVO DEL MODELO LOG-LINEAL

El análisis log-lineal es un método cuyo objetivo fundamental consiste en estudiar las relaciones entre las variables categóricas y así mismo nos orienta a encontrar el modelo más adecuado.

* 1. EL MODELO LOG-LINEAL

Para introducirnos en el análisis teórico de los modelos log-lineales empezaremos, por simplicidad, con los que tienen como fin explicar la estructura de las tablas de contingencias de dos dimensiones (r x c).

El fundamento de los modelos log-lineales radica en el principio de independencia de dos variables aleatorias. Si dos variables aleatorias son independientes entonces la marginal conjunta es igual al producto de sus respectivas marginales, se verifica que la probabilidad conjunta *pag* se expresa,

para todo *i* y para todo y , como

(1)

es decir, la probabilidad de que un elemento o individuo, presente simultáneamente las dos características (factores) en los niveles i-ésimo y j- ésimo, es igual al producto de la probabilidad de qué presente la característica A en el nivel i-ésimo y la B en el j-ésimo; en otras palabras, la probabilidad conjunta *pá* es igual al producto de las probabilidades marginales *(p.* y *p. ).*

Si por el contrario, consideramos la no existencia de independencia podemos

expresar la probabilidad conjunta *pig* de la siguiente manera:

*pj —— p‹.p.php* Siendo k »0 (2)

siendo *ka* la cuantificación del efecto conjunto del i-ésimo del factor A y del j- ésimo del factor B, efecto conjunto al que llamaremos con nombre de interacción de ambas variables o factores.

Postular o no rechazar la hipótesis de independencia de los dos factores, supone atribuir al elemento t el valor de la unidad, para todo *i* y para todo y .

Si tomamos logaritmos neperianos en (1) y (2), se llega a que, en el caso de independencia.

ln *pi1 —— M pr +* mp.; (3)

y cuando la independencia no existe

lnp; —— lnp . + lnp.; + ln li; (4)

Estas dos expresiones constituyen una primera formulación de los modelos log- lineales, para el caso de dos factores, es decir cuando la tabla de contingencia es de dos dimensiones.

Aunque los modelos (1) y (2) son respectivamente equivalentes, a (3) y (4), parece más razonable utilizar estas últimas formulaciones, modelos aditivos, en donde el término que descrimina la existencia o no de independencia es ahora int , que toma el valor cero cuando estamos en presencia de independencia, valor más acorde con la idea de ausencia de asociación que se pretende representar, erreez del valor unidad que se le asignaba al parámetro £ en el modelo multiplicativo (2), para este mismo caso. Goodman (1970,1971) y Haberman (1974) introducen el concepto de descomposición del modelo log- lineal.

4.3.1 EFECTOS PRINCIPALES

En vez de centrar nuestra atención en las probabilidades de cada celda, como es el caso de los modelos (3) y (4), utilizaremos, como base de las estimaciones, las frecuencias observadas, dado que ésta es la información del que disponemos. Si recordamos que las frecuencias esperadas (£.›) se

calculan, en una tabla con un número total de elementos igual a N, como el producto de la probabilidad de cada celda por este número total. *Er —— Np« , y* que en el caso de independencia o ausencia de asociación resulta

y dado que como sabemos,

*Ej. —— Npi ,‘ E.; —— Np.;*



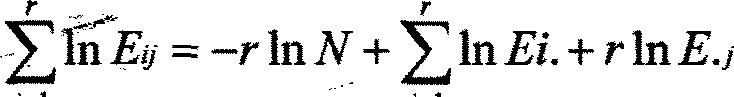
el modelo (3) se transforma en

1ndr = —In *N +* ln ñí. + lm., (5)

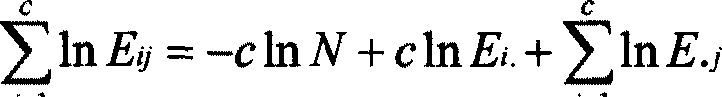
Este modelo tiene semejanza formal con el del análisis de la varianza, por lo cual es aconsejable la utilización de la terminología de esta importante técnica de análisis estadístico.

A continuación efectuaremos en (5) una serie de transformaciones a fin de hacer operativo el modelo. El factor *A* se presenta con *r* niveles: *i* = *1....r,* y el factor *B* con c niveles: j = *1....c.*

Sumando los miembros de la ecuación (5) con respecto a i (factor A), tenemos

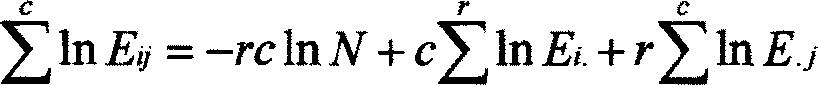
 (6)

A continuación si calculamos la suma de (5) con respecto aj (factor B)

 (7)

Por último la suma en (5) se realiza respecto a *i* y j (los dos factores), resultando

(8)

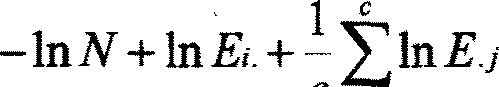


Dividiendo la expresión (6) por r, la (7) por c, y la (8) por el producto re tenemos:

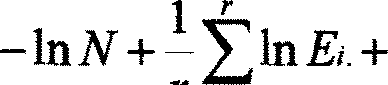
*r* í=1

*=* -InN *+ ¡Z* ln Ai. + lnE J

(10)



(11)



*r*

A partir de las expresiones anteriores definimos los parámetros ,u,›t(,).‹2(,›

(12)



*r c*

No es otra cosa que la media general de los logaritmos de las rxc frecuencias estimadas (recordemos que en una tabla *(AxB)* hay rxc celdas).

 *r c*

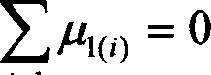


#### Zln £« — *re*



(13)

Es igual a la diferencia de dos términos de los cuales el primero es la media de la distribución marginal, respecto al factor B (columnas), de los logaritmos de las frécuencias esperadas; el segundo término es la media general obtenida en la expresión (12). Esta diferencia nos dice que los efectos producidos por los distintos niveles del factor A se miden a través de las desviaciones de las medias marginales de cada nivel respecto a la media general: se ha eliminado el efecto general que existía si todas las celdas tuvieran el mismo número de elementos.

Si sumamos los r términos „) Ilegamos, fácilmente, a que la suma es igual a cero.

esta conclusión, que coincide plenamente con lo que sucede en el análisis de la varianza, tiene una consecuencia importante como es que los efectos producidoú por los niveles de una característica, tal y como han sido definidos, no son independientes entre sí, pues su suma es igual a cero, lo que implica la relatividad del concepto efecto, no siendo posible separarlo del contexto de la estructura concreta de la tabla que se analiza, puesto que cualquier modificación de dicha estructura conduce a una modificación, cuantitativa y quizás de signo, de las estimaciones de los efectos.

*) r c*

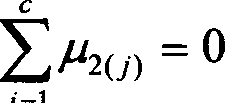


*re*

j£f 2(j)

## —ZZln €3

(14)

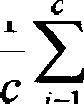
La expresión (14) es exactamente igual que lo realizado con referencia a la expresión (13); sin más que efectuar las transposiciones pertinentes. En particular, los efectos del factor *B* deben cumplir, análogamente a los de *A,* la condición

Entonces, restando a la expresión (11), la suma de (9) y (10) se obtiene:

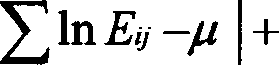
 ln *Ej; —y ——*



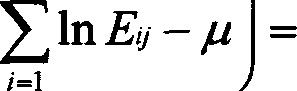
*r*



i



1



*r*

*r*

(15)

resultado que llevado a (5) proporciona la expresión: ”

(16)

Mediante el cual, el logaritmo neperiano de la frecuencia esperada puede descomponerse en la suma de tres parámetros, lo que pone de manifiesto la similitud con el modelo del análisis de varianza a que antes aludíamos, ya que, como vamos a ver a continuación, cada uno de estos tres parámetros va a tener un significado similar al que poseen los del referido modelo del análisis de la varianza.

El sumando ,u cuantifica el valor que adoptarían los logaritmos de las frecuencias esperadas si dos factores no ejercieran ningún efecto. El sumando (,) mide el efecto que produce el nivel de la variable A, factor A, o más

claramente la influencia que la fila í-ésima ejerce sobre el logaritmo del número de elementos o individuos que posean ese nivel de A; de manera análoga el

sumando ‹‹,) evalúa el efecto que el nivel j-ésimo del factor B, efecto que la

columna j-ésima, ejerce sobre la aparición de elementos en ese nivel.

Los parámetros (, y ‹2(,› reciben el nombre de efectos principales o directos. Los valores positivos de los efectos directos (principales) indican que el nivel en cuestión actúa favoreciendo la presencia de individuos en esa fila o columna; dicho con otras palabras, la probabilidad de ocurrencia de individuos en esta fila tiende a ser “alta”.

Los valores negativos indican la situación contraria, esto es, el nivel no favorece, penaliza, la presencia de individuos en esa situación.



Hasta ahora, para exponer las similitudes que existen entre los modelos lineales del modelo del análisis de la varianza, hemós estudiado el modelo log- lineal más simple, suponiendo la existencia de independencia, según el modelo (3), pue implica la condición *pag —— p .p. .* Es inmediata la extensión de este

modelo simple al más complicado por admitir la falta de independencia, falta que se materializa en la presencia del término de las interacciones, como expresa el modelo (4), que implica la condición (2), cuyos cálculos omitimos por considerarlos no necesarios en toda la exposición.

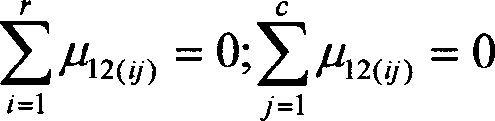
El modelo log-lineal sin independencia, con presencia de interacciones por lo tanto, será

ln ñi, = + und -i- 2(J) -I- ,fJl2(íJ) (17)

los términos primero, segundo y tercero, del segundo miembro, tienen las mismas interpretaciones que en (16). El tercer sumando ‹.2(,› expresa la cuantificación de la acción conjunta del nivel i-ésimo del factor *A, o* efecto de la fila *i,* y del nivel j-ésimo del factor *B, o* efecto de la columna j (en términos del análisis de la varianza la interacción de las dos variables o factores) de tal

forma que si tiene Iugar la interacción su existencia influye en las diferencias de las frecuencias entre celdas, supuesto eliminados los efectos de filas y

columnas y,(,) *y* i•2(,›  y media general (y).

De la misma manera que en los efectos principales existen unas condiciones que deben cumplir los efectos de los niveles de cada factor, en el caso de la interacción aparecen dos condiciones (restricciones) que verifican los efectos de las interacciones (por la forma en que han sido definidos) y son las siguientes:

cada una de las dos sumas marginales (respecto a las filas y a las columnas) son iguales a cero, toque supone la no independencia de los efectos de las interacciones, y su consiguiente relatividad, en línea total con lo que sucedía con los efectos principales.

El modelo especificado para dos factores es fácilmente generalizable para cualquier número de ellos. A continuación vamos a hacerlo para tres y cuatro, lo que además nos dará pie para acercarnos a la problemática de los modelos log-lineales.

Tres factores o características, cuando los elementos o individuos que los poseen se clasifican con respecto a ellos, dan lugar a una tabla de contingencia de tres dimensiones. De forma análoga el caso de dos factores hemos de tener en cuenta una serie de efectos que contribuyen a la tendencia de un elemento a pertenecer a una u otra celda de la tabla. Estos efectos son: tres efectos directos o principales, uno por cada factor ( ,y,yy,); a continuación tenemos

los efectos conjuntos de cada par de características, denominadas interacciones de orden uno(›t .x.›‹23): por último el efecto conjunto de las tres caracteristicas, interacciones de orden dos 6• .›). Expresado todo esto de forma similar al modelo (17) Ileva a:

ln *Eitt —— y + yi( › + 2(y) +* 3(1) -i- 12(¡/) -Í- 13(ii) + 23(já) + 123( l) (18)

para hacernos idea de la dificultad de comprensión directa del concepto de interacción de orden elevado, la de orden dos (›t2.) expresa la influencia que, por ejemplo el factor 3 ejerce sobre la interacción de orden uno de los otros dos factores (•.2) ; sin ninguna dificultad podemos cambiar unos factores por otros para explicar el complejo sentido de la interacción de orden dos, «.23 -

En el caso de cuatro factores el modelo, que incluye la interacción de orden tres ¢•l234) asi como las inferiores, tiene por expresión:

ln j£f -Í- 1(í) + jfJ2(J) + 3(k) *+* **jfZ4(/) -1- 12(y) -Í- )f13(iá) •Í•** *14(el) +* jfJ23(j£) -1- 24(J/) +



34(kf) + *fl* 23(ijk) *+* 124(inf) + 234(já/) + 1234(ydf)

4.3.3 CEROS ESTRUCTURALES Y TABLAS INCOMPLETAS

( 19)

El que una o varias celdas de una tabla de contingencia aparezcan vacías, es decir, el que sus frecuencias observadas sean cero, prtede deberse a una de dos causas completamente distintas y con implicaciones, también, completamente diferentes.

Vemos que la falta de elementos, celdas vacías también llamadas ceros, tienen un origen distinto:

Ceros muestrales: si es posible su aparición por incremento de la muestra, y Ceros estructurales: si es imposible. La presencia o ausencia de ceros estructurales determina una condición de la tabla de contingencia: si no contiene ceros estructurales la tabla se denomina completa y si los contiene incompleta.

Una de las principales implicaciones que presenta una y otra situación es que en las tablas completas (sin ceros estructurales, pero pudiendo tener ceros muéstrales) ningún valor esperado deberá ser cero, mientras que en las incompletas (conteniendo ceros estructurales) allí donde figuren los ceros estructurales los valores esperados necesariamente tienen que ser ceros.

Es preciso salir al paso de una práctica que pudiendo ser lícita en el caso de ceros muéstrales no lo es con los estructurales. Esta practica no es otra que la fusión de filas o columnas. Esto es licito cuando los ceros son muéstrales y no pueden obtenerse una muestra de mayor tamaño, pues con la desaparición de las celdas vacías no se modifica la información que se tiene de la población (excepción hecha de la eliminación de uno o varios niveles de uno y otro factor). Por el contrario una práctica semejante con ceros estructurales implica trabajar sobre una tabla de contingencia que en modo alguno representa la población en cuestión que si contiene, por su propia estructura, celdas vacía, es decir combinaciones de los niveles que no aparecen en la población

* 1. CLASIFICACIÓN DE MODELOS

Hemos visto la expresión general de los modelos log-lineales en los casos de tablas de contingencia de dimensión 2, 3 ó 4 y los problemas que plantea. Antes de comenzar con el análisis de estos modelos, debemos introducir una serie de definiciones que nos permitan distinguir los modelos generales de aquellos que presentan ciertas particularidades, que deben ser tratados de manera especial.

* + 1. MODELO SATURADO

Definimos un modelo saturado si contiene en su formulación tantos parámetros independientes como celdas tiene la tabla a la que es aplicado; dicho de otra forma más operativa, modelo es saturado si incluye todos los posibles efectos principales e interacciones. En caso contrario, si falta algún o algunos parámetros, se le denomina no saturado. Todos los modelos con interacciones contemplados hasta ahora son saturados, y él (16) no saturado.

Tomamos como ejemplo el modelo (17)

ln *Eii* = + yi(i -i- y2(j) -1- j£f12(y)

los subíndices tienen como campo de variación *i* = *1...r;j -— 1...c .*

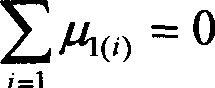
El número de parámetros independientes contenidas en él es el siguiente:

36

* Uno, correspondiente a la media general .
* *r-1* parámetros que recogen los efectos principales ,u, , ), ya que pese a que

son r los niveles de ese factor, uno queda determinado si tenemos en cuenta la restricción impuesta

*r*



* Análogamente, *c-1* parámetros que recogen los efectos principales de «2s,›
* Debido también a las restricciones que ligan a las interacciones, *existirán (r-*

*1)(c-1)* parámetros que recogen los efectos conjuntos.

En total el número de parámetros de este modelo será: Número de parámetros = *1 + (r-1) + (c-1) + (r-1)(c-1) -- rc*

Este número resulta igual al de celdas de la tabla *(AxB).* De manera similar podríamos determinar los parámetros que figuran en losa modelos de 3, 4 o más dimensiones.

Por otra parte, el modelo saturado reproduce exactamente las frecuencias observadas.

* + 1. MODELO JERÁRQUICO

El concepto de modelo jerárquico se refiere a la coherencia interna de los modelos no saturados: si en un modelo falta un término (se supone que se ha establecido sobre ese término la hipótesis de que es igual a cero) para considerarle como jerárquico, deberán estar excluidos todos los parámetros de orden superior que contengan la combinación de subíndices fijos del que no aparece. Estos se harán con más facilidad con un ejemplo: si en el modelo de cuatro factores (19) se suprime el término ‹23(,.) , para que el modelo resultante

sea jerárquico, deberán estar ausentes los términos que contengan conjuntamente la combinación de subindices fijos 2 y 3.

*\23( ijk),* jÉ/234( jM), ,É/l234(iJá/)

con lo cual resulta el modelo log-lineal

lu Pitt = z -1- fbi(í) -I- £f2(j) -1- 3(+) -t- ,ff4ç/) -1- /zi2(iJ ) -Í- il3(i+) + j£f14(d) -I- j£ 24(J/) + j£f34(k/) + j£f124(y/) + jf 134(;£/)

El criterio de jerarquía puede contemplarse en forma inversa ya que, si en el modelo aparece un parámetro con una cierta combinación de subíndices fijos, deberán figurar todos los parámetros que contengan esos subíndices en combinaciones inferiores; es decir, si aparece la combinación 124 es preciso que también figuren las combinaciones fijas 1,2,4, (1,2), (1,4) y (2,4).



2( ) • 14(j/)• Ü24(jf)

24(y/)

Es preciso tener en cuenta que todo modelo saturado es necesariamente jerárquico, lo recíproco, que un modelo sea jerárquico no supone que sea saturado, dado que manteniendo la propiedad de jerarquía establecida pueden estar ausentes los parámetros correspondientes a los efectos de orden superior.

En la mayor parte de los casos los modelos jerárquicos, frente a los no jerárquicos, bastan para el estudio de las situaciones que puedan presentarse.

* 1. INDEPENDENCIA Y COLAPSABILIDAD
     1. INDEPENDENCIA Y ASOCIACIÓN

A fin de exponer, de la manera más asequible, los conceptos de este apartado recurriremos al modelo tridimensional *(AxBxC),* cuyo modelo saturado, y por tanto jerárquico, viene dado por la expresión:



Especificamos las clases de independencia y asociación que pueden presentarse, en términos de los parámetros de este modelo.

1. INDEPENDENCIA TOTAL, COMPLETA O GLOBAL

Los tres factores son total o completamente independientes si lo son tanto simultáneamente como dos a dos, lo que conduce a que, para todo i, j y k, todas las interacciones sean nulas, es decir;

j£f12 = 0 ; ,£r3 = 0 ; 3 = 0 ; //t23 = 0

1. INDEPENDENCIA PARCIAL:

Un factor completamente independiente de los demás. Recurriendo al modelo tridimensional diremos que el factor 1 es completamente independiente de los otros dos factores (2,3) cuando las interacciones en las que aparece ese factor son nulas:

  //iz3 = 0

Esta definición permite que pueda existir algún tipo de relación entre los factores 2 y 3 y, por tanto, si esto sucede deberá resultar 3 x 0.

1. INDEPENDENCIA CONDICIONAL

Siguiendo con el ejemplo del modelo de tres factores, podemos decir que los factores 1 y 2 presentan independencia condicional cuando son independientes entre sí, para cada nivel del tercer factor, pudiendo estar ambos factores asociados a su vez con este tercer factor, teniendo entonces que

j£ft2 = 0 ; j£ft23 = 0

Las otros dos interacciones



39

1. ASOCIACIÓN PARCIAL

El modelo presenta asociación parcial entre los tres factores cuando las interacciones de segundo orden son nulas, ,ut23 — 0, siendo diferentes de cero todas las interacciones de primer orden

pu z 0 ; 13 4 0 ; yi3 z 0

En general, la mayoría de los datos que proceden de clasificaciones cruzadas 2, 3 o más factores, pueden describirse a partir de modelos jerárquicos. No obstante, existen excepciones en las que es necesario utilizar estructuras no jerárquicas, donde puede ser necesario incluir, por ejemplo, el parámetro

,ut23 x 0, siendo las interacciones de orden inferior nulas

,«» = » =,«z› = 0

En estos casos la” interpretación de las interacciones es más compleja y no pueden encuadrarse en los tipos de independencia y asociación antes presentados: el efecto que recogerian estos modelos log-lineales no jerárquicos estaría más relacionado con la sinergia entre factores que con la idea más común de dependencia o asociación.

* + 1. COLAPSABILIDAD

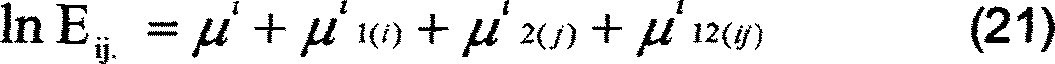
Uno de los problemas que aparecen con frecuencia en la utilización de los modelos log-lineales hace referencia a la posibilidad de combinar categorías de uno o más factores. Tal forma de actuar conduce a la reducción de la dimensión de la tabla de contingencia resultante y a hacerla más manejable, sobre todo cuando la tabla es de orden o dimensión elevado; es decir, a la obtención por ejemplo de una tabla *(AxB) a* partir de una tridimensional *(AxBxC),* sumando para todos los niveles del tercer factor las correspondientes frecuencias de celdas, o dicho de otra forma, llegar a la distribución marginal de los dos primeros factores.

Esta práctica no siempre es lícita ya que es imprescindible, a la hora de no tener en cuenta un factor, que la estructura del modelo no varíe con su eliminación. El concepto de colapsabilidad hace referencia a esta posibilidad.

Se dice que una tabla de contingencia es colapsable cuando se pueden combinar todos o parte de los niveles de un factor, mediante suma de sus correspondientes frecuencias, sin que cambie la estructura significativa del modelo de la tabla inicial.

Partiendo del modelo saturado con dimensión *(AxBxC)*

ln Eíjk + (f) 2(y) + 3(ñ) 2(y) + 3(ti) + 25(yá) + 123(9i) (20)

Si la tabla es colapeable respecto, por ejemplo al tercer factor, sumando dentro de los niveles de este obtendríamos el siguiente modelo

y para que la estructura o comportamiento que describe el nuevo modelo (21) entre los factores 1 y 2 sea el mismo (no varíe) que en el modelo saturado (20),

es ne ario que los parámetros u l2( y /z„ , , de uno y otro modelo, sean

iguales.

Si, por el contrario, al condensar la tabla, todos los términos correspondientes a

las interacciones de primer orden («)2 \* «.3 \*«23) siguen estando presentes, se modifica la estructura del modelo debido a que los parámetros varían.

Las condiciones de colapsabilidad de una tabla de contingencia son las siguientes:

1. Si en una tabla (AxBxC), por ejemplo, el factor 1 es condicionalmente independiente de uno de los otros dos {2 ó 3), la tabla es colapsable para cualquiera de los niveles del factor 2 o del factor 3.
2. Si un factor es completamente independiente respecto a los demás, es decir, cuando existe independencia parcial, se puede condensar la tabla de contingencia sumando respecto a sus niveles sin cambiar los parámetros

, y por tanto la estructura del modelo.

1. Por último, si todos los factores son completamente independientes, se puede reducir la dimensión de la tabla mediante el colapso de cualquiera de los factores.

En todos los supuestos anteriores es preciso tener en cuenta la condición inicial de que el modelo sea jerárquico y, además, siempre se tendrá que, para el ejemplo de la tabla *(AxBxC),* «23 '0

Por otro Iado, domprobamos como la independencia de la tabla expandida, referente al modelo saturado, se traslada a la tabla condensada, de manera que la independencia no se pierde por colapso de los niveles de uno o más factores. Pero, desgraciadamente el supuesto ínvefso no se cumple y, por tanto, la práctica de estudiar todas las tablas marginales (AxB) es engañosa si algunos de los factores se hallan interrelacionados.

Qüeremos reiterar lo dicho sobre el colapso por fusión de filas o columnas de una tabla a fin de eliminar ceros estructurales, es una práctica no lícita, debido a que la presencia de tales ceros está proporcionando Mna información necesaria e imprescindible para describir el modelo que corresponda a la tabla incompleta, y que al colapsar las filas y/o columnas que permitan la desaparición de los ceros estructurales, se está distorsionando la idea de la población de la cual procede la muestra, población que sí contiene, por su propia esencia, esas celdas vacías. En el caso de ceros muéstrales este problema no es relevante y para el colapso pueden seguirse las normas generales que hemos establecido anteriormente.

* + 1. SISTEMA DE NOTACIÓN ABREVIADA EN LOS MODELOS LOG- LINEALES

Según va aumentando la dimensión de los modelos log-lineales crece más rápidamente él número total de términos que aparecen, haciendo cada vez más tediosa su formulación expandida. Por esta razón, se ha venido aceptando de manera mas generalizada en la literatura especializada una notación abreviada que resume y explícita el número de términos relativos a efectos principales e interacciones que se contengan en cada modelo concreto.

Esta notación abreviada tiene su justificación en la idea de modelo jerárquico reflejado entre corchetes el subíndice o subíndices de los términos de mayor orden que impliquen la existencia de términos de orden inferior.

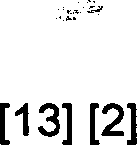
Por ejemplo, el modelo [123] es un modelo jerárquico saturado que contiene el término de la interacción de segundo orden «›2. y todos.los de orden inferior en donde puedan combinarse estos factores, además de los efectos principales, es de r, el modelo [123] sería:



El modelo [12] [13] [23] sería:



y así sucesivamente. En la tabla N° 2 se recogen las equivalencias de los distintos modelos tridimensionales respecto a su notación abreviada, y su adecuada clasificación en función de los términos que presente:

Tabla N° 2: Notación abreviada de los Modelos Log-lineales



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| NOTACION ABREVIADA | | MODELO EXPANDIDO | | | DESCRIPCIÓN | CLASIFICACIÓN |
| [123] | | 1 2 3  12 13 23 123 | | | Los tres factores están  asociados | Asociación |
| [12] [13] [23] | | ln *E —— u + u + u*  *ijk 1* 2  12 13 23 | *+ u +*  3 | | No existe asociación  simultánea entre los tres factores | Asociación parcial |
| [12] [13] | | In *E* = u + ii + ii + u + | | | Los factores 2 y 3 son | Independencia |
|  | | *ijk* 1 2 3 | | | independientes entre sí, | condicional |
|  | | *u + u*  12 13 | | | para cada nivel del factor 1. |  |
| [12] {23] | | 1» | | | Los factores 1 y 3 son |  |
|  | | If -1- If | | | independientes entre sí, |  |
|  | | 12 23  ln *E* = u + u + ii + u + | | | para cada nivel del factor 2. |  |
| [13] [23] | | *ijk* 1 2 3 | | | Los factores 1. y 2 son |  |
|  | | 13 23 | | | independientes entre sí, |  |
|  | |  | | | para cada nivel del factor 3. |  |
|  |  |  |  |  | El factor 3 es totalmente | Independencia |
|  |  |  |  |  | independiente de los | parcial |
|  |  | 12 |  |  | factores 1 y 2. |  |
|  |  | ln £ = u + u + u | + u | + | El factor 2 es totalmente |  |
|  |  | yk 1 | 2 3 |  | independiente de los |  |
|  |  | 13 |  |  | factores 1 y 3. |  |
| [23] [1] | - | In *E* = u + ii + u 1 1 | + ii  2 3 | + | El factor 1 es totalmente |  |
|  |  |  |  |  | independiente de los |  |
|  |  | 23 |  |  | factores 2 y 3. |  |
| [1] [2] [3] | | In *E —- u + u + u + u*  £1 2 3 | | | No existe interacciones | Independencia  total |

* 1. ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO Y BONDAD DEL AJUSTE

Una vez establecido el modelo en principio más adecuado al problema, cuyo reflejo cuantitativo muestral es una determinada tabla de contingencia, se procede a la estimación de los parámetros y a la evaluación del grado de ajuste de los datos observados a dicho modelo, fases que resultan imprescindibles para explicar mediante un modelo la estructura tabular de las variables bajo estudio.

Para este análisis las etapas que hay que seguir son:

1. Estimación de las frecuencias esperadas.
2. Estimación datos parámetros especificados en el modelo log-lineal y,
3. Valoración de en que medida es capaz de reproducir de forma idónea la estructura de la tabla.
   * 1. ESTIMACIÓN MÁXIMO VEROSÍMIL DE LOS PARÁMETROS
        1. TABLAS DE CONTINGENCIA BIDIMENSIONAL

Se detalla el método directo y exacto en tablas de contingencia *(AxB)* y sin ceros estructurales. (ver Ruiz Maya, F.J. Martín Pliego,1998)

Llamamos n„ a la frecuencia observada correspondiente a la combinación del nivel i-ésimo del factor *A* y del nivel j-ésimo del factor *B: n,* y » son las

frecuencias marginales de cada uno de los dos factores, siendo *N* el número

total de elementos que han sido clasificados en la tabla de contingencia; su calculo, recordamos, es el siguiente

Partimos del modelo saturado de la tabla *(AxB)*

ln *Ej; ——* + yi(i) -i- y2(j) -b 12(y)

Y establecemos la hipótesis de independencia, es decir que para todo *i* y j las

interacciones son nulas, ,ut2(») = 0.

Llamamos E, a la estimación de la frecuencia esperada E, . Las estimaciones se realizan por el método de la máxima verosimilitud y la solución general es;

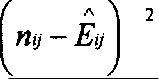
 (1)

*N*

Una vez conseguidas las estimaciones de las frecuencias esperadas estamos en condiciones de contrastar la hipótesis de independencia.

El principal problema reside en la necesidad de disponer de un estadístico que cuantifique en que medida un modelo concreto se ajusta adecuadamente a la estructura de una tabla de contingencia concreta, o lo que es análogo, cuanto explica el modelo la estructura de la tabla. Las dos medidas más usuales de la Prueba Chi-cuadrado son: el estadístico y'de Pearson y la razón de

verosimilitud (G2) La primera de las dos tiene como expresión, como se ha

indicado anteriormente,

(2)

siendo la expresión de la segunda medida

G2 = 2Z

"

(3)

Estos dos estadísticos 2 y ', para la situación que la hipótesis cierta sea la

de independencia, se distribuyen como una -cuadrado con *(r-1)(c-1)* grados de libertad.

Si él número total de elementos de la tabla de contingencia, N, es elevado los dos test son equivalentes. La ventaja del primero sobre el segundo radica en el

hecho de la posible partición de 2 en tantos sumandos como grados de

libertad tenga.

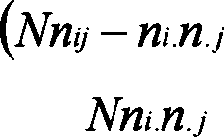
La generalización de las expresiones (2) y (3) a una tabla multidimensional es inmediata.

En la dos formulas n, representa las frecuencias observadas, las que

componen la tabla de contingencia, y E, las frecuencias esperadas en cada celda bajo las diferentes hipótesis que se establezcan en cada modelo log- lineal. En la practica estas frecuencias esperadas se sustituyen por sus estimaciones máxio›o verosímiles, y es en este proceso de estimación donde radica la menor o mayor comple dad del calculo y, por consiguiente, la solución del modelo.

Para obtener los valores concretos de 2 y 2 sustituiremos en ambas expre”sones, (2) y (3), la estimación E, obtenida en (1), resultando

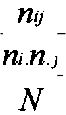
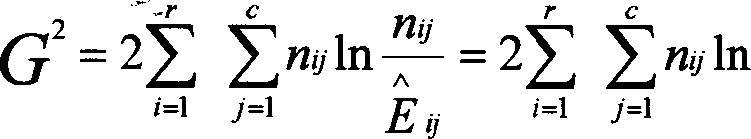
(4)



~~' '~~

Análogamente para el estadístico G2

r 6



= 2Z Zuv ln

Si el resultado de la constrastación, fijado el nivel de significación adecuado, a sido no rechazar la hipótesis de independencia (y„ = 0)equivaIente, como sabemos, aceptar que el modelo log-lineal tiene por expresión

El paso siguiente será la estimación de los efectos principales, debiendo tener

presente que, al haber aceptado la hipótesis de independencia, las

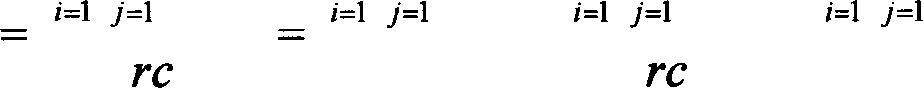
estimaciones de los efectos deberán ser calculados aplicando esta condición, expresión (1). Recurriendo a las formulas (12), (13) y (14) de la sección

anterior, y sustituyendo en ellas E, por E:, en (1), llegamos a los siguientes resultados:

Estimación de la media general

*r c*

ZZ •\*, ZZi=. +ZZi=, —II',=



— In N (6)



*r*

*rc*

Estimación del efecto principal del factor *A*

(7)

Estja›ación del efecto principal del factor *B*

## . .

. Zln £v

p2(j) = —' ' —

*r*

ZZln €¢ *r Ann.; +*

*rc*

*r*

1nn.‹ — *r* ln *N*

*r*



Se comprueban fácilmente las dos condiciones que deben cumplir los efectos.



Si la contrastación condujera a no aceptar la existencia de independencia, es decir que no todas las interacciones son distintas de cero, el modelo saturado al que se ajustan las frecuencias, deberá ser:

**lnfv'ç + ç+)+ ç2Ü›+ çiqy**

La no aceptación de la independencia de las dos características, o lo que es equivalente la presencia de asociación, Ileva consigo que la expresión (1) no puede ser utilizada en los cálculos posteriores a la hora de obtener las estimaciones máximo verosímiles de los efectos, puesto que solo es valida dicha expresión en el supuesto de independencia, por lo cual las estimaciones máximo verosímiles precisas se han de basar en el conocimiento que tenemos de la estructura poblacional de los dos factores, y este conocimiento no es otro que la propia tabla de contingencia, por lo cual la estimación máximo verosímil

de las frecuencias esperadas será en este caso E, = *no, ,* que sustituido en (12),

(13) y (14) de la sección anterior, proporciona las siguientes estimaciones de los parámetros

(9)

*r'( --* ¡Zlnn; *— y*

1

(10)



SU 2(j) —

Z

*r*

lnn« — (11)

Obtenidas las estimaciones de estos parámetros, el paso siguiente consiste en él calculo de las interacciones .. , utilizando para ello la relación

y iz( ) = lnnb — y+ , i( ) + y z(; (12)

las estimaciones ,u ‹,› verifican, las dos condiciones

J t2(y) = 0

Efectuados los cálculos pertinentes es posible comparar entre sí el efecto que producen los diferentes niveles de cada factor sobre la cuantía de las

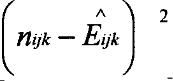
frecuencias, así como, si no se ha aceptado la hipótesis de independencia, los distintos efectos que inducen conjuntamente las combinaciones de los niveles de las dos características.

* + - 1. TABLAS DE CONTINGENCIA TRIDIMENSIONAL

Los cálculos precisos para el análisis de una tabla de contingencia bidimensional mediante los modelos log-lineales son llevados en lo referente a la penosidad del cálculo, aunque sea tedioso. Las cosas cambian cuando trabajamos con tablas multidimensionales, incluso con solo tres factores porque, junto a las dificultades de cálculo, aparecen otras que lo complican excesivamente.

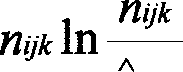
* + - * 1. TIPOS DE MODELOS

Enumeramos, siguiendo a Haberman (1970), los modelos log-lineales tridimensionales jerárquicos básicos que responden a las diferentes hipótesis que podemos establecer, ya presentados en la tabla 1, indicando en el de tipo 2 la necesidad de utilizar un método de estimación aproximado.

El primer paso a dar en el proceso de estimación de un modelo es el de la contrastación de la hipótesis de partida. Esto Ileva a cabo, mediante el test 2 o el G2 cuyas expresiones son, para el caso de tres factores,

*r*

2

G 2 

*i)”k*

distribuyéndose como una y' , dependiendo él número de grados de libertad de

cada tipo de modelo (cada hipótesis establecida y aceptada). Ruiz Maya, F.J. Martín Pliego (1998)

El esquema de presentación de los modelos es el siguiente: hipótesis sobre los efectos, modelo, estimadores máximo verosímiles de las frecuencias

esperadas, E, , necesarias para contrastar la hipótesis y estimar el modelo, número de grados de libertad, y expresiones generales de los estimadores de los efectos.

En la tabla de tres dimensiones él número de niveles de los factores es: Factor

*A, i* = *1...r* factor *B,* j =1...c; factor C, *k =1...s.*

TIPO 1: MODELO SATURADO 123]

Hipótesis: ninguna.

Modelo:

In *Ei;k —— y + yi(! +* ,fJ2(j) + 3(I) -1-

+ 12( ) -Í- 13(ii) -1- ,iz23( i)

-1- / t23(/Jk)

Estimadores máximo verosímiles de las frecuencias esperadas:



Grados de libertad:

GB = *res* —[1+ *(r* —1)+ (c —1) +(s —1) + (r —1)(c —1) +(r —1)(s —1) +

+(c —1)(s —1) +(r — 1)(c —1)(s — l) = *res — res ——* 0

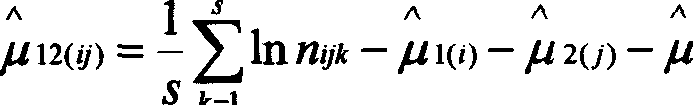
Estimadores de“ los efectos:











1

—*r* Zln *n!;k* — , f 2(j) — ,rf 3(k) — r



TIPO 2: INTERACCIÓN DE ORDEN DOS NULA, MODELO DE ASOCIACIÓN PARCIAL [12) (13] 23)

Hipótesis:

@\)3(i)k) *——* 0

Modelo:

ln *Eijk — fi + fi I(i) +* j£f2(J) + jfJ3(l) +



Esti dores máximo verosímiles de las frecuencias esperadas:

Las estimaciones máximo verosímiles de las frecuencias esperada se obtienen mediante métodos aproximados, pues no las hay directas.

Grados de libertad:

fi€ = *res* —[1 +(r —1)+(c — l)+(s—1) + *(r* —1)(c —1) + *(r* —1)(s —1) +

+(c—1)(s —1)] = *res — r(c + s* —1) — (c—1)(s —1) =

*—— res -rc -re — ce + r + c + s —* 1

Estimadores de los efectos:

i *rr s »*

—*rs*

ZZln *E i;k*

i=1 *k=i*



*5 k--\*



TIPO 3: MODELO DE INDEPENDENCIA CONDICIONAL 12] 13)

Hipótesis

Modelo

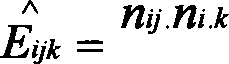
j£f23(j£) —— 0 ¡

*fl 23(ijk) ——* 0



+ jf212(y) -1- jtJ13(rá)

Estimadores máximo verosímiles de las frecuencias esperadas:

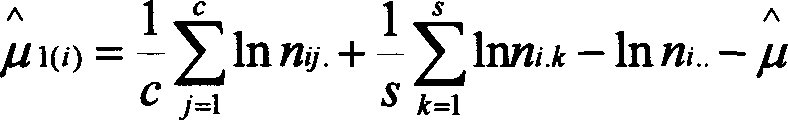


Grados de libertad:

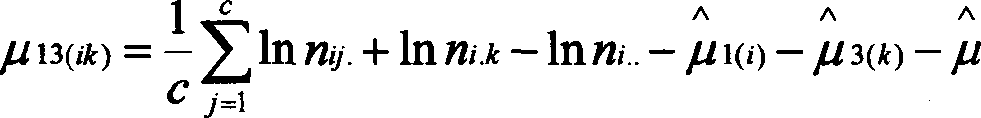
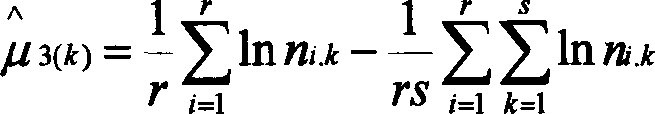
*GB-—— res —(I +(r —* 1)+ (c— 1) +(s— l)+ *r —* 1)(r — 1) + *{r — i) s —* 1)] =

= *res* — r(c + s — l) = *r cs — c -s +* l)

Estimadores de los efectos:







TIPO 4: DOS FACTORES INDEPENDIENTES DEL TERCERO. MODELO

DE INDEPENDENCIA PARCIA [12) (3)

Hipótesis:

Modelo

*fi15(ik) ——* 0 ; y23(ji) —— 0 ; ,f2l23( á) = 0

ln *Eijk — fi +* j£fl(i) -Í- jfJ2(J) -1- j£j3(l) -I-



Estima@res máximo verosímiles de las frecuencias esperadas:



Grados de libertad:

GB = *res -[I + (r —* 1)+(c — 1) + (s — 1) + *r —* 1)(c — 1)] =

*-— res - re + s -*1) = *re -*1)(s — 1)

Estimadores de los efectos:

l *rr c*

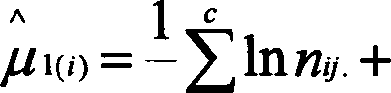
lnn

ZZ

*re*

+ 'Zlnn.» — lnN



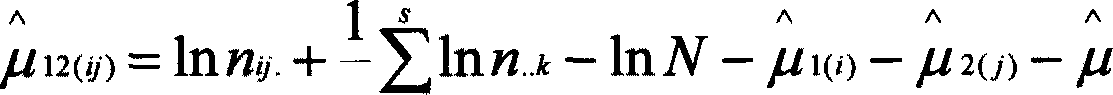


*( rr*



—*r* Zlnnr

(f3( ) *—- X n..k -* Zlnn..i



**TIPO 5: TODOS LOS FACTORES INDEPENDIENTES. MODELO** DE

**INDEPENDENCIA TOTAL** 1) (2)(3)

Hipótesis:

Modelo

yiiii,) = 0 ; 13(ii) = 0 ¡ 23(ji) = 0 ; , ii3(iji) — 0



Estimadores”máximo verosímiles de las frecuencias esperadas:

*n›..n. .n..›*

**Grados de libertad:**

*Eijk —*

*N’*

GB = *res* —[1+ *(r* —1)+ (c —1) + (s —1)] =

*—— res — r — c — s +* 2

,Estimadores de los efectos:

1



*r*

/=t



1





C j=1 ›5' *k*

yip,› = mm — — lnn.

*r* -

1 •

j(J 2(J› = In n , — — In ii.,.

i

y 3( › = Inn i — — f inn

TIPO 6: DOS FACTORES Y SU INTERACCIÓN

Hipótesis:





Modelo

ln *Erik ——* ,u + yi(i) + (f3(l) + l3(i£)

Estimadores máximo verosímiles de las frecuencias esperadas:

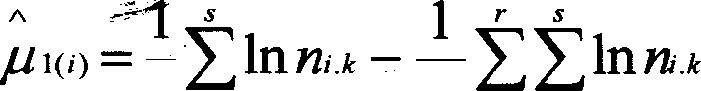
*E.,k -— ni.k*

Grados de libertad:

GB = *res* —[1+ *(r — I) + s —* I) *+ r* —1)(s — 1)] =

*—— res — re —— rs(c —* 1)

Estimadores de los efectos:



*r*

*re*

*y* i3(Ok) *——* Inn *.k —* lnc — u» — 3(l› —

**TIPO** Ç‹ DOS FACTORES SIN INTERACCIÓN

Hipótesis:

j£f2(j) = 0 ; jIJl2((¡) 0 ; yi›(ii) = 0 ; ,ff23(jk) = 0 ; **,(fl23(yk)** = 0

Modelo



Estimadores máximo verosímiles de las frecuencias esperadas:

*n:..n..›*

*cN*

Grados de libertad:

GB = res —[1+ (r — *I)(s —* 1)] = *res — rs + r + s —* 2

56

Estimadores de los efectos:

, 1(›) = ln *ni. —* lRf2‹..

y 3( )= lnn —

l ' lnn ›

*5 k--\*

—Z

TIPO 8: UN SOLO FACTOR

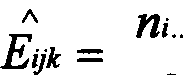
Hipótesis:



Modelo:



Estimadores máximo verosímiles de las frecuencias esperadas:



Grados de libertad:

*GL —— res —[I + r —* 1)]= *res - r —— r cs — I)*

P8timadores de los efectos:

*( rr*

###### —Zlnn

— Inc — lris — In *N*

TIPO 9: NINGÚN FACTOR

Hipótesis:

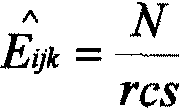


Modelo:



57

Estimadores máximo verosímiles de las frecuencias esperadas:



Grados de libertad:



Estimador de ,u

*y =* In *N -* Inr *—* lno —Inc

* + - 1. MÉTODOS APROXIMADOS DE ESTIMACIÓN

El modelo de tipo 2 no tiene estimaciones máximo verosímil directas de las frecuencias esperadas, lo que exige el recurso a métodos de estimación aproximados.

Estos métodos se basan en algoritmos iterativos de aproximación que converjan mas o menos rápidamente a estimaciones ciertamente estables. Aunque se han propuestos varios procedimientos, detallaremos dos métodos: el mido iterativo de ajuste proporcional de Deming-Stephan y el método de Newton-Raphson, dado que los otros algoritmos no son sino variantes de estos dos y además son estos procedimientos los que desarrollan la mayoría de los paquetes informáticos.

* + - * 1. MÉTODO ITERATIVO DE AJUSTE PROPORCIONAL DE DEMING STEPHAN (1940)

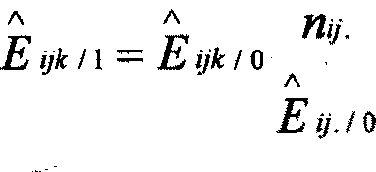
Independientemente del método de cálculo concreto, el procedimiento consiste en tomar como elementos de referencia, las condiciones que se deben cumplir, las distribuciones de frecuencias marginales bidimensionales. Estas distribuciones se denominan configuraciones, existiendo en el caso tridimensional tres configuraciones bidimensionales. *C, »* \*3 y C’,. La

configuración *C„* es el conjunto de frecuencias *n, , C„ e\* conjunto de frecuencias *no* 3 y C',3 el de las frecuencias •.3 -

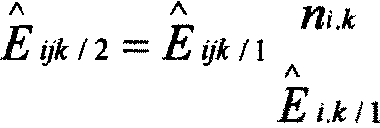
El procedimiento de cálculo se inicia atribuyendo a las frecuencias esperadas el valor unidad *£ Jk IO ——* I , a continuación se obtiene el valor de la configuración

+ 2 *n, y* \*„ .  partiendo de los valores de esta configuración se inicia la

primera iteración que, como todas las que se realicen, tiene tres pasos. En el primer paso se utiliza la expresión



Con lo que nos ceñimos a la restricción del cumplimiento de la configuración *Co.* dado que si sumamos tos *E jfk › ,* respecto a K, el resultado es igual a *n, .* Obtenidos los valores *Ei)k I›* se calcula la configuración C'„ constituida por C' . y per *El. i* (segundo paso de la iteración) para lo cual la fórmula es:



De la mismá manera que el primer paso *(C„ )* la suma de la aproximación de las frecuencias esperadas (<.3 )\*,«z es igual a n, .

Por último, y como tercer paso, se calcula la configuración *Cp n y E ,*

mediante la siguiente expresión:



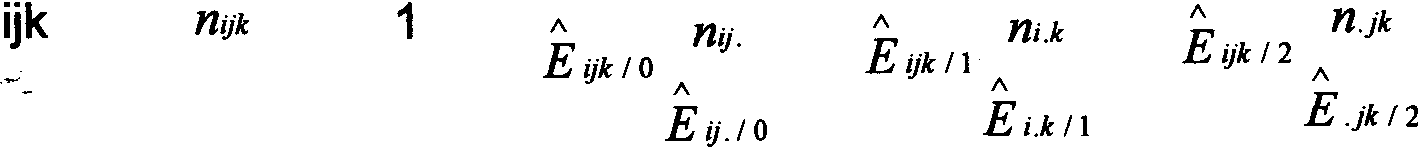
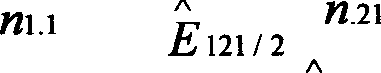
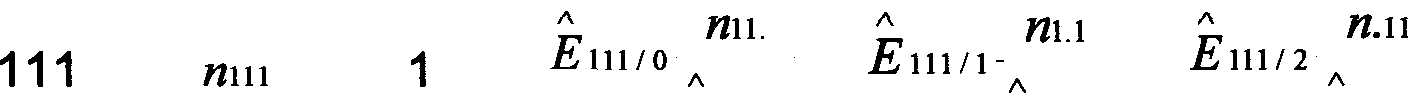
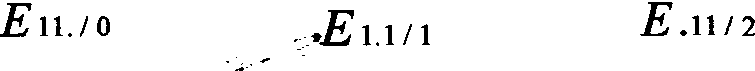
*E , jk / 2*



Con lo cual se termina la primera iteración y la obtención de la primera aproximación a la estimación de las frecuencias esperadas A,« 3 . El paso a la iteración siguiente se realiza sustituyendo en la tabla los valores de A 0 por

En la tabla N° 3, hemos representado de una manera sintética los tres pasos de la primera iteración.

Tabla N° 3: Representación de la primera iteración del Método iterativo



*ijk* / 1

*E ijk 1 2*

121

FGS

*ni zi*

ri<•

1

'

72l2.

t2i / 0

*E i z i o*

*E,.... nos*

*E re. 1* 0

*'E ...ti ~~p""~~*

*E*

*E rcsl2 ~~‘!~~*

*E los 1* 2

Configuración C1›

*Celdas nij. E i;. i o*

11. *no. E ii.i s*

12. nia. A iz. / o

 rió. *E t.i* o

 *nrc. E* re. / 0

Configuración C1«

*Celdas ni.x E i.k l i*



i.k 

r.k 

Configuración C›3

*Celdas n › E ;k 12*

.11 *n.i i* 6.it / 2

. 12 rt.iz *E .i i z*

.jk 

 72.rx .rx / 2

La segunda iteración sigue los pasos anteriores con la única salvedad que ahora se comienza el primer paso, con los últimos valores obtenidos en la iteración anterior procediendo igual se continúan las siguientes iteraciones.

Una vez que se ha acabado una iteración se plantea el problema de sí seguimos con otra o no, es decir, si las estimaciones calculadas satisfacen

nuestras necesidades de acuracidad o es preciso calcular una iteración más. Para dar respuesta a esta pregunta se recurre al procedimiento siguiente:

Fijado el error absoluto que estamos dispuestos a cometer se calcula la expresión

A ijz / 3/ — A *ijk I el -* 3 < 6

siendo / él número de la iteración. Cuando para todo *ijk* se verifica la desigualdad podemos detener él calculo de las estimaciones de las frecuencias esperadas.

* + - * 1. MÉTODO NEWTON-RAPHSON

El algoritmo de Newton-Raphson se basa en una aproximación cuadrática de la función de verosimilitud, aproximando el valor de dicha función a través del desarrollo en serie de Taylor, alrededor del valor óptimo que se desea encontrar

En la aplicación supondremos que él número de observaciones de cada celda de la tabla sigue una distribución de Poisson, pues es él supuesto que no impide restricciones a priori en los parámetros.

Por tanto, en este caso, la función de verosimilitud será, para una tabla *AxBxC:*



Donde el logaritmo del núcleo de la verosimilitud es:

*r*

n«i ln £

— Z

f l

(13)

Por otra parte, siendo E y U, los vectores columna de las E„ y de los

parámetros del modelo log-lineal, este puede expresarse matricialmente a través de:

ln € —— X *U* (14)

Siendo X la matriz de diseño que va a caracterizar cada modelo particular.

Para ilustrar como puede definirse X, supongamos que para una tabla 2x2 propugnamos un modelo de independencia, es decir, suponemos que:

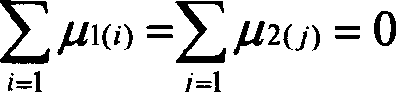
Sin interacciones, por tanto. En este caso sencillo, la expresión matricial del modelo será:

(15)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| In *lii* | *1* | *1* | 1 |  |
| ln *Ei2* | 1 | 1 | — 1 |
| ln *Ez* | 1 | —1 | l | "" |
|  |  |  |  | j£f 2(1) |
| In€22 | 1 | — 1 | 1 |  |

Donde la matriz Xesta formada por 1 y -1, de manera que la codificación establecidWpermita que se cumplan las condiciones

2 2



En (15) podemos obtener, por ejemplo:



Dado que, como antes habiamos establecido:

2

Z «°(•)'/22(I) -1- /22(2) = 0

v I

Siendo, por tanto, la segunda fila de la matriz de diseño X, el vector (1,1,-1).

Por supuesto, en cada caso concreto hay que definir X de manera adecuada.

Volviendo a las tablas *AxBxC,* el modelo saturado viene dado por

In *Ei;k —— y -+- y I(i) +* j£f2(j) + 3(£) + £f l2(i ) -i- j£f13(1£) -I•

-t- j£f 23(yI) -Í- /2t23(yk) (16)

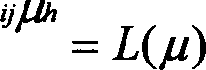
Denotado ahora por (E„ E .--- ...Ej) y por *(y,n •-----•N )* las frecuencias poblacionales y observadas, E,, y *•,›k .* con objeto de simplificar las expresiones y teniendo en cuenta la expresión matricial (14) del modelo, podemos establecer que cada elemento, puede expresarse como

*h*

Donde los son los elementos de la fila i-ésima de la matriz de diseño X, y ,u, representa cada uno de los parámetros que el modelo incluya.

A partir de esta expresión (17), y adoptando la nueva anotación simplificada tendremos que el logaritmo del núcleo de la verosimilitud que obteníamos en

(13) queda:

*h * (18)

Z•

Es decir, se ha reparametizado el núcleo de la función de verosimilitud expresando ecta en función directa de los parámetros en vez de las frecuencias esperada E .

El algoritmo de Newton-Raphson para la estimación aproximada de los parámetros, viene dado por la expresión iterativa



(19)

Siendo *DE y e\* vector de primeras derivadas de 1(,u), valorado en y

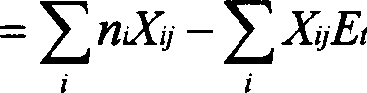
*D'£ y* la matriz hessiana correspondiente a sus segundas derivadas,

valorado también en el resultado obtenido en la iteración anterior.

En nuestro modelo (18), el vector *DE y* y la matriz hessiana *D'£*

pueden obtenerse teniendo en cuenta que:

*d; ——*







Y, por tanto, sus valores en la t-ésima iteración serán:

*d' —— Z •!; ni — E')*

Siendo ET t-ésima aproximación a E que puede obtenerse a través de:

.

\*'= e “

De acuerdo con la expresión (14).

Estos resultados Ilevan a que la expresión matricial (19) en el modelo log-lineal, teniendo en cuenta que:

Quedaría

*DL( ) —— (n — E')*

*D 2 L ) —— — Diag E' )J*

*o+ ! —— i•+ [x'Diag( z’)X ) X'(• — E'* (20)

Donde X es la matriz de diseño, n el vector de frecuencias observadas, E’ el vector de frecuencias esperada obtenido en la t-ésima iteración y *IJiag(E' )es*

una matriz diagonal cuyos elementos no nulos de la diagonal principal serian

los valores de ET

El proceso iterativo puede empezar con



Si no hay ninguna celda vacía, o con



Es el caso de que hubiera alguna n, = 0.

La estimación de la matriz de covarianzas de los parámetros es precisamente



Que se obtiene como subproducto de este algoritmo de Newton —Raphson

Por otra parte, la expresión (20) puede reformularse adoptando forma de solución del método de mínimos cuadrados ponderados (Agresti A., 1990), por lo que este método también recibe el nombre de método de aproximación de los miramos cuadrados iterativamente moderados, pues las ponderaciones van cambiando err cada iteración.

4.7 ANÁLISIS DE RESIDUOS POR CASO

Los tests 2 y c' proporcionan una idea global de cómo los datos (frecuencias observadas) de la tabla de contingencia proceden de una población que obedece al modelo log-lineal que hallamos considerado, es decir, permite establecer la bondad del ajuste de esos datos al modelo establecido, pero no informan sobre la relevancia particular en ese ajuste de los términos ,« del modelo, ni de los casos extremos que puedan darse en cada

una de las celdas de la clasificación cruzada.

Dos tipos, pues, de análisis deben completar el diagnóstico que, sobre las estimaciones, proporcionan los conocidos tests a' y G2 en el estudio de una

tabla, que se pretenda describir a través de un modelo log-lineal. Estas dos clases de análisis se basan en la estimación de las diferencias entre las frecuencias observadas y las proporcionadas por el modelo en cuestión, diferencias denominadas residuos, y en su grado concreto de significación estadística. Veamos cual es el modo de proceder.

4.7.1 ANALISIS DE RESIDUOS DE LOS PARÁMETROS DEL MODELO

Este es estrictamente aplicable al caso de modelos saturados y completos, es decir, donde necesariamente todas las estimaciones de las frecuencias esperadas sean mayores que cero, en ausencia de ceros estructurales.

La significación estadística de los parámetros de un modelo saturado puede

estudiarse a través de su estimación tipificada, bajo el supuesto de que dicho parámetro sea nulo, es decir, usando los residuos tipificados.



Doqde la estimación del error típico 5,, también se obtiene de acuerdo al método de estimación directa o aproximada que se halla utilizado.

Para grandes muestras » sigue una distribución asintóticamente normal con media cero y varianza unidad. Por ejemplo, un nivel de significación del 10% un

valor de w mayor que 1,65 indica que dicho parámetro ,u es significativamente distinto de cero.

El uso de este análisis residual permite ordenar tanto los efectos principales e interacciones de acuerdo con su grado de significación y magnitud como, en su caso, reducir el modelo cuando algún parámetro ,u sea significativamente igual

a cero marcándonos la pauta del modelo no saturado que parezca mas adecuado, así como la posibilidad de colapso de la tabla (ver la sección 4.5.2) Goodman (1986) obtuvo una estimación asintótica del error típico *Se ,* igual para todos los términos de un modelo saturado correspondiente a una tabla

bidimensional (AxB): dicha estimación es





*(rc)2*

de manera que, en este caso, los residuos de los parámetros son:





*Su*

En modelos no saturados todas las expresiones anteriores, relativas al error típico de los parámetros , pueden considerarse como cotas superiores de las

estimaciones asintóticas de dichos errores, y los correspondientes tests pueden utilizarse también como cotas superiores de los correspondientes a estos modelos no saturados.

4.7.2 ANÁLISIS DE RESIDUOS EN LAS CELDAS

Resulta de interés el estudio del grado de ajuste en cada una de las celdas de una tabla de contingencia, con objeto de verificar si, a pesar de que el ajuste global del modelo sea satisfactorio, existen algunas celdas donde no lo sea tanto, así como para detectar la presencia de valores extremos, o para establecer el patrón positivo o negativo de las desviaciones de las celdas.

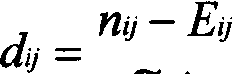
AI calcular estas desviaciones entre los valores observados, *ni ,* y las

estimaciones de los esperados, *E ,* quizá ocurra que se obtengan diferencias

grandes, lo que indicara que las combinaciones de los niveles donde tal cosa suceda presentan particularidades de interés, o bien que los valores de las celdas son, como decíamos, valores extremos. Por ello, el análisis de tales desviaciones es capaz de arrojar, en muchas ocasiones, importante Iuz sobre la estructura de la tabla de contingencia.

Como casos límites una desviación pueden ser alta simplemente porque lo sean los dos términos que la integran, o pequeña, por serlo también los dos términos: en cualquier circunstancia, estos valores límites pueden enmascarar situaciones extremas. En los dos casos el “elemento perturbador" es la propia magnitud, alta o baja, de las dos frecuencias (observada y esperada), por lo cual se hace preciso eliminar esta influencia.

Uno de los procedimientos de la determinación de la significación de tales desviaciones se basa en el análisis de los residuos tipificados



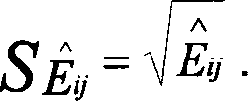
en do¥de *Ç ,* es error típico de la estimación de las frecuencias esperadas

Si tomamos como base de referencia del grado de bondad del ajuste el test y', los residuos tipificados son





La estimación del error típico implícitamente supone una distribución de Poisson para el número de observaciones de cada una de las celdas, por lo cual



Estos residuos tipificados *de* se distribuyen asintóticamente Goodman (1986) como una normal de media cero y varianza uno, por lo que, al igual que en el caso anterior de los residuos tipificados de los parámetros, un valor de *de*

mayor que, por ejemplo, 1,65 indica que la desviación »« — 6« es significativamente distinta de cero, a nivel de significación del 1OO/o.

4.8 SELECCIÓN DEL MODELO

En una estructura jerárquica pueden establecerse distintos modelos sin saber, de antemano, de forma razonablemente segura cual de ellos será él mas adecuado a los datos observados. Existe el criterio de optar, de entre varios modelos igualmente aceptables, por él más sencillo, el que contenga menos términos (criterio de parsimonia). Es evidente que la aplicación de este criterio solo es factible cuando se conoce cuales son los modelos “mejores” siendo, por consiguiente, este aspecto el primero a dilucidar por lo qMe, para ello debe investigarse en que medida se ajustan a los datos los diferentes modelos, de manera tal que podamos elegir aquel o aquellos que mejor expliquen la estructura de la tabla de contingencia.

Para determinar la inclusión o no de los parámetros en el modelo log-lineal pued seguirse un procedimiento paso a paso (stepwise) que consiste en propugnar un modelo y, a partir de él se van añadiendo o suprimiendo parámetros.

MÉTODOS COMPUTACIONALES:

1. El método de selección Forward, que parte de un modelo sencillo, añadiendo parámetros mientras que estos no sean significativamente iguales a cero, según su nivel de significación fijado de antemano.

2. El método de selección Backward, supone un análisis en sentido contrario al caso anterior, puesto que normalmente se parte de un modelo saturado, simplificándolo, según se pueda, de aquellos parámetros que no sean estadísticamente significativos.

Para admitir o rechazar un parámetro, en un paso cualquiera en estos procesos, es necesario determinar su grado de significatividad según advertíamos y, además, la posible ganancia en términos en algún estadístico que mida la bondad de ajuste, por lo que vamos a exponer los métodos mas frecuentemente utilizados en la selección de modelos.

* + 1. MEDIANTE ANÁLISIS DE LOS RESIDUOS

En el acápite anterior exponíamos el procedimiento que se basaba en la utilización de los residuos tipificados:



bajo la hipótesis nula de que el parámetro estudiado fuera nulo, ,u = 0.

Al tener w una distribución asintóticamente normal, para• un determinado nivel

de significación si " era mayor que su correspondiente valor crítico se

rechazaba el que el parámetro fuera igual a cero, siendo significativo su valora, por tanto, su presencia en el modelo.

Este procedimiento puede utilizarse indistintamente en procesos forward o backward en la inclusión o eliminación de parámetros.

* + 1. ESTADÍSTICA 2 AJUSTADO

Sean dos modelos (1 y 2) que difieren en uno o más parámetros cuyos estadísticos razón de verosimilitud vienen dados por cx' y G 2 cuyos grados de libertad son *ra* y *r2,* respectivamente. Una medida de su relativa bondad del ajuste esta definida por:

2 = 1 — G2'

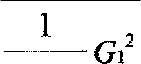
Gi'

El estadístico 2 ajustado, que designaremos por Jt ', es semejante al jt° ajustado de los modelos de regresión, definiéndose como:

2

G2

*q -rz*

*q — rt*

siendo q él número total de celdas de la tabla de contingencia, coincidente con los grados libertad del modelo saturado.

Un valor alto de £" indicara que el modelo 2 se ajusta bien, puesto que

G2' / *q — rá ,* toma un valor pequeño.

* + 1. CRITERIO DE INFORMACIÓN DE AKAIKE

Akaike (1973) sugiere un criterio para selección del modelo. Este criterio recomienda elegir el modelo que minimice la expresión:

*2 (*

*A —— G*

2 y

*q —* 2F)

dc›ede G

*r* tienen la misma interpretación que el criterio anterior.

Si tenemos una relación de posibles modelos log-lineales para explicar la tabla,

elegiremos aquel que proporcione el menor valor de A.

* + 1. CRITERIO DE SELECCIÓN BACKWARD DE AITKIN

Aitkin (1978,1979) sugiere un método de selección del modelo, cuyo procedimiento es el siguiente:

Sean dos modelos 1 y 2, cuyos estadísticos razón de verosimilitud son ci g,', y en donde se supone que el modelo 1 esta anidado en el 2 (es decir,

2 y

todos los parámetros del modelo 1 están contenidos en el modelo 2), se puede demostrar que el estadístico

sigue, asintóticamente, una distribución ji-cuadrado con un número de grados de libertad igual a la diferencia de los grados de libertad de cada uno de los dos modelos.

Si el valor del estadístico 2 es significativo, al nivel fijado, entonces el modelo que incluye los parámetros adicionales, es decir, el modelo 2, proporciona una mejor explicación de la estructura de la tabla.

4.9 LA COMPLEMENTARIEDAD ENTRE EL ANÁLISIS DE

CORREÉPONDENCIAS Y EL ANÁLISIS LOG-LINEAL

En esta sección mostramos cómo pueden relacionarse eIAC y el AL. Esto Ileva a tener un buen conocimiento del AC. Y cómo esto puede llevar a un uso combinado de ambas técnicas.

LAS RELACIONES ENTRE EL AC Y MODELOS CON LAS RESTRICCIONES ELLAS INTERACCIONES

Una manera de superar el problema de interpretar un número grande de parámetros log-lineales es restringir los parámetros de la interacción de alguna u otra forma; esto se hace en el modelo de asociación (Andersen, 1980; Goodman, 1979, 1981, 1985, 1986). Cuando el número de categorías es grande, el número de parámetros a ser interpretado puede reducirse considerablemente de esta manera.

Se sabe bien que el AC y el modelo de asociación están relacionados de la siguiente manera. La representación de AC en las dimensiones de k puede escribirse en una versión adaptada como: (Van del Heijden, P.G.M.; De Leeuw,

J. (1985)"Correspondence analysis used complementary to log-Iinear analysis")

*p! -— p!. p.* 1+ Z \*• ,« (1) Reconstitución de los datos



donde:

(//) Un elemento de la matriz *D (tt' + R C . D, :* Matriz diagonal fila

*D, :* Matriz diagonal columna

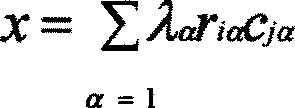
t: . Vector unitario

A : Matriz diagonal con valores singulares 1,

Jt = *D, \* 2 U y *= D,""' F* ; donde *U U = I = K K*

Así *R' D,R —— —— C D,C*

donde: *po* es la proporción reconstituida para la celda (i,j). Escoufier (1982) notó que si

*k*

es pequeño comparado la unidad (entonces el log(1 + x)= x) entonces nosotros podemos volver a escribir la ecuación (1) como:

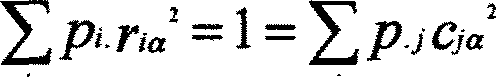
*k*

lOg *piJ* Ofi + ,tf1(i) + 2(J) + ioú,«

(2)



donde = 0, *y›« ) = logpr.,* yi( i —— logp . Desde *nz y c;z* son normalizados tal que



La condición de Escoufier reduce bruscamente a la situación que la salida de la independencia no es demasiado grande. Compare con la ecuación de relación entre ' y los valores singulares al cuadrado, en 2

(3)

(La ecuación (3) muestra que el AC descompone el valor de la chi-cuadrado para testar independencia en la matriz).

Otro resultado es que, si k = 1 y las proporciones vienen de una distribución normal bivariante discreto (o una distribución que es normal bivariante después de una transformación conveniente de las filas y columnas), entonces no importa que tan grande sea el valor singular , la ecuación (2) se relaciona estrechamente al modelo de asociación.



(Goodman, 1981). En esta expresión, - J, , *ni* » /z und y *ni* « /z 2( › donde

(.› 2s,› son normalizados en la misma forma como », y c' . La

diferencia entre la ecuación (2) y la ecuación (4) es qúe la ecuación (2) es una aproximación que involucra k factores, mientras que la ecuación (4) tiene sólo un factor. Comparado con el modelo saturado de log-lineal (ecuación (17) de la sección 4.Ç.2), el modelo (4) puede interpretarse como un modelo que tiene en centa la interacción dónde:

«) =$// '›‹•›,«'2‹»

Dados los objetivos de ambos métodos y sus posibilidades o condicionamientos, la situación ideal es aquella en que el AL nos oriente hacia el modelo más adecuado para las variables cualitativas en estudio (es decir, responde a la cuestión de las variables que están relacionadas) y el *AC* nos orienta sobre la relación entre los valores, lo que es equivalente a la interpretación de los parámetros de distintos órdenes (es decir, responde a la cuestión de cómo se relacionan las variables: Qué categorías de ellas se asocian más o menos de lo esperado). Uno se refiere al nivel de las variables, otro, al de los valores o categorías de las mismas. Por tanto, lo primero es

asegurarse del modelo, por medio del AL con el 1' de máxima verosimilitud y,

posteriormente, ver las relaciones entre categorías, interpretación de los parámetros, por el AC.

P.G.M. Van del Heijden y J. de Leeuw (1985: 435-446) estudian y aplican esta relación. Su análisis y ejemplificación los hacen solamente para dos y tres variables.

El AC descompone la diferencia (n, — A, ), donde los *E, —* valores

esperados bajo un modelo determinado, se calculan a partir de los marginales de *mi .* El contenido de la relación establecida entre el AL y el AC y la lógica de

su demostración (P.G.M. Van del Heijden y J. de Leeuw, 1985) consiste en mostrar que los valores *n,* y A, son iguales a la estimación de máxima

verosimilitud de las frecuencias esperadas bajo un modelo log-lineal especificado.

El resultado hace posible interpretar la solución del análisis de correspondencias en términos de modelos de log-lineal y salvar los problemas de interpretación de los parámetros de log-lineal que surgen frecuentemente. Es decir, que tanto para análisis de tablas de contingencia de dos variables, de tres o de más alto nivel, la solución del AC se toma como la diferencia entre dos modelos log-lineales. Por ejemplo, para dos variables la diferencia entre el modelo saturado y el modelo de independencia, es decir, el parámetro de interacción de primer orden se interpreta como la solución del primer eje del AC. En el caso de tres variables, si uno de los primeros órdenes de interacción (asociación de dos variables) no parece interesante se puede construir una sola variable interactivamente entre las dos.

Cuando se está especialmente interesado en la relación de una variable con las otras dos, ésta no debe codificarse interactivamente con ninguna de las otras dos. La solución del AC en este caso muestra los dos primeros órdenes de interacción, de esta variable con las otras dos, y el segundo orden de interacción.