UNIVERSIDAD DE JAÉN

Material del curso "Análisis de datos procedentes de investigaciones mediante programas informáticos"

Manuel Miguel Ramos Álvarez

Índice

MATERIAL XVI "DESCRIPCIÓN CON ANÁLISIS CATEGÓRICO"

1. Conceptos fundamentales del análisis Categórico

- **Objetivo**: para investigaciones en las que todas las variables están medidas en una escala nominal u ordinal (variables categóricas o categorizadas artificialmente) y el objetivo es determinar si hay relaciones entre variables o mejor dicho si hay asociaciones.
- **Lógica estadística**. Se parte de tablas de contingencia de múltiples entradas, una por cada dimensión de la asociación (por cada variable que se ha medido), en la que aparecen frecuencias conjuntas. Estas pueden ser bidimensionales o multidimensionales (más de dos vías). Alternativamente las casillas podrían contener proporciones –probabilidades- o razones.
    - En principio, los casos que confirman la asociación aparecen en la diagonal principal, mientras que en la secundaria aparecen los casos que no la confirman. En caso de que predominen los casos favorables, podría pensarse que hay relación entre ambas variables.
- **Conclusión**: Encontrar un modelo óptimo de las frecuencias a partir de las variables por separado y/o la interacción entre ambas (las frecuencias de una variable son una función de la otra variable). Con frecuencia se empieza por el modelo más complejo (cada una de las variables más las interacciones) y a continuación se intenta eliminar el máximo posible sin que la capacidad de explicar las frecuencias observadas disminuya significativamente, lo que lleva a ir probando modelos de complejidad decreciente. Dicho planteamiento en realidad implica que pretendemos explicar las frecuencias (Criterio) a partir de los factores y su interacción (predictores).

- Variantes. Según el enfoque concreto del investigador hay una gran diversidad de técnicas que derivan de este planteamiento.
    - **Descriptivo. Logarítmico-lineal.** En el modelo, se predicen las frecuencias conjuntas a partir de las variables incluidas en el estudio, sin diferenciar entre predictores y criterio. Sirve, pues, para estudiar la asociación de variables en general.
    - **Aproximación explicativa. Logit/Probit.** Se pretende realizar la regresión o predicción de alguna(s) variable(s) en función de otras. Es decir, éstas aparecen claramente diferenciadas.

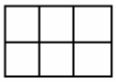
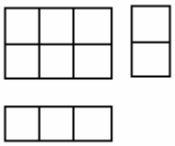
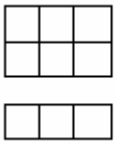
2. Planteamiento computacional del análisis Categórico

- Se establece la predicción de las frecuencias a partir de las variables de la investigación, "n" se introduce para pasar desde las frecuencias absolutas a las relativas (o proporciones) y se plantea en términos logarítmicos para trabajar bajo el supuesto de linealidad, a partir de un modelo original que es multiplicativo.

Modelo Log.Lineal: $\log(m_{jk}) = \lambda + \lambda_A + \lambda_B + \lambda_{AB}$; $m_{jk} = e^{\lambda} e^{\lambda_A} e^{\lambda_B} e^{\lambda_{AB}}$

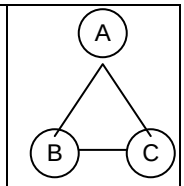
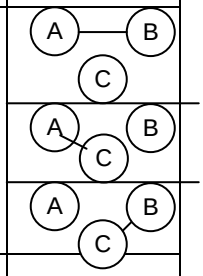
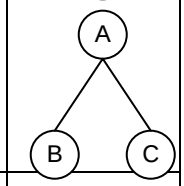
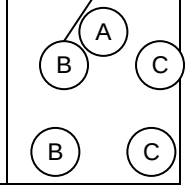
3. Introducción al Análisis Categórico log-lineal

1. Fase de **ajuste** del modelo. Habrá que buscar el modelo óptimo para explicar los datos. Pero ¿Qué tipos de modelos pueden ponerse a prueba mediante análisis categórico? Lo más usual es seguir un enfoque jerárquico, que si un modelo incluye un término de orden superior lo estarán también todos los términos de orden inferior. Por ejemplo, si se incluye la interacción entonces también debe incluir los efectos principales.
    a. Los modelos en orden de complejidad pueden ser:
        i. Para dos variables (tablas bidimensionales o de 2-vías):

	MODELO	SÍMB.	ESTRUCTURA	INFORMACIÓN
(3)	Interacción (saturado)	{AB}	$\log(m_{jk}) = \lambda + \lambda_A + \lambda_B + \lambda_{AB}$	
(2)	Independencia	{A,B}	$\log(m_{jk}) = \lambda + \lambda_A + \lambda_B$	
(1)	Marginal	{A} {B}	$\log(m_{jk}) = \lambda + \lambda_A$ $\log(m_{jk}) = \lambda + \lambda_B$	
(0)	Nulo o de Equiprobabilidad	{}	$\log(m_{jk}) = \lambda$	---------

ii. Para tres o más variables (tablas multidimensionales o de múltiples vías):

Modelos completos

(5) Asociación Completa (saturado)	{ABC}	•Asociación compleja entre todas las var. •La relac entre dos de ellas es modulada por la tercera var.	
(4) Asociación Homogénea	{AB,AC, BC}	•Asociación entre todos los pares de var.	
(3) Independencia Condicional	{AB, AC} {AB, BC} {AC, BC}	•Dos var. son indep a través de todos los niveles de la tercera (i.e. B y C son indep a través de todos los a_j).	
(2) Independencia en un factor (asociación parcial)	{AB, C} {AC, B} {BC, A}	•Una var. es indep. de la combinac. de las restantes.	
(1) Independencia Completa	{A, B ,C}	•Ninguna var. depende de las otras ni de combinaciones de las mismas.	

Modelos incompletos y marginales

(3) Asociación dos variables	{BC} {AC} {AB}	• No Ef. var. que se colapsa. • No relac. entre var. que se colapsa y las especificadas.
(2) Independencia Marginal dos variables	{B, C} {A, C} {A, B}	• Lo anterior. • No relac. entre las var. specificadas.
(1) Marginal una variable	{C} {B} {A}	• No Ef. var. que se colapsan. • No relac. entre var. que se colapsan y la especificada.
(0) Nulo o de equiprobabilidad	{}	• Ninguno de los efectos anteriores.

2. **Fase de evaluación del modelo:**
   a. El enfoque es evaluar a los diferentes modelos posibles y **seleccionar aquél que proporcione** el mejor ajuste de los datos con el menor número de parámetros posible. Esto es equivalente a una prueba clásica de **independencia**: si el modelo de independencia no se ajusta de manera significativa entonces puede admitirse que las variables están relacionadas o que son dependientes.
   b. En realidad se recomienda operar en dos pasos sucesivos (algunos programas como Statistica incluyen un algoritmo para automatizar el proceso pero que no funciona con modelos muy simples):
      i. Evaluación del **ajuste global**. la evaluación de cada modelo se efectúa mediante Chi-Cuadrado de Pearson (ajuste frecuencias esperadas según el modelo y frecuencias observadas). La **regla de decisión**: si la complementaria de la probabilidad (acumulada) asociada al estadístico es **mayor o igual que 0,10** entonces el ajuste es aceptable (equivalente a no rechazar la hipótesis nula de buen ajuste).
      ii. Evaluación del ajuste **condicional**. La evaluación global de los diferentes modelos jerárquicos debe ir seguida por una evaluación condicional, en la que se comparan directamente los modelos que producen un buen ajuste. En este caso se recomienda restar el valor del **estadístico G (la razón de verosimilitud)** de sendos modelos y comprobar si la diferencia resulta o no significativa según la distribución Chi-cuadrado. El requisito es que el modelo más compacto esté completamente incluido en el modelo ampliado.
      iii. **Otras medidas de comparación** de modelos alternativos son: el criterio de Información Bayesiano (BIC), el criterio de información de Akaike (AIC), el índice de disimilaridad o proporción de elementos que habría que cambiar para que los valores predichos coincidieran con los observados.
      iv. **A nivel exploratorio.** comprobar que el patrón de residuales es aleatorio, es decir que no sigue ningún patrón sistemático (i.e. gráfico de frec. residuales vs ajustadas).

En el programa Statistica

Statistics -> Advanced linear/Non linear models -> Log-linear analysis of frequency tables -> Input File: Frequencies with coding variables -> Variables: with freq. Counts: Freq, Variables with codes: A_Enton, B_Sign, C_TipoEstim -> Ok -> Ok -> Test of all marginal & partial associations models, Automatic selection of best model

- En 2-vías: forzar a comenzar la búsqueda desde un modelo no saturado.
- En p-vías: probar 1º a relajar el criterio inicial a 0.05.

Ejemplos:
2-vías:

Modelo	G^2	gl	$p(G^2)$	p(Dif)
{AB}	**0,000**		**1,000***	
{A, B}	**0,925**	**1**	**0,336***	**0,092**
{A}	**3,761**	**2**	**0,153***	
{B}	30,194	2	0,000	
{ }	33,030	3	0,000	

*p>0,10 **p<0,05

Modelo Marginal: $Ef : \log(m_{jk}) = 3,76 - 0,41 Enton\phi_j$

Frecuencias

	Sig	SinSig
Enton	70	60
	(65)	(65)
SinEnton	35	22
	(28,5)	(28,5)

Parámetros

λ		**3,762**
λ_{A1}	4,174	**0,412**
λ_{A2}	3,350	**-0,412**

Ajuste: $G^2(2)= 3,761$; p>0,10

3-Vías:

MODELO	G^2	gl	$p(G^2)$	p(Dif.)
{AB,AC, BC}	**2,587***	**1**	**0,108**	
{AB,AC}	**4,543***	**2**	**0,103**	
{AB,BC}	5,044	2	0,080	
{AC,BC}	**2,645***	**2**	**0,266**	
{AC,B}	**4,673***	**3**	**0,197**	0,112
{A,B,C}	**7,203***	**4**	**0,126**	
{AC}	**5,610***	**4**	**0,230**	
{A,C}	**8,140***	**5**	**0,149**	0,112
{A}	11,104	6	0,085	
{C}	42,421	6	0,000	

*p>0,10 **p<0,05

Modelo Marginal: $Ef : \log(m_{jkl}) = 3,71 - 0,32 Enton\phi_j - 0,09 TipEstim$

Frecuencias

Visual

	Sig	SinSig
Enton	70	60
	(62,00)	(62,00)
SinEnton	35	22
	(32,50)	(32,50)

Auditivo

	Sig	SinSig
Enton	50	45
	(51,50)	(51,50)
SinEnton	25	35
	(27,00)	(27,00)

Parámetros (Sis.Ef.)

λ		**3,711**
λ_{A1}	4,034	**0,323**
λ_{A2}	3,389	**-0,323**
λ_{B1}	3,804	**0,093**
λ_{B2}	3,619	**-0,093**

Ajuste: $G^2(5)= 8,140$; $p>0,10$

3. Fase de **Estimación de parámetros**. Una vez decidido cuál es el modelo óptimo y que éste se puede considerar significativo desde las pruebas de bondad de ajuste, se procede a la explicitación de los parámetros del mismo, así como la posible significación de tales parámetros. El modelo ajustado debería incluir parámetros que fueran todos significativos. Para ello se puede evaluar la significación estadística de los parámetros uno a uno mediante una prueba Z. Para más detalles ver el capítulo dedicado al análisis categórico en Ramos, Catena y Trujillo (2004) y en Catena, Ramos y Trujillo (2003).
4. **Interpretación** de la solución obtenida. Igualmente se deberían estimar los parámetros para adoptar decisiones detalladas en torno a los valores posibles de cada una de las variables del análisis, mediante un enfoque de análisis detallado tipo contrastes. Para ello hay que tener presente cuál es el sistema de codificación empleado (sistema de efectos vs. sistema *dummy*). Para más detalles ver el capítulo dedicado al análisis categórico en Ramos, Catena y Trujillo (2004) y en Catena, Ramos y Trujillo (2003).
    a. En este punto interesará también estimar las frecuencias esperadas a partir del modelo óptimo, de cara a la predicción.

Volver Principio