UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

Distribución Bernoulli Multivariada. Una aplicación a la salud oral

Ramón Alvarez - Fernando Massa
Diciembre 2014

Documentos de Trabajo

Serie DT (14 / 03) - ISSN : 1688-6453

Distribución Bernoulli Multivariada. Una aplicación a la salud oral

Ramón Alvarez [1]
Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.

Fernando Massa [2]
Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.

RESUMEN

Existen muchas situaciones en muy variadas disciplinas como la economía, el marketing, la epidemiología, donde la matriz de datos de la que se dispone está formada por datos binarios (unos y ceros) que surgen de trabajar con varias variables aleatorias resultantes de un experimento con 2 resultados posibles en cada caso. Muchas veces interesa analizar la relación entre variables y formar a su vez grupos que den cuenta de esas relaciones.En este documento se presenta la distribución Bernoulli multivariada **(BM)**, la que puede ser caracterizada por un vector de intensidades y una matriz de asociaciones entre las variables binarias. Es importante entonces ver como queda parametrizado este modelo probabilístico y como puede ser estimado.

Se presenta una aplicación en salud bucal para evaluar la enfermedad periodontal en la población adulta uruguaya. Los datos surgen del primer relevamiento nacional de salud bucal, llevado a cabo durante los años 2011 y 2012 en diversos departamentos de Uruguay, donde se entrevistó a personas de 3 grupos etarios (jóvenes, adultos y adultos mayores).

Palabras claves: asociación, distribución Bernoulli multivariada, enfermedad periodontal,intensidad, variable latente.

[1] ramon@iesta.edu.uy
[2] fmassa@iesta.edu.uy

1. Introducción

En este documento se presenta y caracteriza una distribución de probabilidad multivariada que solo puede adoptar los valores cero o uno y que se denomina *Bernoulli Multivariada* (BM). Esta distribución equivale a considerar los vértices de un hipercubo en R^k, cuyas coordenadas son los valores 0 y 1. Una de las primeras aproximaciones a la temática se puede encontrar en (Marshall y Olkin, 1985) donde se plantea la distribución de Bernoulli bivariada.

En primera instancia, la distribución BM podría definirse como el producto de k distribuciones marginales cada una acorde al modelo Bernoulli (Tamhane et al., 2009), sin embargo dicha parametrización solo contempla el caso en el que las variables en cuestión son independientes. Es por esto que aquí se opta por una formulación donde se incluyen la opción de modelar las asociaciones entre las variables. Para ello se siguen las ideas expuestas en (Dai, 2012). Pese a que la naturaleza categórica de las variables permite pensar en asociaciones entre dos, tres o mas de ellas simultáneamente, se toma la decisión de contemplar solamente las asociaciones "dos a dos" de modo de construir modelos mas parsimoniosos.
Sin embargo, la metodología aquí propuesta puede extenderse para tener en cuenta asociaciones de orden superior. El método empleado en este trabajo difiere de la parametrización basada en la dicotomización de la distribución gaussiana multivariada (Cox y Wermuth, 2002) (Tamhane et al., 2009) debido a que, a diferencia de esta, no asume la existencia de variables latentes, lo cual supone una ventaja en cuanto a la simplicidad del modelo probabilístico.

El documento se compone de la siguiente manera. En la sección 2 se considera la construcción de la distribución, comenzando desde el caso univariado, pasando por el bivariado y llegando finalmente al modelo general, presentando las principales propiedades de cada caso. En esta sección se culmina describiendo el proceso de estimación haciendo énfasis en la situación donde los datos presentan valores ausentes. En la sección 3 se presenta una aplicación en salud oral de esta metodología. Se plantean algunos estadísticos para explorar la independencia o asociación entre las variables.

2. Modelo probabilístico

A continuación se plantea la distribución BM comenzando como una reparametrización de la distribución de Bernoulli, para luego extenderla al caso bivariado y finalmente al caso general. En cada etapa se exploran las principales características de la función de masa de probabilidad.

2.1. Caso univariado

La distribución de Bernoulli es utilizada para modelar las variables aleatorias resultantes de un experimento binario (considerando $Rec(X) = \{0, 1\}$) mediante un único parámetro, el cual se interpreta como la probabilidad de obtener un éxito en dicho experimento. La función de cuantía es la siguiente:

$$P(X = x) = p^x(1 - p)^{1-x} \tag{1}$$

La variable aleatoria definida de esta manera tiene esperanza p y varianza $p(1 - p)$. También es sencillo apreciar que esta función de cuantía puede expresarse como un miembro de la familia exponencial.

$$P(X = x) = e^{xlog(\frac{p}{(1-p)})+log(1-p))} \tag{2}$$

De esta manera surge que el "parámetro natural" de esta distribución es el logaritmo del *odd*. Tras llevar a cabo el cambio de variable $\phi_1 = \frac{p}{(1-p)}$, se llega a la siguiente parametrización:

$$P(X = x) = \phi_0\phi_1^x \tag{3}$$

Donde ϕ_1 representa el *odd* de éxito y ϕ_0 es una constante que normaliza la distribución y que se interpreta como la probabilidad de obtener un fracaso. En el caso univariado, esta constante es $\phi_0 = \frac{1}{1+\phi_1}$. Las nuevas expresiones para la esperanza y varianza de la distribución son $E(X) = \frac{\phi_1}{1+\phi_1}$ y $Var(X) = \frac{\phi_1}{(1+\phi_1)^2}$. Pese a que, en un principio, esta reparametrización solo parece complicar la caracterización de la distribución, en dimensiones superiores probará ser de gran utilidad ya que proporcionará gran flexibilidad para incluir las asociaciones entre variables.

2.2. Caso bivariado

En el caso bivariado, si las variables X_1 y X_2 son independientes, su cuantía conjunta podría definirse de la siguiente manera:

$$P(X_1 = x_1, X_2 = x_2) = p_1^{x_1}(1-p_1)^{1-x_1} p_2^{x_2}(1-p_2)^{1-x_2} \tag{4}$$

Luego de realizar el mismo cambio de variable sugerido en el apartado anterior, la cuantía conjunta se expresa de la siguiente manera:

$$P(X_1 = x_1, X_2 = x_2) = \phi_0 \phi_1^{x_1} \phi_2^{x_2} \tag{5}$$

En este caso, la constante de normalización ϕ_0 equivale a $\frac{1}{1+\phi_1+\phi_2+\phi_1\phi_2}$ y se interpreta como la probabilidad de obtener un fracaso en ambas variables. El siguiente paso en la construcción de la distribución BM es el de incluir en la ecuación (5) la asociación entre X_1 y X_2. Para ello se introducirá un nuevo parámetro α_{12} de la siguiente manera:

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2) = \phi_0 \phi_1^{x_1} \phi_2^{x_2} \alpha_{12}^{x_1 x_2} \\ &\phi_0 = \tfrac{1}{1+\phi_1+\phi_2+\phi_1\phi_2\alpha_{12}} \end{aligned} \tag{6}$$

Tras la modificación propuesta, ϕ_0 continúa siendo la probabilidad de obtener dos fracasos.

En cuanto al parámetro α_{12}, es sencillo demostrar que equivale al *Odds Ratio* entre X_1 y X_2. Sin embargo, la interpretación de ϕ_1 y ϕ_2 cambia ligeramente, ya que en este caso pasan a ser los *odds* de éxito de cada variable *condicional* a que la otra variable valga cero. Ya en presencia de ambos tipos de parámetros, nos referiremos al conjunto de valores ϕ_i como *intensidades* o *fuerzas* y al conjunto de valores α_{ij} como *asociaciones*.

El siguiente paso es definir las distribuciones marginales y condicionales de cada variable. En cuanto a las marginales, se puede demostrar que son Bernoulli con la siguiente función de cuantía:

$$\begin{aligned} &P(X_1 = x) = \phi_0^* \phi_1^{*x} \\ &\phi_1^* = \phi_1(1+\phi_2\alpha_{12}) \\ &\phi_0^* = \tfrac{1}{1+\phi_1^*} \end{aligned} \tag{7}$$

La distribución marginal de ϕ_2 es análoga. En cuanto a las distribuciones condicionales, se puede demostrar que éstas también son Bernoulli:

$$\begin{array}{l} P(X_1 = i|X_2 = j) = \frac{P(X_1=i,X_2=j)}{P(X_2=j)} = \frac{\phi_0\phi_1^i\phi_2^j\alpha_{12}^{ij}}{\phi_0\phi_2^j(1+\phi_1\alpha_{12}^j)} = \frac{\phi_1^i\alpha_{12}^{ij}}{(1+\phi_1\alpha_{12}^j)} = \phi_{0|j}\phi_{1|j}^i \\ \phi_{1|j} = \phi_1\alpha_{12}^j \\ \phi_{0|j} = \frac{1}{1+\phi_1\alpha_{12}^j} \end{array} \tag{8}$$

Vale la pena mencionar que al fijar $j = 0$ se obtiene el caso particular de donde surge la interpretación de ϕ_1 como odd condicional. La cuantía condicional de X_2 se obtiene de la misma manera.

2.3. Caso general

La función de cuantía del vector $X = (X_1, X_2, \ldots, X_p)$ en el caso de p variables binarias posiblemente asociadas entre sí es la siguiente:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \phi_0 \prod_{i=1}^{p} \phi_i^{x_i} \prod_{j=i+1}^{p} \alpha_{ij}^{x_i x_j} \tag{9}$$

En este caso, la especificación de ϕ_0 se vuelve un poco mas compleja y para ello se define la matriz de configuraciones H. Esta matriz, que consta de $\frac{p(p+1)}{2}$ columnas y 2^p filas, contiene cada una de las posibles configuraciones del vector aleatorio X en las primeras p columnas y los productos de estas coordenadas en las siguientes $\frac{p(p-1)}{2}$. Adicionalmente se define el vector γ, el cual contiene los $\frac{p(p+1)}{2}$ parámetros del modelo. Se trabaja entonces con $\Gamma = (log\phi_1, log\phi_2, \ldots, log\alpha'_{p-1,p})$. De esta manera, se reescribe la cuantía en función de los elementos de Γ.

$$\begin{array}{rcl} P(\underline{X} = \underline{x}) & = & \phi_0 \prod_{i=1}^{p} \phi_i^{x_i} \prod_{j=i+1}^{p} \alpha_{ij}^{x_i x_j} \\ & = & exp(log(\phi_0 \prod_{i=1}^{p} \phi_i^{x_i} \prod_{j=i+1}^{p} \alpha_{ij}^{x_i x_j})) \\ & = & \phi_0 exp(\sum x_i log\phi_i + \sum x_i x_j log\alpha_{ij}) \end{array} \tag{10}$$

Y al sumar todos los elementos de la cuantía:

$$
\begin{aligned}
1 &= \sum_{x\in H} P(\underline{X}=\underline{x}) = \phi_0 \sum_{x\in H} exp(\sum x_i log\phi_i + \sum x_i x_j log\alpha_{ij}) \\
\Rightarrow \phi_0 &= \frac{1}{\sum_{x\in H} exp(\sum x_i log\phi_i + \sum x_i x_j log\alpha_{ij})} \\
\phi_0 &= \frac{1}{\mathbf{1}e^{H\phi}}
\end{aligned}
\tag{11}
$$

A modo de ejemplo se presenta el caso particular de $p = 2$. En dicho caso Γ y H adoptan la siguiente forma:

$$
\Gamma = \begin{vmatrix} log\phi_1 \\ log\phi_2 \\ log\alpha_{12} \end{vmatrix} \qquad H = \begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{vmatrix}
$$

De esta manera:

$$
\begin{aligned}
\phi_0 &= \frac{1}{\mathbf{1}e^{H\Gamma}} \\
\phi_0 &= \frac{1}{e^{<(0,0,0)\Gamma>}+e^{<(1,0,0)\Gamma>}+e^{<(0,1,0)\Gamma>}+e^{<(1,1,1)\Gamma>}} \\
\phi_0 &= \frac{1}{e^0+e^{log\phi_1}+e^{log\phi_2}+e^{log\phi_1+log\phi_2+log\alpha_{12}}} \\
\phi_0 &= \frac{1}{1+\phi_1+\phi_2+\phi_1\phi_2\alpha_{12}}
\end{aligned}
$$

tal como se vio en la ecuación (6).

En cuanto a la interpretación de los parámetros, ϕ_0 continúa interpretándose como la probabilidad de obtener el valor cero en todas las variables. En cuanto los parámetros ϕ_i y α_{ij} se interpretan como los *odds* y *odds ratio* condicionales a que el resto de las variables sean cero. Pese a que sería deseable que la interpretación de dichos coeficientes no fuese parcial, es fácil construir estimadores incondicionales a partir de los elementos del vector Γ.

El siguiente paso es definir las distribuciones condicionales y marginales de subconjuntos del vector X. Sin pérdida de generalidad se asumirá que se quiere obtener la distribución marginal del vector $X^M = (X_1, X_2, \ldots, X_M)$, la cual se obtendrá sumando sobre los 2^{p-M} valores posibles del vector $X^m = (X_{M+1}, \ldots, X_p)$, donde $m = p - M$.

$$
\begin{aligned}
P(X^M = x) &= \sum_{X_{M+1}=0}^{X_{M+1}=1} \cdots \sum_{X_p=0}^{X_p=1} \phi_0 \prod_{i=1}^{p} \phi_i^{x_i} \prod_{j=i+1}^{p} \alpha_{ij}^{x_i x_j} \\
&= \phi_0^* \prod_{i=1}^{M} \phi_i^{x_i} \prod_{j=i+1}^{M} \alpha_{ij}^{x_i x_j} F(x, \phi^m, \phi^{Mm}) \\
&= \phi_0^* \prod_{i=1}^{M} \phi_i^{*x_i} \prod_{j=i+1}^{M} \alpha_{ij}^{*\, x_i x_j}
\end{aligned}
$$

De aquí se puede concluir que todas las distribuciones marginales también pertenecen a la familia de distribuciones BM. En cuanto a $F(x, \phi^m)$, es una función que involucra a los elementos de x, a los intensidades $(\gamma^{(m)})$ correspondientes a las variables sobre las cuales se suma y a las intensidades (ϕ^M) que "vinculan" los elementos de X^M y X^m. Para la construcción de los parámetros marginales se utiliza el resultado anterior conjuntamente con la definiciones de *odd* y *odds ratio* marginales. La definición de las intensidades marginales ϕ_i^* en (13) es la siguiente:

$$
\phi_i^* = \phi_i \widetilde{\gamma}_i^{(p-M)} \tag{14}
$$

donde $\widetilde{\phi}_i^{(m)} = \frac{e^{H\phi^{(m)}}}{\mathbf{1}e^{H\phi^m}} \phi^{M(m)}$. La interpretación de estas intensidades marginales corresponde a una corrección de las intensidades originales, donde dicha corrección se construye como un promedio ponderado de las asociaciones $(\alpha^{M(m)})$ entre X_i y las variables contenidas en X^m, con ponderadores dados por las intensidades y asociaciones (α^m) de las variables sobre las cuales se sumó.
El caso de las asociaciones marginales en la ecuación (13) es similar:

$$
\alpha_{ij}^* = \alpha_{ij} \phi_i \phi_j \frac{\widetilde{\gamma}_{ij}^{(m)}}{\widetilde{\gamma}_i^{(m)} \widetilde{\gamma}_j^{(m)}} \tag{15}
$$

Las distribuciones condicionales son mas sencillas y se construyen a partir de la siguiente relación:

$$
P(X^M = x^M | X^{(m)} = x^{(m)}) = \frac{P(X^M = x^M, X^{(m)} = x^{(m)})}{P(X^{(m)} = x^{(m)})} \tag{16}
$$

donde el numerador no es otra cosa que la cuantía que ya se definió en (9) y el denominador corresponde a la marginal del vector X_p que se acaba de presentar. Finalmente la cuantía condicional es la siguiente:

$$\begin{aligned} P(X^M = x^M | X^m = x^m) &= \phi_{0|j} \prod_{i=1}^{M} \phi_{i|m}^{x_i} \prod_{j=i+1}^{M} \alpha_{ij}^{x_i x_j} \\ \phi_{0|j} &= \frac{1}{\mathbf{1}e^{H\phi_{M|m}}} \\ \phi_{i|m} &= \phi_i \prod_{x_j \in m} \alpha_{ij}^{x_j} \end{aligned} \tag{17}$$

Hay que tener en cuenta como el proceso de condicionar en los valores de las variables contenidas en X^m solo afecta las intensidades y no las asociaciones.

2.4. Estimación

Dado que la función de verosimilitud es no lineal en los parámetros, se opta por realizar la estimación de los parámetros del modelo BM mediante técnicas de optimización numérica. Para ello se define la función de log-verosimilitud de una muestra de n observaciones como:

$$\ell(\underline{x}|\underline{\phi}) = n \log(\phi_0) + \sum_{j=1}^{p} S_j \log \phi_i + \sum_{j=1}^{p} S_{jk} \log \alpha_{ik} \tag{18}$$

donde $S_j = \sum_{i=1}^{n} x_{ij}$ y $S_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik}$.

La maximización de esta función se lleva a cabo por algunos de los métodos iterativos comunmente utilizados. La mayoría de los mismos requiere del gradiente (o *score*) y la matriz Hessiana de la ecuación (18). Los elementos del primero (al que denotamos como $U(\underline{\phi})$) tienen la siguiente forma:

$$\begin{aligned} U_j(\underline{\gamma}) &= \frac{\partial \ell(\underline{x}|\underline{\phi})}{\partial \phi_j} = \frac{S_j}{\phi_j} - n \frac{\mathbf{1}e^{H_{(j)}\Gamma}}{\mathbf{1}e^{H\Gamma}} \\ U_{jk}(\underline{\gamma}) &= \frac{\partial \ell(\underline{x}|\underline{\phi})}{\partial \alpha_{jk}} = \frac{S_{jk}}{\alpha_{jk}} - n \frac{\mathbf{1}e^{H_{(jk)}\Gamma}}{\mathbf{1}e^{H\Gamma}} \end{aligned} \tag{19}$$

donde $H_{(j)}$ es la matriz compuesta por las filas de H que contienen unos en la j-ésima columna (correspondiente a ϕ_j), luego esta columna es reemplazada por un vector de ceros. El caso de $H_{(jk)}$ es análogo al anterior pero con la columna correspondiente a α_{jk} reemplazada por un vector de ceros.

3. Una aplicación a la salud oral

Una posible aplicación de esta distribución es en al análisis de la enfermedad periodontal. La enfermedad periodontal, es una de las enfermedades más prevalentes en Odontología, teniendo un peso muy importante en la carga mundial de enfermedades crónicas no trasmisibles, las que afectan al 40 % de la población mundial (Lorenzo et al., 2013b). Desde el punto de vista de la salud colectiva el estudio de su distribución, explicación, prevención y tratamiento deben abordarse integralmente y considerase en el contexto de la salud general de los colectivos humanos. Desde el punto de vista biológico, la enfermedad periodontal está asociada al biofilm, matriz de microorganismos (incluidos los patógenos en una baja proporción) adheridos a la superficie del diente que en condiciones normales, se encuentran en armonía con el huésped sano. Los signos asociados con esta patología son sangrado gingival, sarro, bolsa patológica, pérdida de inserción de los tejidos periodontales, pérdida ósea y movilidad dentaria. Los índices que pretenden dar cuenta de la enfermedad periodontal tienen limitaciones derivadas del número de signos involucrados así como de los instrumentos utilizados y la subjetividad del observador. A nivel internacional se habla de enfermedad periodontal, cuando existen bolsas periodontales iguales o mayores a 4 mm, la que se mide a través del índice CPI.

3.1. Datos de sangrado

A continuación se presenta la aplicación de la distribución BM para el análisis del sangrado peridontal que es uno de los componentes de la enfermedad periodontal. Los datos utilizados corresponde al "1er Relevamiento Nacional de Salud Bucal en población joven y adulta uruguaya" en el que se entrevistaron 1483 personas a las que se les relevó entre otras cosas la presencia/ausencia de sangrado (gingivitis) en los sextantes (Lorenzo et al., 2013a). Los sextantes refieren a una forma de dividir las distintas piezas dentales en función a como están ubicadas en la boca.

	nrounico	Depart	sexo	edad	salud	univ	sang1617	sang11	sang2627	sang31
1	2	PAY	M	40	privada	SI	sano	sano	sano	sano
2	5	PAY	M	21	privada	NO	sano	sano	sano	sano
4	10	PAY	M	21	publica	SI	sano	sano	presente	presente
5	11	PAY	F	66	publica	NO	presente	presente	presente	presente

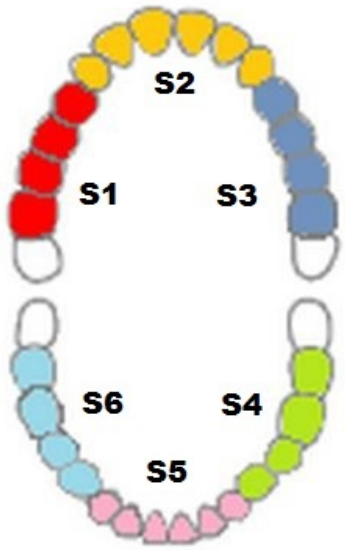

Figura 1: Distribución de los sextantes en la boca

	sang3637	sang4647	tramo_eta_rec	inse	sext1	sext2	sext3	sext4	sext5	sext6
1	sano	sano	de 35 a 44	41	0	0	0	0	0	0
2	sano	sano	de 15 a 24	36	0	0	0	0	0	0
4	sano	sano	de 15 a 24	62	0	0	1	0	1	0
5	presente	presente	de 65 a 74	19	1	1	1	1	1	1

Vemos como ejemplo 4 registros de la tabla de datos que muestran algunas características (sexo,edad,etc) y demás (en las últimas seis variables), como es el perfil individual de sangrado para los 6 sextantes.
En la figura 1 puede verse que hay sextantes vinculados al maxilar superior (sextantes 1, 2 y 3) e inferior (sextantes 4, 5 y 6) y a su vez si están en la parte derecha (sextantes 1 y 6) o izquierda (sextantes 3 y 4) de la boca.
En un análisis preliminar de estos datos se observó que gran parte de la muestra carecía de esta patología. Se consideró "sano" a un individuo con sus 6 sextantes sanos. Estos constituyen el 62 % de los datos, por lo tanto conformaron un perfil claro de individuos los cuales se dejaron de lado para trabajar sobre el resto, de modo de poder determinar distintos perfiles de carga de enfermedad.

	presencia	ausencia	%
S1	195	1288	13,1
S2	171	1312	11,5
S3	209	1274	14,1
S4	223	1260	15,0
S5	364	1119	24,5
S6	211	1272	14,2

Cuadro 1: Presencia de sangrado por sextantes

A partir de estos datos se van a estimar modelos donde se supone que no hay restricciones entre las relaciones de las 6 variables y luego modelos donde hay independencia local y homogeneidad local de las asociaciones.

3.2. Modelo estimado

Las subrutinas de cálculos fueron desarrolladas por el Prof. Ayu. Fernando Massa en el sistema R (R Core Team, 2014) usando, para la optimización los algoritmos de optimización no lineal implementados en la librería *nloptr* (Johnson, 2014) y que aparecen comentados por Ypma en el reporte técnico (Ypma, 2014).
A continuación se muestran las subrutinas de estimación creadas en **R** especialmente con los resultados de las estimaciones puntuales y por intervalo, por ejemplo para el modelo simple (sin restricciones).

```
y<-datos[,c(15:20)]
modelo1<-estim(y)
L0<-c(modelo1$intensidades,modelo1$asociaciones)
int.conf(modelo1,0.05)
repar(modelo1$intensidades,modelo1$asociaciones)
```

Vemos entonces los valores estimados $\hat{\phi}_i$ y $\hat{\alpha}_{ij}$ que retorna la función *estim*. Por otra parte, para una mejor interpretación de lo resultados, se reparametrizan los $\hat{\phi}_i$ y los $\hat{\alpha}_{ij}$ para ser presentados como *proporciones*,*Odds* y *OR*
Los intervalos de confianza para los parámetros que surgen del modelo (intensidades y asociaciones) se calculan utilizando la normalidad asintótica de los estimadores máximo verosímiles con la siguiente formulación:

$$[\phi - Z_{(1-\alpha/2)} * s.e; \phi + Z_{(1-\alpha/2)} * s.e] \tag{20}$$

donde *s.e.* es la raíz cuadrada de la varianza de cada parámetro del modelo, la que se estima para cada caso, a través de la descomposición QR de la hessiana asociada al modelo.

intensidades					
S1	S2	S3	S4	S5	S6
0.042	0.033	0.043	0.031	0.137	0.044
asociaciones					
-	4.00	7.39	1.77	1.65	1.64
	-	2.29	1.16	3.87	2.30
		-	2.70	2.08	2.03
			-	6.25	8.97
				-	1.95
					-

(a) Parámetros estimados

proporciones					
S1	S2	S3	S4	S5	S6
0.131	0.115	0.141	0.150	0.245	0.142
OR					
-	12.83	18.54	9.11	6.73	8.23
	-	10.7	8.4	9.8	9.07
		-	11.4	7.7	9.5
			-	14.5	21.8
				-	8.3
					-

(b) Reparametrización a proporciones y OR

```
> int.conf(modelo1,0.05)
-----------------------------------------------------
intervalos de confianza al 95% para las intensidades
-----------------------------------------------------
   int.inf   int   int.sup
1   0.031   0.042   0.054
2   0.023   0.033   0.043
3   0.031   0.043   0.055
4   0.022   0.032   0.042
5   0.113   0.137   0.161
6   0.033   0.045   0.057
-----------------------------------------------------
intervalos de confianza al 95% para las asociaciones
-----------------------------------------------------
      asoc.inf  asoc    asoc.sup
1-2    2.151    4.010    5.868
1-3    4.338    7.392   10.446
1-4    0.845    1.780    2.714
1-5    0.899    1.651    2.402
1-6    0.807    1.643    2.478
2-3    1.222    2.298    3.374
2-4    0.542    1.161    1.781
2-5    2.161    3.872    5.583
2-6    1.160    2.307    3.455
3-4    1.376    2.700    4.024
3-5    1.193    2.085    2.978
3-6    1.050    2.035    3.020
4-5    3.642    6.256    8.870
4-6    5.376    8.978   12.580
5-6    1.124    1.951    2.777
```

3.3. Discusión

Puede verse en este caso que según el modelo estimado, el sextante con mayor intensidad (parcial) es el $S5$ con un valor de $\hat{\phi}_5 = 0{,}137$ mientras que los sextantes que presentan mayor asociación (parcial) son el $S4, S6$ y el $S1, S3$ que son los sextantes posteriores inferiores y superiores respectivamente, con valores de $\hat{\alpha}_{4,6} = 8{,}97$ y $\hat{\alpha}_{1,3} = 7{,}39$.
Si se opta por reducir el número de parámetros del modelo mediante restricciones de igualdad, surgen diferentes alternativas. Una posibilidad es el modelo de "independencia", en dicho caso se impone $\alpha_{ij} = 1 \forall i, j$, logrando así que solo se estimen las p intensidades del modelo. Otro caso donde se simplifica la dimensionalidad del modelo es el caso de "homogeneidad", en este caso se asume $\alpha_{ij} = \alpha kl$ de modo que se estimen p intensidades y una sola asociación, común a todos los pares de sextantes. Utilizando una prueba de cociente de verosimilitud, se pudieron contrastar las hipótesis de estos modelos. A continuación se presentan las líneas de código:

```
# para testear la hipotesis de asociaciones=1 (independencia)
 1-pchisq(-2*(modelo1.indep$Logv-modelo1$Logv),df=p*(p-1)/2)
# para testear la hipotesis de asociaciones iguales (homogeneidad?)
 1-pchisq(-2*(modelo1.rest$Logv-modelo1$Logv),df=p*(p-1)/2-1)
```

Para el caso de la independencia entre los sextantes, se pudo rechazar la independencia ya que el valor del estadístico (cuya distribución era χ^2_{15}) arrojó un valor $p = 1{,}37e^{-8}$. Para el caso del modelo de homogeneidad de asociaciones, el estadístico de prueba tiene un grado de libertad menos debido a que se estima un parámetro de asociación. En este caso el p-valor fue de $3.4e^{-10}$, rechazándo así, que todas las asociaciones fuesen iguales a un único valor desconodico. Por lo tanto en ambos casos se rechazan la independencia y la homogeneidad de asociaciones.
En última instancia, retomando que en el modelo sin restricciones se observó que las estimaciones de las asociaciones posteriores (sextantes S1-S3 y sextantes S4-S6) eran mucho mayores al resto, se decidió poner a prueba la siguiente hipótesis:

$$\alpha_{13} = \alpha_{46}$$

Para esto se estimó un nuevo modelo bajo esta restricción. Al comparar las verosimilitudes, el p-valor encontrado fue de 0.26, lo que sugirió que las asociaciones posteriores eran efectivamente, de la misma magnitud.

4. Conclusiones y futuros pasos

En este trabajo se presenta una metodología de análisis para varias variables binarias diferente a la que habitualmente se usa y que está basada en una descomposición de una distribución Bernouilli multivariada en términos que reflejan intensidades de cada variable y asociaciones entre estas.

Para el caso de una aplicación en salud oral se analizan las asociaciones entre sextantes en el sangrado.

1. Se descarta la hipótesis de que la presencia de sangrado es independiente entre algunos sextantes.

2. Se constata que la asociación de presencia de sangrado entre los sextantes posteriores no difiere entre mandíbula y maxilar.

A futuro se intentará establecer diferentes tipologías que den cuenta del gradiente de infección usando diferentes técnicas a ser combinadas con la distribución Bernoulli multivariada (BM) .

1. Creación de tipologías de sangrado gengival a través de variables latentes que indican la pertenencia a diferentes grupos usando el algoritmo (EM) .

2. Clustering jerárquicos sobre distancia de datos binarios (Jurasinski y Retzer, 2012).

3. Clustering a partir de particiones difusas mediante medidas de entropía:
(Álvarez et al., 2012),(Moustaki y Papageorgiou, 2004),(Tsekouras et al., 2005)

Por otra parte resta estudiar cómo hacer el proceso de estimación de los modelos al trabajar con valores faltantes. A su vez poder implementar el cálculo de los intervalos de confianza para las reparametrizaciones de los componentes del modelo (Odds, y OR), en donde la varianza debe ser estimada mediante simulación montecarlo.

A. Apéndice

Para el cálculo de la matriz Hessiana se calculan las derivadas parciales del *score* según ϕ_j y ϕ_{jk}, de este modo surgen varios casos:

- caso A) derivada segunda de las intensidades.
  $\frac{\partial^2 \ell(\underline{x}|\underline{\phi})}{\partial \phi_j^2} = \frac{\partial U_j(\phi)}{\partial \phi_j} = \frac{\partial}{\partial \phi_j}[\frac{S_j}{\phi_j} - n\frac{\mathbf{1}e^{H_{(j)}\log\phi}}{\mathbf{1}e^{H\log\phi}}] = n[\frac{\mathbf{1}e^{H_{(j)}\log\phi}}{\mathbf{1}e^{H\log\phi}}]^2 - \frac{S_j}{\phi_j^2}$
- caso B) derivadas cruzadas entre las intensidades.
  $\frac{\partial^2 \ell(\underline{x}|\underline{\phi})}{\partial \phi_j \phi_k} = \frac{\partial U_j(\phi)}{\partial \phi_k} = \frac{\partial}{\partial \phi_k}[\frac{S_j}{\phi_j} - n\frac{\mathbf{1}e^{H_{(j)}\log\phi}}{\mathbf{1}e^{H\log\phi}}] = n[\frac{\mathbf{1}e^{H_{(j)}\log\phi}\mathbf{1}e^{H_{(k)}\log\phi} - \mathbf{1}e^{H_{(j)}^{(k)}\log\phi} \not\in e^{H\log\phi}}{[\mathbf{1}e^{H\log\phi}]^2}]$
- caso C) derivadas cruzadas entre intensidades y asociaciones.
  $\frac{\partial^2 \ell(\underline{x}|\underline{\phi})}{\partial \phi_j \phi_{kl}} = \frac{\partial U_j(\phi)}{\partial \phi_{kl}} = \frac{\partial}{\partial \phi_{kl}}[\frac{S_j}{\phi_j} - n\frac{\mathbf{1}e^{H_{(j)}\log\phi}}{\mathbf{1}e^{H\log\phi}}] = n[\frac{\mathbf{1}e^{H_{(j)}\log\phi}\mathbf{1}e^{H_{(kl)}\log\phi} - \mathbf{f}e^{H\log\phi}\mathbf{1}e^{H_{(j)}^{(kl)}\log\phi}}{[\mathbf{1}e^{H\log\phi}]^2}]$
- caso D) derivadas segundas de las asociaciones.
  $\frac{\partial^2 \ell(\underline{x}|\underline{\phi})}{\partial \phi_{jk}^2} = \frac{\partial U_{jk}(\phi)}{\partial \phi_{jk}} = \frac{\partial}{\partial \phi_{jk}}[\frac{S_{jk}}{\phi_{jk}} - n\frac{\mathbf{1}e^{H_{(jk)}\log\phi}}{\mathbf{1}e^{H\log\phi}}] = n[\frac{\mathbf{1}e^{H_{(jk)}\log\phi}}{\mathbf{1}e^{H\log\phi}}]^2 - \frac{S_{jk}}{\phi_{jk}^2}$
- caso E) derivadas cruzadas entre las asociaciones.
  $\frac{\partial^2 \ell(\underline{x}|\underline{\phi})}{\partial \phi_{jk} \phi_{lm}} = \frac{\partial U_{jk}(\phi)}{\partial \phi_{lm}} = \frac{\partial}{\partial \phi_{lm}}[\frac{S_{jk}}{\phi_{jk}} - n\frac{\mathbf{1}e^{H_{(jk)}\log\phi}}{\mathbf{1}e^{H\log\phi}}] = n[\frac{\mathbf{1}e^{H_{(jk)}\log\phi}\mathbf{1}e^{H_{(lm)}\log\phi} - \mathbf{1}e^{H_{(jk)}^{(lm)}\log\phi}\mathbf{1}e^{H\log\phi}}{[\mathbf{1}e^{H\log\phi}]^2}]$

Donde:

- $H_{(j)}^{(k)}$ se construye reteniendo las filas de $H_{(j)}$ que tienen unos en la k-ésima columna (correspondiente a ϕ_k) y luego se reemplazan dichos unos por ceros.
- $H_{(j)}^{(kl)}$ se construye reteniendo las filas de $H_{(j)}$ que tienen unos en la columna correspondiente a ϕ_{kl}, esta columna es luego reemplazada por un vector de ceros.
- $H_{(jk)}^{(lm)}$ se construye reteniendo las filas de $H_{(jk)}$ que tienen unos en la columna correspondiente a ϕ_{lm}, esta columna es luego reemplazada por un vector de ceros.

Esta matriz también suele utilizarse como una aproximación a la matriz de covarianzas de los estimadores máximo verosímiles. En eso caso, se utliza la matriz de información esperada de Fisher, calculada como el inverso, del opuesto del valor esperado de la matriz Hessiana. Para el cálculo de la misma, es necesario calcular la esperanza de los términos S_j y S_{jk} que aparecen en los puntos A y D.

- $\mathbf{E}(S_j) = \mathbf{E}(\sum_{i=1}^{n} X_{ij}) = n\mathbf{E}(X_{ij}) = n\mathbf{P}(X_{ij} = 1) = n\phi_0^*\phi_1^*$
- $\mathbf{E}(S_{jk}) = \mathbf{E}(\sum_{i=1}^{n} X_{ij}X_{ik}) = n\mathbf{E}(X_{ij}X_{ik}) = n\mathbf{P}(X_{ij}X_{ik} = 1) = n\phi_0^*\phi_1^*\phi_2^*\phi_{12}^*$

Referencias

Álvarez, F., Alvarez Vaz, R., y Massa, F. (2012). Determinación de tipologías de infecciones parasitarias intestinales, en escolares mediante, técnicas de clustering sobre datos binarios. En *CLATSE 2012*. Congreso Latinoamericano de Sociedades de Estadística.

Cox, D. R. y Wermuth, N. (2002). On some models for multivariate binary variables parallel in complexity with the multivariate gaussian distribution. *Biometrika*, 89:462–469.

Dai, B. (2012). Multivariate Bernoulli Distribution Models. Technical Report 1171, Department of Statistics, University of Wisconsin, Madison WI, 1300 University Ave.

Johnson, S. G. (2014). *The NLopt nonlinear-optimization package*. Rpackage version 1.0.4.

Jurasinski, G. y Retzer, V. (2012). *simba: A Collection of functions for similarity analysis of vegetation data*. R package version 0.3-4.

Lorenzo, S., Alvarez Vaz, R., Blanco, S., y Peres, M. (2013a). Primer Relevamiento Nacional de Salud Bucal en población joven y adulta uruguaya: Aspectos metodológicos. *Odontoestomatología*, 15:8 – 25.

Lorenzo, S., Piccardo, V., Alvarez, F., Massa, F., y Alvarez Vaz, R. (2013b). Enfermedad Periodontal en la población joven y adulta uruguaya del Interior del pais: Relevamiento Nacional 2010-2011. *Odontoestomatología*, 15:35 – 46.

Marshall, A. y Olkin, I. (1985). A family of bivariate distribution generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association*, 80:332–338.

Moustaki, I. y Papageorgiou, I. (2004). Latent class models for mixed variables with applications in archaeometry. *Elsevier Computational Statistics & Data Analysis*, page 17.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tamhane, A. C., Qiu, D., y Ankenman, B. E. (2009). A parametric mixture model for clustering multivariate binary data. *Wiley InterScience*, pages 3–19.

Tsekouras, G., Papageorgiou, D., Kotsiantis, S., Kalloniatis, C., y Pintelas, P. (2005). Fuzzy clustering of categorical attributes y its use in analyzing cultural data. *World Academy of Science, Engineering y Technology*, 1:87–91.

Ypma, J. (2014). Introduction to NLOPTR: an R interface to nloptr. Technical report.