Modelo Lineal Generalizado

Trabajo Práctico 1

1. Sea $\mathbf{X} = (X_1, \ldots, X_N)' \sim M(n, \pi_1, \ldots, \pi_N)$.

   a) Si $Z = X_1 + X_2$, entonces

   $$X^* = (Z, X_3, \ldots, X_N)' \sim M(n, \pi_1 + \pi_2, \pi_3, \ldots, \pi_N)$$

   Es decir, si se colapsa una multinomial sumando celdas la distribución sigue siendo multinomial.

   b) Dado que $Z = z$ los subvectores $(X_1, X_2)'$ y $(X_3, \ldots, X_N)'$ son condicionalmente independientes y multinomiales:

   $$\begin{aligned} (X_1, X_2)' &\sim M(z, \frac{\pi_1}{\pi_1 + \pi_2}, \frac{\pi_2}{\pi_1 + \pi_2}) \\ (X_3, \ldots, X_N)' &\sim M(n - z, \frac{\pi_3}{\pi_3 + \ldots + \pi_N}, \ldots, \frac{\pi_N}{\pi_3 + \ldots + \pi_N}) \end{aligned}$$

2. Si $X_1, \ldots, X_N$ son variables independientes tales que $X_j \sim \mathcal{P}(\lambda_j)$, entonces

   $$(X_1, \ldots, X_N)'|_{\sum_{j=1}^N X_j = n} \sim M(n, \pi_1, \ldots, \pi_N)$$

   donde

   $$\pi_j = \frac{\lambda_j}{\lambda_1 + \ldots + \lambda_N}$$

   Por lo tanto, la distribución de $X_1, \ldots, X_N$ puede ser factorizada en el producto de

   $$n = \sum_{j=1}^N X_j \sim \mathcal{P}(\sum_{j=1}^N \lambda_j)$$

   y

   $$(X_1, \ldots, X_N)'|_{n=n^*} \sim M(n^*, \pi_1, \ldots, \pi_N)$$

3. Sea $\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22})'$ un vector con distribución $M(n, \boldsymbol{\pi})$, donde $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})'$ que corresponde a una tabla de 2×2. Bajo el supuesto de **simetría** $\boldsymbol{\pi}$ puede expresarse como:

   $$\mathbf{\Pi} = \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \\ \beta \\ 1 - \alpha - 2\beta \end{bmatrix},$$

con $\theta = (\alpha, \beta)'$. Probar que los EMV bajo este modelo son

$$\begin{aligned} \widehat{\alpha} &= \frac{X_{11}}{n} \\ \widehat{\beta} &= \frac{X_{12} + X_{21}}{2n}. \end{aligned}$$

4. Sea $\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22})'$ un vector con distribución $M(n, \boldsymbol{\pi})$, donde $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})'$ que corresponde a una tabla de 2×2. Bajo el supuesto de **simetría** e **independencia** $\boldsymbol{\pi}$ puede expresarse como:

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} \theta \\ \theta(1-\theta) \\ \theta(1-\theta) \\ (1-\theta)^2 \end{bmatrix},$$

con $\theta \in (0, 1)$. Hallar el EMV de $\boldsymbol{\pi}$ bajo este modelo.

5. Everitt (1977) considera una muestra de 97 niños escolarizados de 10 años que fueron clasificados según los siguientes factores:
Conducta en Clase (X), *Riesgo Hogareño (Y)* y *Adversidad Escolar (Z)*. La Conducta en Clase fue clasificada por los maestros como descarriada o no descarriada, el Riesgo Hogareño como riesgo (R) o no riesgo (N) y Adversidad Escolar en baja, media o alta. Los datos se hallan en la siguiente tabla.

		Adversidad Escolar						
		Bajo		Medio		Alto		
	Riesgo	N	R	N	R	N	R	Total
Conducta	No Descarriado	16	7	15	34	5	3	80
en Clase	Descarriado	1	1	3	8	1	3	17
	Total	17	8	18	42	6	6	97

Cuadro 1: Valores Observados

a) En forma genérica llamemos n_{ijk} a la frecuencia observada en la casilla ijk, es decir en el nivel i de la variable genérica X, j de la variable genérica Y y k de la variable genérica Z, siendo $1 \leq i \leq I$, $1 \leq j \leq J$ y $1 \leq k \leq K$. De manera que el vector $(n_{111}, \ldots, n_{IJK})$ tiene distribución multinomial $M(n, \pi_{111}, \ldots, \pi_{IJK})$.Deduzca los estimadores de máxima verosimiltud de $(\pi_{111}, \ldots, \pi_{IJK})$ bajo el modelo saturado.

b) Deduzca los estimadores de máxima verosimilitud de $(\pi_{111}, \ldots, \pi_{IJK})$ en el modelo de independencia.

c) Puede ocurrir que una variable$X$ sea independiente de las otras dos, de manera que sea razonable el modelo

$$\pi_{ijk} = \pi_{i++}\, \pi_{+jk}\, \forall i, j, k \tag{1}$$

Halle los estimadores de π_{ijk} bajo el modelo (1).

d) Estime las π_{ijk} usando el modelo de independencia y los estimadores hallados en el ítem anterior a partir de la taba dada.

6. Como parte de un estudio longitudinal una muestra de 3182 individuos sin enfermedad cardiovascular fueron seguidos durante cuatro años y medio. Un total de 2121 individuos que no practicaban ningún tipo de ejercicio, no desarrollaron ninguna enfermedad cardiovascular durante el tiempo que duró el estudio. Estos 2121 individuos fueron clasificados según tres factores: *tipo de personalidad* (A o B), *nivel de colesterol* (alto o bajo) y *presión sistólica* (alta o normal). Las personas de tipo A muestra signos de stress, preocupación e hiperactividad, mientras que las tipo B son relajadas, despreocupadas y de actividad normal.

   Se obtuvo la siguiente tabla:

		Presión	
Personalidad	Colesterol	Normal	Alta
A	Normal	716	79
	Alto	207	25
B	Normal	819	67
	Alto	186	22

Cuadro 2: Valores Observados

a) Llamemos n_{ijk} a la frecuencia observada en la casilla ijk, es decir en el nivel i de la variable genérica X, j de la variable genérica Y y k de la variable genérica Z, siendo $1 \leq i \leq I$, $1 \leq j \leq J$ y $1 \leq k \leq K$. De manera que el vector $(n_{111}, \ldots, n_{IJK})$ tiene distribución multinomial $M(n, \pi_{111}, \ldots, \pi_{IJK})$.

Cuando dado un nivel de una variable las otras dos son independientes, estamos ante un **modelo de independencia condicional**. Sabemos que

$$P(X = i, Y = j | Z = k) = \frac{\pi_{ijk}}{\pi_{++k}} .$$

La independencia condicional entre X e Y significa que para todo i, j y k

$$P(X = i, Y = j | Z = k) = \frac{\pi_{i+k}}{\pi_{++k}} \frac{\pi_{+jk}}{\pi_{++k}} .$$

Por lo tanto, podemos expresar el **modelo de independencia condicional** como

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} .$$

Halle los estimadores de π_{ijk} bajo el modelo de independencia condicional.

b) Estime las π_{ijk} usando el modelo de independencia y los estimadores hallados en el ítem anterior a partir de la taba dada, tomando como Z la variable *Personalidad*.