UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

ANÁLISIS BAYESIANO DE DATOS CATEGÓRICOS

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRA EN CIENCIAS MATEMÁTICAS

P R E S E N T A

LIZBETH NARANJO ALBARRÁN

DIRECTOR DE TESIS: DR. EDUARDO ARTURO GUTIÉRREZ PEÑA

MÉXICO, D.F. ENERO, 2009

Índice general

Introducción

Motivación

En general, al analizar datos, algunas veces es de interés tomar en cuenta la información inicial disponible que se tenga sobre las cantidades bajo estudio. El enfoque Bayesiano de la inferencia estadística permite hacer esto basándose en la interpretación subjetiva de la probabilidad. Dada una distribución inicial (típicamente un modelo paramétrico) que describe la información inicial acerca del valor de una cantidad desconocida, el teorema de Bayes (sobre el cual se basa la estadística Bayesiana) permite actualizar esa información con la de los datos observados. La distribución final resultante resume la información disponible de las cantidades de interés, las cuales están condicionadas al modelo planteado.

El análisis de datos categóricos se ha basado principalmente en el enfoque de la estadística Clásica, haciendo sus inferencias con aproximaciones basadas en la teoría asintótica. Sin embargo, como en muchas áreas en donde se aplica la estadística, el enfoque Bayesiano ha venido desarrollándose en los últimos años. Esto debido al mejoramiento de herramientas y programas computacionales eficientes, que han hecho posible la solución de problemas más complejos.

El interés principal de este texto es el análisis Bayesiano de datos categóricos. Además se presenta el análisis de las tablas de contingencia y su modelación utilizando modelos loglineales, así como el análisis de series de tiempo de datos discretos.

Descripción de la Tesis

El texto consta de cinco capítulos. El primer capítulo pretende introducir al lector a los conceptos básicos de estadística (Bayesiana), tales como análisis conjugado, predicción, intercambiabilidad, suficiencia y familias exponenciales. Además se exponen métodos de simulación y una breve explicación del *software* WinBUGS que permite la simulación de modelos Bayesianos. En el

segundo capítulo se estudian los datos categóricos desde el punto de vista de la estadística Clásica, enfocándose en el análisis de tablas de contingencia, sus propiedades y su modelación a través de modelos loglineales. El tercer capítulo analiza los datos categóricos pero ahora utilizando las herramientas de estadística Bayesiana. Se estudian los modelos loglineales Bayesianos y brevemente se resumen otros tipos de modelos que pueden ser de mucha utilidad al analizar datos categóricos. En el cuarto capítulo se muestran otros modelos utilizados para variables discretas, una aplicación de las pruebas estadísticas Bayesianas en tablas de contingencia y principalmente, un análisis de datos discretos de series de tiempo. Finalmente, en el capítulo 5 se presentan algunas conclusiones y se discuten posibles temas para trabajo futuro.

Capítulo 1

Preliminares

Este capítulo es una breve introducción a los conceptos básicos de la estadística Bayesiana, así como a uno de los *software* estadísticos de gran utilidad para la simulación de modelos Bayesianos y en particular para el estudio de datos categóricos.

En la sección 1.1 estudiamos los conceptos básicos de la estadística Bayesiana. En la sección 1.2 damos una breve descripción de WinBUGS, *software* para el análisis Bayesiano de modelos estadísticos a través de métodos de Monte Carlo vía cadenas de Markov. Finalmente, en la sección 1.3 estudiamos las propiedades de las estadísticas suficientes y las familias exponenciales.

1.1. Inferencia Bayesiana

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución desconocida F. En Estadística es común considerar una familia paramétrica de densidades,

$$\mathcal{P} = \{p(x|\theta) : \theta \in \Theta\},$$

y proceder como si la distribución F correspondiera a alguno de los modelos en $\mathcal{P}$. De esta manera, el problema se reduce a hacer inferencias sobre el supuesto valor del parámetro θ que corresponde al "modelo verdadero". Desde el punto de vista Bayesiano, la información previa sobre el valor desconocido de θ se describe a través de una *distribución inicial* $p(\theta)$. El teorema de Bayes,

$$p(\theta|x_1, \ldots, x_n) = \frac{p(\theta)p(x_1, \ldots, x_n|\theta)}{\int p(\theta)p(x_1, \ldots, x_n|\theta)d\theta},$$

permite entonces incorporar la información contenida en la muestra, produciendo una descripción de la incertidumbre sobre el valor del parámetro a través de la *distribución final* $p(\theta|x_1, \ldots, x_n)$.

Es común escribir el teorema de Bayes como

$$p(\theta|x_1,\ldots,x_n) \propto p(\theta)\mathcal{L}(\theta|x_1,\ldots,x_n),$$

donde

$$\mathcal{L}(\theta|x_1,\ldots,x_n) \propto p(x_1,\ldots,x_n|\theta),$$

es la función de verosimilitud.

1.1.1. Análisis Conjugado

Tanto $p(\theta)$ como $p(\theta|x_1,\ldots,x_n)$ son distribuciones de probabilidad sobre el parámetro θ. La primera distribución sólo describe la información inicial y la segunda distribución actualiza dicha información usando también la información muestral que se pueda obtener. Resulta conveniente tanto para el análisis como desde el punto de vista computacional, que $p(\theta)$ y $p(\theta|x_1,\ldots,x_n)$ pertenezcan a la misma familia paramétrica.

Definición 1.1.1. *Sea $\mathcal{P} = \{p(x_1,\ldots,x_n|\theta) : \theta \in \Theta\}$ una familia paramétrica. Una clase (o colección) de distribuciones de probabilidad $\mathcal{F}$ es una familia conjugada para $\mathcal{P}$ si para todo $p(x_1,\ldots,x_n|\theta) \in \mathcal{P}$ y $p(\theta) \in \mathcal{F}$ se cumple que $p(\theta|x_1,\ldots,x_n) \in \mathcal{F}$.*

Para garantizar que $p(\theta)$ y $p(\theta|x_1,\ldots,x_n)$ pertenezcan a la misma familia general de funciones de distribución, se elige a $p(\theta)$ de tal manera que tenga la misma "estructura"de $p(x_1,\ldots,x_n|\theta)$ vista como una función de θ.

A continuación se estudiarán casos particulares de familias conjugadas de las distribuciones más usadas en el análisis de datos categóricos y tablas de contingencia.

Datos Binomiales

Muchas aplicaciones hacen referencia a un número fijo n de observaciones binarias. Sean $y_1,\ldots,y_n$ los resultados de n ensayos independientes e idénticos tal que $p(Y_i = 1) = \pi$ y $p(Y_i = 0) = 1 - \pi$. Generalmente se etiquetan como "éxitos" y "fracasos" los resultados 1 y 0 respectivamente. Las variables aleatorias $\{Y_i\}$ independientes e idénticamente distribuidas se conocen como variables aleatorias Bernoulli. El número total de éxitos, $Y = \sum_{i=1}^{n} Y_i$, tiene una distribución binomial con índice n y parámetro π, y se denota por $Bin(y|n,\pi)$. La función de masa de probabilidad de Y es

$$p(y|\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \qquad y = 0,1,\ldots,n,\ 0 < \pi < 1,$$

donde el coeficiente binomial es

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}.$$

La verosimilitud es

$$\mathcal{L}(\pi|y) \propto \pi^y(1-\pi)^{n-y}.$$

Una distribución inicial conjugada para una distribución binomial es la distribución beta con parámetros α y β (ambos positivos), $\pi \sim Beta(\pi|\alpha,\beta)$, cuya función de densidad de probabilidad de π está dada por

$$p(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, \qquad 0<\pi<1,\ \alpha>0,\ \beta>0,$$

donde $\Gamma(\cdot)$ es la función gamma dada por

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

Note que una distribución inicial simétrica sobre π se obtiene haciendo $\alpha = \beta$ y cuando $\alpha = \beta = 1$ se reduce a una distribución inicial uniforme. La distribución final de π también es una distribución beta, con parámetros $\alpha + y$ y $\beta + n - y$, cuya función de densidad de probabilidad es

$$p(\pi|y,n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)}\pi^{\alpha+y-1}(1-\pi)^{\beta+n-y-1}, \qquad 0<\pi<1.$$

Suponga que $Y_1,\ldots,Y_G$ son variables aleatorias independientes con distribución $Bin(y_i|n_i,\pi)$ para $i = 1,\ldots,G$. Entonces, si la distribución inicial para π es $Beta(\pi|\alpha,\beta)$, con α y β conocidas, la distribución final de π es

$$\pi \sim Beta\left(\pi|\alpha+\sum_{i=1}^G y_i, \beta+\sum_{i=1}^G n_i - \sum_{i=1}^G y_i\right).$$

Datos Poisson

Algunas veces los datos de conteo no resultan de un número fijo de ensayos y su rango son los números enteros no negativos. En estos casos, y bajo ciertas condiciones, un modelo que puede utilizarse es la distribución Poisson. Sea Y una variable aleatoria Poisson con parámetro $\mu > 0$; entonces su función de masa de probabilidad es

$$p(Y=y) = e^{-\mu}\frac{\mu^y}{y!}, \qquad y = 0,1,\ldots,\ \mu>0.$$

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ son variables aleatorias con distribución Poisson, con media y varianza común μ. La distribución inicial de μ se puede tomar como una distribución gamma con parámetros α y β (ambos positivos), con media α/β, y cuya función de densidad de probabilidad está dada por

$$p(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu}, \qquad \mu > 0,\ \alpha > 0,\ \beta > 0.$$

La verosimilitud de la distribución Poisson es

$$\mathcal{L}(\mu|y_1, \ldots, y_n) = \mathcal{L}(\mu|\mathbf{y}) \propto \prod_{i=1}^{n} e^{-\mu}\mu^{y_i}.$$

La distribución final de μ es

$$\begin{aligned} p(\mu|\mathbf{y}) &\propto \left[\prod_{i=1}^{n} \exp(-\mu)\mu^{y_i}\right] \mu^{\alpha-1}\exp(-\beta\mu) \\ &= \mu^{\alpha+\sum_{i=1}^{n} y_i - 1}\exp[-\mu(\beta+n)], \end{aligned}$$

de tal manera que la distribución final para μ es también una distribución gamma, específicamente,

$$p(\mu|y) = Gamma\left(\mu|\alpha + \sum_{i=1}^{n} y_i, \beta + n\right).$$

Debido a que la distribución inicial y la final pertenecen a la misma familia de distribuciones, entonces la familia de distribuciones gamma es conjugada para la familia de distribuciones Poisson.

Datos Multinomiales

Algunos ensayos tienen más de dos posibles categorías. Suponga que cada uno de N ensayos idénticos e independientes pueden clasificarse en cualquiera de k categorías; en el caso binomial $k = 2$. Sea $y_{ij} = 1$ si el ensayo i se clasifica en la categoría j y sea $y_{ij} = 0$ en otro caso. Entonces $Y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ representa un ensayo multinomial con $\sum_j y_{ij} = 1$; por ejemplo, $(0, 0, 1, 0)$ denota una clasificación en la categoría 3 de cuatro posibles categorías. Note que y_{ik} es redundante, debido a que es linealmente dependiente de y_{ij} con $j = 1, \ldots, k-1$.

Sea $n_j = \sum_i y_{ij}$ el número de ensayos que se clasifican en la categoría j. Se dice que los conteos $(n_1, n_2, \ldots, n_k)$ tienen una distribución multinomial.

Sea $\pi_j = p(Y_{ij} = 1)$ la probabilidad de que se clasifique en la categoría j, con $j = 1, \ldots, k$, para cada ensayo; necesariamente $0 < \pi_j < 1 \; \forall j$, y $\sum_{j=1}^{k} \pi_j = 1$.

La función de probabilidad de la distribución multinomial está dada por

$$p(n_1, \ldots, n_k | \pi_1, \ldots, \pi_k, N) = N! \prod_{j=1}^{k} \frac{\pi_j^{n_j}}{n_j!}, \qquad n_j = 0, 1, \ldots, \; 0 < \pi_j < 1,$$

con $N = \sum_{j=1}^{k} n_j$ y $\sum_{j=1}^{k} \pi_j = 1$.

Note que si $n_j, \; j = 1, \ldots, k$, fueran variables independientes Poisson con medias $\mu_j, \; j = 1, \ldots, k$, entonces su distribución condicional, dada $N = \sum_{j=1}^{k} n_j$, sería multinomial con parámetro $\pi_j = \mu_j / \sum_{j=1}^{k} \mu_j$. La demostración es inmediata, dado que N tiene una distribución Poisson con media $\sum_{j=1}^{k} \mu_j$ y de aquí

$$\begin{aligned} p(n_1, \ldots, n_k | N) &= p(n_1, \ldots, n_k)/p(N) \\ &= \frac{e^{-\sum \mu_i} \prod_{j=1}^{k} (\mu^{n_j}/n_j!)}{e^{-\sum \mu_j} (\sum \mu_j)^N / N!} \\ &= N! \prod_{j=1}^{k} \frac{\pi_j^{n_j}}{n_j!}. \end{aligned} \tag{1.1}$$

Una distribución inicial conjugada para la distribución multinomial con vector de parámetros $(\pi_1, \ldots, \pi_k)$ puede ser la distribución Dirichlet, con función de probabilidad

$$p(\pi_1, \ldots, \pi_k | \alpha) = \frac{\Gamma(\alpha_*)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{l=1}^{k} \pi_l^{\alpha_l - 1}, \qquad 0 < \pi_l < 1, \; \alpha_j > 0,$$

con $\sum_{j=1}^{k} \pi_j = 1$, $\alpha = (\alpha_1, \ldots, \alpha_k)$ y $\alpha_* = \sum_{i=1}^{k} \alpha_i$.

La distribución está parametrizada por un vector $\alpha = (\alpha_1, \ldots, \alpha_k)$ tal que $E(\pi_i) = \alpha_i / \alpha_*$, $Var(\pi_i) = E(\pi_i)(1 - E(\pi_i))/(1 + \alpha_*)$ y $Cov(\pi_i, \pi_j) = -E(\pi_i)E(\pi_j)/(1 + \alpha_*)$. El valor de α_* se interpreta como el "tamaño de muestra inicial hipotético", y determina la cantidad de información contenida en la distribución inicial: una α_* pequeña implica información vaga mientras que una α_* grande indica una distribución inicial robusta para $(\pi_1, \ldots, \pi_k)$.

La distribución final de $(\pi_1, \ldots, \pi_k)$ es

$$\begin{aligned} p(\pi_1, \ldots, \pi_k | n_1, \ldots, n_k, \alpha, N) &\propto p(n_1, \ldots, n_k | \pi_1, \ldots, \pi_k, N) p(\pi_1, \ldots, \pi_k | \alpha) \\ &= N! \prod_{i=1}^{k} \frac{\pi_i^{n_i}}{n_i!} \frac{\Gamma(\alpha_*)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{l=1}^{k} \pi_l^{\alpha_l - 1} \\ &\propto \prod_{i=1}^{k} \pi_i^{n_i + \alpha_i - 1}. \end{aligned}$$

Este último es el kernel de una distribución Dirichlet con vector de parámetros $(n_1 + \alpha_1, \ldots, n_k + \alpha_k)$, por lo tanto esa es la distribución final de $(\pi_1, \ldots, \pi_k)$. Esta distribución contiene toda la información disponible sobre las probabilidades $(\pi_1, \ldots, \pi_k)$ de las celdas, condicional a las observaciones $(n_1, \ldots, n_k)$.

A falta de información inicial se usa una distribución inicial no informativa. Una de las distribuciones iniciales más usadas para parámetros multinomiales es precisamente la distribución de Dirichlet con vector de parámetros $\alpha = (1/2, \ldots, 1/2)$.

Teniendo en cuenta que α_* se interpretó como el tamaño de muestra inicial, la cantidad $I = \alpha_*/(N+\alpha_*)$ puede considerarse como la proporción de la información total que contribuye a la distribución inicial. De esta manera, un valor de α_* que permita obtener $I = 0.01$ produciría alrededor del 1 % de la información total, mientras que $I \approx 1$ implicaría que los datos están completamente dominados por la distribución inicial.

1.1.2. Predicción

En muchas ocasiones el propósito de un análisis estadístico es *predecir* el valor de una observación futura X con base en la información disponible. El problema de inferencia sobre θ puede considerarse como un paso intermedio en la solución al problema de predicción, aunque en ciertas situaciones puede ser de interés en sí mismo. Por otro lado, debido a resultados de consistencia, un parámetro puede verse como el límite de una sucesión de *estadísticas* (*i.e.* funciones de las observaciones) cuando el tamaño de la muestra tiende a infinito (ver teorema 1.1.2). De esta manera, hacer inferencias acerca del valor del parámetro θ puede considerarse como una forma límite de hacer inferencias predictivas acerca de las observaciones.

Dado el valor del parámetro θ, la distribución que describe el comportamiento de la observación futura x es $p(x|\theta)$; sin embargo, el valor de θ generalmente es desconocido. Algunos métodos estadísticos tradicionales atacan este problema *estimando* a θ con base en la muestra observada, y en muchos casos simplemente sustituyen el valor de θ con la estimación resultante.

Desde la perspectiva Bayesiana, el modelo paramétrico $p(x|\theta)$, junto con la distribución inicial $p(\theta)$, inducen una distribución conjunta para (X, θ), dada por

$$p(x, \theta) = p(x|\theta)p(\theta).$$

La distribución marginal

$$p(x) = \int p(x|\theta)p(\theta)d\theta,$$

describe nuestro conocimiento acerca de X dada la información inicial disponible. Dicha distribución se conoce comúnmente como la *distribución predictiva (inicial).*

De manera similar, una vez obtenida la muestra, el modelo $p(x|\theta)$ y la distribución final inducen una distribución conjunta para (X, θ) *condicional en los valores observados* $x_1, \ldots, x_n$; *i.e.*

$$p(x, \theta|x_1, \ldots, x_n) = p(x|\theta, x_1, \ldots, x_n)p(\theta|x_1, \ldots, x_n) = p(x|\theta)p(\theta|x_1, \ldots, x_n),$$

donde la última igualdad se da siempre y cuando haya independencia condicional de X y $(X_1, \ldots, X_n)$ dado θ. Así, la distribución

$$p(x|x_1, \ldots, x_n) = \int p(x|\theta)p(\theta|x_1, \ldots, x_n)d\theta,$$

describe el comportamiento de X dada toda la información disponible y se conoce como la *distribución predictiva (final).*

Ejemplo

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución Bernoulli con función de masa de probabilidad

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}, \qquad x \in \{0, 1\},\ 0 < \theta < 1,$$

y supongamos que θ tiene una distribución inicial Beta(α_0, β_0).

Como ya se mencionó, la distribución beta es conjugada para el modelo Bernoulli, por lo que la distribución final de θ también es beta. De hecho,

$$p(\theta|x_1, \ldots, x_n) = Beta(\theta|\alpha_1, \beta_1),$$

donde $\alpha_1 = \alpha_0 + \sum_{i=1}^n x_i$ y $\beta_1 = \beta_0 + n - \sum_{i=1}^n x_i$. La distribución predictiva final está dada por

$$p(x|x_1, \ldots, x_n) = \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(\alpha_1 + x)\Gamma(\beta_1 + 1 - x)}{\Gamma(\alpha_1 + \beta_1 + 1)}, \qquad x = 0, 1.$$

Esta distribución se conoce como *Beta-Binomial.* ■

Ejemplo

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución normal con función de densidad

$$p(x|\theta) = N(x|\theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\}, \qquad x \in \Re,\ \theta \in \Re,$$

con $\sigma > 0$ conocido. Supongamos que θ tiene una distribución inicial conjugada,

$$p(\theta) = N(\theta|\mu_0, \tau_0^2).$$

Entonces la distribución final de θ está dada por

$$p(\theta|x_1, \ldots, x_n) = N(\theta|\mu_1, \tau_1^2),$$

donde

$$\mu_1 = (1/\tau_0^2 + n/\sigma^2)^{-1}(\mu_0/\tau_0^2 + n\bar{x}/\sigma^2)$$

y

$$\tau_1^2 = (1/\tau_0^2 + n/\sigma^2)^{-1}.$$

La distribución predictiva final es entonces

$$p(x|x_1, \ldots, x_n) = N(x|\mu_1, \tau_*^2),$$

con

$$\tau_*^2 = \sigma^2\tau_1^2(1/\tau_1^2 + 1/\sigma^2).$$

■

1.1.3. Intercambiabilidad

Consideramos a $X_1, \ldots, X_n$ variables aleatorias, cuyo comportamiento se describe a través de la especificación de una distribución conjunta, digamos $p(x_1, \ldots, x_n)$. Esta distribución define de manera implícita otras especificaciones que pueden ser de gran interés. Por ejemplo, para $1 \leq m \leq n$,

$$p(x_1, \ldots, x_m) = \int p(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)dx_{m+1} \cdots dx_n,$$

es la distribución marginal de $(X_1, \ldots, X_m)$, mientras que

$$p(x_{m+1}, \ldots, x_n|x_1, \ldots, x_m) = \frac{p(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)}{p(x_1, \ldots, x_m)}$$

es la distribución condicional de las variables, $X_{m+1}, \ldots, X_n$, dados los datos $X_1 = x_1, \ldots, X_m = x_m$.

La especificación directa de $p(x_1, \ldots, x_n)$ puede llegar a ser muy difícil en la práctica. Es por eso que conviene examinar con cuidado el proceso de selección de una medida de probabilidad específica que describa adecuadamente nuestro conocimiento.

Definición 1.1.2. *Un modelo predictivo para una sucesión de variables aleatorias $X_1, X_2, \ldots$ es una medida de probabilidad P, la cual especifica la forma de la distribución conjunta que describe nuestro conocimiento acerca de cualquier subconjunto de la sucesión.*

Considere una sucesión de variables aleatorias $X_1, X_2, \ldots$, y supongamos un modelo predictivo que especifica, para todo $n \in \mathbb{N}$, una distribución conjunta de la forma

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i),$$

de manera que las variables aleatorias X_i son independientes. Claramente, para cualquier $1 \leq m \leq n$ se tiene entonces que

$$p(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m) = p(x_{m+1}, \ldots, x_n),$$

por lo que bajo este modelo las observaciones $x_1, \ldots, x_m$ no proporcionan información alguna sobre las observaciones futuras; no hay aprendizaje a partir de la experiencia dada.

Un modelo predictivo con una estructura de independencia es inapropiado en contextos donde creemos que los datos darán información acerca de eventos futuros. En estos casos, la densidad conjunta debe tener estructura de dependencia entre las variables aleatorias.

Veremos ahora una forma particular de juicio subjetivo acerca de ciertas estructuras simples de dependencia pero que pueden corresponder a juicios reales en situaciones de interés práctico.

Definición 1.1.3 (Intercambiabilidad finita)**.** *Se dice que las variables aleatorias $X_1, \ldots, X_n$ son (finitamente) intercambiables bajo una medida de probabilidad P si la distribución inducida por P satisface*

$$p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

para toda permutación π definida sobre el conjunto $\{1, 2, \ldots, n\}$.

En otras palabras, las "etiquetas"que identifican a cada una de las variables no proporcionan información alguna.

Es claro que si las variables $X_1, \ldots, X_n$ son independientes e idénticamente distribuidas entonces son intercambiables. El inverso no siempre es verdad.

Ejemplo

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ un vector aleatorio con distribución Normal multivariada $N_n(\mathbf{0}, \mathbf{\Sigma})$ y supongamos que los elementos de la diagonal de $\mathbf{\Sigma}$ son todos idénticos. Consideremos los siguientes dos casos:

i) La matriz $\mathbf{\Sigma}$ es diagonal. Entonces las variables aleatorias $X_1, \ldots, X_n$ son independientes e idénticamente distribuidas, y por lo tanto intercambiables.

ii) La matriz $\mathbf{\Sigma}$ *no* es diagonal. En este caso las variables aleatorias $X_1, \ldots,$ X_n no son independientes, pero siguen siendo intercambiables.

■

Definición 1.1.4 (Intercambiabilidad infinita)**.** *La sucesión infinita de variables aleatorias $X_1, X_2, \ldots$ es (infinitamente) intercambiable si toda subsucesión finita es intercambiable en el sentido de la definición 1.1.3.*

No toda colección finita de variables aleatorias intercambiables puede anidarse en una sucesión infinita de variables aleatorias intercambiables definidas de manera similar. Más aún, una colección finita de variables aleatorias intercambiables no necesariamente puede anidarse en una colección finita más grande de variables aleatorias intercambiables.

Ejemplo

Sean X_1, X_2 y X_3 variables aleatorias que toman valores en el conjunto $\{0, 1\}$ y con función de probabilidad conjunta dada por

$$\begin{aligned} P(X_1 = 0, X_2 = 1, X_3 = 1) &= P(X_1 = 1, X_2 = 0, X_3 = 1) \\ &= P(X_1 = 1, X_2 = 1, X_3 = 0) \\ &= 1/3, \end{aligned}$$

y donde cualquier otra combinación tiene probabilidad cero, de manera que X_1, X_2 y X_3 son intercambiables. Se puede mostrar que no existe una variable X_4 que tome valores en $\{0, 1\}$ y tal que X_1, X_2, X_3 y X_4 sean intercambiables. ■

La importancia del concepto de intercambiabilidad queda expresada en el teorema de Representación de De Finetti, el cual proporciona una representación de la función de probabilidad conjunta de n variables aleatorias de una sucesión infinita de variables aleatorias intercambiables.

Teorema 1.1.1 (Teorema de Representación de De Finetti)**.** *Si $X_1, X_2, \ldots$ es una sucesión infinita de variables aleatorias definidas sobre $\Re$ e intercambiables con respecto a la medida de probabilidad P, entonces existe una función de distribución Q definida sobre $\mathcal{F}$ (el espacio de todas las distribuciones sobre $\Re$) tal que la distribución conjunta de $X_1, \ldots, X_n$ tiene la forma*

$$P(x_1, \ldots, x_n) = \int_{\mathcal{F}} \left\{ \prod_{i=1}^{n} F(x_i) \right\} dQ(F),$$

donde $Q(F) = \lim_{n\to\infty} P(F_n)$, *y* F_n *es la función de distribución empírica de la muestra* $x_1, \ldots, x_n$.

A manera de ilustración, a continuación se presenta el caso más simple del Teorema de Representación.

Teorema 1.1.2. *Si* $X_1, X_2, \ldots$ *es una sucesión infinita de variables aleatorias definidas sobre* $\{0, 1\}$ *e intercambiables con respecto a la medida de probabilidad* P, *entonces existe una función de distribución* Q *tal que la función de probabilidad* $p(x_1, \ldots, x_n)$ *tiene la forma*

$$p(x_1, \ldots, x_n) = \int_0^1 \left\{ \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right\} dQ(\theta),$$

donde $Q(\theta) = \lim_{n\to\infty} P(Y_n/n \leq \theta)$, *con* $Y_n = X_1 + \cdots + X_n$, *y* $\theta = \lim_{n\to\infty} Y_n/n$ *(c.s.).*

Este teorema tiene un significado muy profundo desde el punto de vista de la modelación subjetiva. El teorema nos dice que el modelo predictivo para una sucesión intercambiable de variables aleatorias binarias puede describirse en términos de una situación en la que, condicional en el valor de una variable aleatoria θ, las variables aleatorias X_i se consideran independientes con distribución Bernoulli y a θ se le asigna una distribución de probabilidad Q.

Por la ley Fuerte de los Grandes Números, $\theta = \lim_{n\to\infty} Y_n/n$ (c.s.), de manera que la distribución de probabilidad Q puede interpretarse como una descripción de los *juicios acerca del límite de la frecuencia relativa de los "éxitos"en la sucesión de ensayos Bernoulli.*

Corolario 1.1.1. *Si* $X_1, X_2, \ldots$ *es una sucesión infinita de variables aleatorias definidas sobre* $\{0, 1\}$ *e intercambiables con respecto a la medida de probabilidad* P, *entonces la distribución condicional* $p(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m)$ *es de la forma*

$$p(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m) = \int_0^1 \left\{ \prod_{i=m+1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right\} dQ(\theta | x_1, \ldots, x_m)$$

donde $1 \leq m < n$,

$$dQ(\theta | x_1, \ldots, x_m) = \frac{\{\prod_{i=1}^{m} \theta^{x_i} (1-\theta)^{1-x_i}\} dQ(\theta)}{\int_0^1 \{\prod_{i=1}^{m} \theta^{x_i} (1-\theta)^{1-x_i}\} dQ(\theta)} \tag{1.2}$$

y $Q(\theta) = \lim_{n\to\infty} P(Y_n/n \leq \theta)$.

La expresión (1.2) no es más que una versión del Teorema de Bayes. Notemos que la forma de la representación no cambia. En la terminología usual, la distribución inicial $Q(\theta)$ ha sido actualizada a través del Teorema de Bayes, obteniéndose la distribución final $Q(\theta|x_1, \ldots, x_m)$.

La distribución predictiva (final) $p(x_{m+1}, \ldots, x_n|x_1, \ldots, x_m)$ nos permite derivar la correspondiente distribución predictiva de cualquier otra variable definida en términos de las observaciones futuras.

Para más detalles de esta sección ver Bernardo y Smith (1994) y Gutiérrez Peña (1998).

1.1.4. Monte Carlo vía Cadenas de Markov

Como se mencionó anteriormente, en el enfoque Bayesiano de la Estadística, la incertidumbre presente en un modelo dado, $p(x|\theta)$, se representa a través de una distribución de probabilidad $p(\theta)$ sobre el espacio parametral Θ (generalmente multidimensional) que define al modelo. El teorema de Bayes,

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)},$$

permite entonces incorporar la información contenida en un conjunto de datos $x = (x_1, \ldots, x_n)$, produciendo una descripción conjunta de la incertidumbre sobre los valores de los parámetros del modelo a través de la distribución final $p(\theta|x)$. Desafortunadamente, la implementación de los métodos Bayesianos ocasionalmente requieren de un esfuerzo computacional muy alto. La mayor parte de este esfuerzo se concentra en el cálculo de ciertas características de la distribución final del parámetro de interés. Además, en la práctica es común que la dimensión de θ sea muy grande. Por otro lado, excepto en aplicaciones muy sencillas tanto $p(x|\theta)$ como $p(\theta)$ pueden llegar a tener formas muy complicadas. En la mayoría de los problemas las integrales requeridas no pueden resolverse analíticamente, por lo que es necesario contar con métodos numéricos eficientes que permitan calcular o aproximar integrales en varias dimensiones. Algunos de estos métodos de integración son las técnicas de Monte Carlo vía cadenas de Markov. Para mayores detalles consultar Gilks, Richardson y Spiegelhalter (1996).

La integración de Monte Carlo evalúa $E[g(\theta)]$ obteniendo muestras denotadas por $\{\theta^{(t)},\ t = 1, \ldots, n\}$ de la distribución $p(\theta|x)$ y entonces aproximando

$$E[g(\theta)] \approx \frac{1}{n}\sum_{t=1}^{n} g(\theta^{(t)}).$$

De esta manera, la media poblacional de $g(\theta)$ se estima por medio de una

media muestral. Cuando las muestras $\{\theta^{(t)}\}$ son independientes, la ley de los grandes números asegura que la aproximación puede hacerse tan precisa como se desee incrementando el tamaño de la muestra, n. Note que aquí n está bajo el control del analista: no es el tamaño fijo para una muestra de datos.

En general, seleccionar muestras $\{\theta^{(t)}\}$ independientes de $p(\theta|x)$ no es factible, ya que $p(\theta|x)$ puede ser no estándar. Sin embargo, las $\{\theta^{(t)}\}$ no necesariamente necesitan ser independientes. Las $\{\theta^{(t)}\}$ pueden ser generadas por cualquier proceso que selecciona muestras por todo el soporte de $p(\theta|x)$ en las proporciones correctas. Una forma de hacer esto es a través de una cadena de Markov teniendo a $p(\theta|x)$ como su distribuión estacionaria. Esto es entonces un método de Monte Carlo vía cadenas de Markov.

En resumen, las técnicas de Monte Carlo vía cadenas de Markov permiten generar, de manera iterativa, observaciones de distribuciones multivariadas que difícilmente podrían simularse utilizando métodos directos. La idea básica es muy simple: construir una cadena de Markov que sea fácil de simular y cuya distribución de equilibrio corresponda a la distribución final que nos interesa.

Proposición 1.1.1. *Sea $\theta^{(1)}, \theta^{(2)}, \ldots$ una cadena de Markov homogénea, irreducible y aperiódica, con espacio de estados Θ y distribución de equilibrio $p(\theta|x)$. Entonces, conforme $t \to \infty$,*

$$(i) \quad \theta^{(t)} \xrightarrow{\mathcal{D}} \theta, \qquad \text{donde } \theta \sim p(\theta|x);$$

$$(ii) \quad \frac{1}{t}\sum_{i=1}^{t} g(\theta^{(i)}) \to E(g(\theta)|x).$$

Algoritmo de Metropolis-Hastings

Este algoritmo construye una cadena de Markov definiendo las probabilidades de transición de la siguiente manera.

Sea $Q(\theta^*|\theta)$ una densidad de transición (arbitraria) y definamos

$$\alpha(\theta^*, \theta) = \text{mín}\left\{\frac{p(\theta^*|x)Q(\theta|\theta^*)}{p(\theta|x)Q(\theta^*|\theta)}, 1\right\}.$$

Algoritmo:
Dado un valor inicial $\theta^{(0)}$, la t-ésima iteración consiste en:

1. generar una observación θ^* de $Q(\theta^*|\theta^{(t)})$;

2. generar una variable $u \sim U(0,1)$;

3. si $u \leq \alpha(\theta^*, \theta^{(t)})$, hacer $\theta^{(t+1)} = \theta^*$; en caso contrario, hacer $\theta^{(t+1)} = \theta^{(t)}$.

Este procedimiento genera una cadena de Markov con distribución de transición

$$\begin{aligned} p(\theta^{(t+1)}|\theta^{(t)}) &= \alpha(\theta^{(t)}, \theta^{(t+1)}) Q(\theta^{(t+1)}|\theta^{(t)}) I_{\{\theta^{(t+1)} \neq \theta^{(t)}\}} \\ &+ \left[1 - \int \alpha(\theta^{(t)}, \theta^*) Q(\theta^*|\theta^{(t)}) d\theta^*\right] I_{\{\theta^{(t+1)} = \theta^{(t)}\}}. \end{aligned}$$

La probabilidad de aceptación $\alpha(\theta^*, \theta)$ sólo depende de $p(\theta|x)$ a través de un cociente, por lo que la constante de normalización no es necesaria.

Comentario. La versión original del algoritmo de Metropolis-Hastings toma $Q(\theta^*|\theta) = Q(\theta|\theta^*)$, en cuyo caso

$$\alpha(\theta^*, \theta) = \text{mín}\left\{\frac{p(\theta^*|x)}{p(\theta|x)}, 1\right\}.$$

Dos casos particulares en la práctica son:

- *Caminata aleatoria.* Sea $Q(\theta^*|\theta) = Q_1(\theta^* - \theta)$, donde $Q_1(\cdot)$ es una densidad de probabilidad simétrica centrada en el origen. Entonces

$$\alpha(\theta^*, \theta) = \text{mín}\left\{\frac{p(\theta^*|x)}{p(\theta|x)}, 1\right\}.$$

- *Independencia.* Sea $Q(\theta^*|\theta) = Q_0(\theta^*)$, donde $Q_0(\cdot)$ es una densidad de probabilidad sobre Θ. Entonces

$$\alpha(\theta^*, \theta) = \text{mín}\left\{\frac{\omega(\theta^*)}{\omega(\theta)}, 1\right\},$$

con $\omega(\theta) = p(\theta|x)/Q_0(\theta)$.

En la práctica es común utilizar, después de una reparametrización apropiada, distribuciones de transición normales ó t de Student ligeramente sobredispersas, por ejemplo

$$Q(\theta^*|\theta) = N_d(\theta^*|\theta, \kappa V(\hat{\theta})) \qquad \text{(caminata aleatoria)}$$

ó

$$Q_0(\theta^*) = N_d(\theta^*|\hat{\theta}, \kappa V(\hat{\theta})) \qquad \text{(independencia)},$$

donde $\hat{\theta}$ y $V(\hat{\theta})$ denotan a la media y a la matriz de varianzas-covarianzas de la aproximación normal asintótica para $p(\theta|x)$, respectivamente, y $\kappa \geq 1$ es una factor de sobredispersión.

Ejemplo. (Coeficiente de correlación)
Suponga que $D = \{\mathbf{y}_i = (y_{1i}, y_{2i})^t, i = 1, \ldots, n\}$ es una muestra aleatoria de una distribución normal bivariada $N_2(0, \Sigma)$, donde

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Suponga una distribución inicial uniforme $U(-1, 1)$ para ρ; la densidad final para ρ está dada por

$$p(\rho|D) \propto (1-\rho^2)^{-n/2} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(S_{11} - 2\rho S_{12} + S_{22}\right)\right\},$$

donde $-1 < \rho < 1$, y $S_{rs} = \sum_{i=1}^{n} y_{ri} y_{si}$ para $r, s = 1, 2$. Generar valores de ρ de la densidad $p(\rho|D)$ no es algo trivial, ya que no es una función log-cóncava. Por tanto, consideramos el algoritmo de Metropolis-Hastings para generar ρ. Como $-1 < \rho < 1$, consideremos la transformación

$$\rho = \frac{-1 + e^{\xi}}{1 + e^{\xi}}, \qquad -\infty < \xi < \infty.$$

Entonces

$$p(\xi|D) = p(\rho(\xi)|D)\frac{2e^{\xi}}{(1+e^{\xi})^2}.$$

En vez de obtener un muestreo directo de ρ, generamos ξ eligiendo una distribución proporcional a $N(\hat{\xi}, \hat{\xi}^2_{\hat{\xi}})$, donde $\hat{\xi}$ es el máximo del logaritmo de $p(\xi|D)$, que puede obtenerse mediante el algoritmo de Newton-Raphson o el algoritmo de Nelder-Mead, y $\hat{\sigma}^2_{\hat{\xi}}$ es menos el inverso de la segunda derivada del $\log p(\xi|D)$ evaluada en $\xi = \hat{\xi}$, es decir,

$$\hat{\sigma}^{-2}_{\xi} = -\left.\frac{d^2 \log p(\xi|D)}{d\xi^2}\right|_{\xi=\hat{\xi}}.$$

El algoritmo para generar ξ opera de la manera siguiente:
Dado un valor inicial $\xi^{(0)}$,

1. generar ξ^* de $N(\hat{\xi}, \hat{\sigma}^2_{\hat{\xi}})$;

2. generar una variable $u \sim U(0, 1)$;

3. si $u \leq \alpha(\xi^*, \xi^{(t)})$, entonces $\xi^{(t+1)} = \xi^*$; en caso contrario, $\xi^{(t+1)} = \xi^{(t)}$ donde

$$\alpha(\xi^*, \xi^{(t)}) = \text{mín} \left\{ \frac{p(\xi^*|D)\phi\left(\frac{\xi^{(t)}-\hat{\xi}}{\hat{\sigma}_{\hat{\xi}}}\right)}{p(\xi^{(t)}|D)\phi\left(\frac{\xi^*-\hat{\xi}}{\hat{\sigma}_{\hat{\xi}}}\right)}, 1 \right\}$$

y ϕ es la función de densidad de probabilidad normal estándar.

Después de obtener ξ, podemos obtener ρ aplicando la transformación correspondiente. ■

Muestreo de Gibbs

Al igual que el algoritmo de Metropolis-Hastings, el algoritmo de Gibbs permite simular una cadena de Markov $\theta^{(1)}, \theta^{(2)}, \ldots$ con distribución de equilibrio $p(\theta|x)$. En este caso, sin embargo, cada nuevo valor de la cadena se puede obtener a través de un proceso iterativo que sólo requiere generar muestras de distribuciones cuya dimensión es menor que d y que en la mayoría de los casos tienen una forma más sencilla que la de $p(\theta|x)$.

Sea $\theta = (\theta_1, \ldots, \theta_k)$ una partición del vector θ, donde $\theta_i \in \Re^{d_i}$ y $\sum_{i=1}^{k} d_i = d$. Las densidades

$$\begin{gathered} p(\theta_1|\theta_2, \ldots, \theta_k, x) \\ \vdots \\ p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k, x) \quad (i = 2, \ldots, k-1) \\ \vdots \\ p(\theta_k|\theta_1, \ldots, \theta_{k-1}, x) \end{gathered}$$

se conocen como *densidades condicionales completas* y en general pueden identificarse fácilmente al inspeccionar la forma de la distribución final $p(\theta|x)$. De hecho, para cada $i = 1, \ldots, k$,

$$p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k, x) \propto p(\theta|x),$$

donde $p(\theta|x) = p(\theta_1, \ldots, \theta_k|x)$ es vista sólo como función de θ_i.

Algoritmo:

Dado un valor inicial $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$, el algoritmo de Gibbs simula una cadena de Markov en la que $\theta^{(t+1)}$ se obtiene a partir de $\theta^{(t)}$ de la siguiente manera:

generar una observación $\theta_1^{(t+1)}$ de $p(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}, x)$;

generar una observación $\theta_2^{(t+1)}$ de $p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}, x)$;

$$\vdots$$

generar una observación $\theta_k^{(t+1)}$ de $p(\theta_k|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)}, x)$.

La sucesión $\theta^{(1)}, \theta^{(2)}, \ldots$ así obtenida es entonces una realización de una cadena de Markov cuya distribución de transición está dada por

$$p(\theta^{(t+1)}|\theta^{(t)}) = \prod_{i=1}^{k} p(\theta_i^{(t+1)}|\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \ldots, \theta_k^{(t)}, x).$$

Comentario. En ocasiones la distribución final implica cierta estructura de independencia condicional entre algunos de los elementos del vector θ. En estos casos es común que muchas de las densidades condicionales completas se simplifiquen.

Ejemplo
Consideremos el modelo jerárquico definido por

$$\begin{aligned} \text{I.} \quad p(x|\omega) &= \prod_{i=1}^{m} p(x_i|\omega_i); \\ \text{II.} \quad p(\omega|\phi) &= \prod_{i=1}^{m} p(\omega_i|\phi); \\ \text{III.} \quad & p_0(\phi). \end{aligned}$$

Esta estructura define un modelo para x parametrizado por $\theta = (\omega, \phi) = (\omega_1, \ldots, \omega_m, \phi)$ y con distribución inicial $p(\theta) = p_0(\phi)p(\omega|\phi)$, de manera que la distribución final está dada por

$$p(\theta|x) \propto p_0(\phi) \prod_{i=1}^{m} \{p(x_i|\omega_i)p(\omega_i|\phi)\}.$$

Entonces $k = m + 1$ y las densidades condicionales completas toman la forma

$$\begin{aligned} p(\theta_1|\theta_2, \ldots, \theta_{k-1}, \theta_k, x) &= p(\omega_1|\phi, x_1) \\ &\vdots \\ p(\theta_{k-1}|\theta_1, \ldots, \theta_{k-2}, \theta_k, x) &= p(\omega_m|\phi, x_m) \\ p(\theta_k|\theta_1, \ldots, \theta_{k-2}, \theta_{k-1}, x) &= p_0(\phi|\omega), \end{aligned}$$

donde $p_0(\phi|\omega) \propto p_0(\phi)p(\omega|\phi)$. ■

Ejemplo. (Modelo normal bivariado)

El propósito de este ejemplo es examinar la estructura de la correlación exacta de una cadena de Markov inducida por el muestreo de Gibbs. Suponga que la distribución posterior $p(\theta|D)$ es una distribución normal bivariada $N_2(\mu, \Sigma)$ con

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad y \qquad \sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

donde μ_j, σ_j, $j = 1, 2$, y ρ son conocidos. Entonces el muestreo de Gibbs requiere muestrear de

$$\theta_1 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(\theta_2 - \mu_2), \sigma_1^2(1-\rho^2)\right)$$

y

$$\theta_2 \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_1 - \mu_1), \sigma_2^2(1-\rho^2)\right).$$

Sea $\{\theta_i = (\theta_{1,i}, \theta_{2,i})^t,\ i \geq 0\}$, que denota una cadena de Markov inducida por el muestreo de Gibbs para la distribución normal bivariada anterior. Si iniciamos a partir de la distribución estacionaria, es decir, $\theta_0 \sim N_2(\mu, \Sigma)$, entonces cada una de las sucesiones $\{\theta_{1,i},\ i \geq 0\}$ y $\{\theta_{2,i},\ i \geq 0\}$ es un proceso AR(1).

Veamos este resultado. Sea $\{z_{1,i}, z_{2,i},\ i \geq 0\}$ una sucesión de variables aleatorias i.i.d. $N(0, 1)$. Entonces la estructura del muestreo de Gibbs implica

$$\begin{aligned} \theta_{1,0} &= \mu_1 + \sigma_1 z_{1,0}, \\ \theta_{2,0} &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_{1,0} - \mu_1) + \sigma_2\sqrt{1-\rho^2}z_{2,0}, \end{aligned}$$

y

$$\begin{aligned} \theta_{1,i+1} &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(\theta_{2,i} - \mu_2) + \sigma_1\sqrt{1-\rho^2}z_{1,i+1}, \\ \theta_{2,i+1} &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_{1,i+1} - \mu_1) + \sigma_2\sqrt{1-\rho^2}z_{2,i+1}, \end{aligned} \tag{1.3}$$

para $i \geq 0$. Ahora consideremos el primer componente $\theta_{1,i+1}$. De la ecuación (1.3), para $i \geq 0$,

$$\begin{aligned} \theta_{1,i+1} &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left[\rho\frac{\sigma_2}{\sigma_1}(\theta_{1,i} - \mu_1) + \sigma_2\sqrt{1-\rho^2}z_{2,i}\right] + \sigma_1\sqrt{1-\rho^2}z_{1,i+1} \\ &= \mu_1 + \rho^2(\theta_{1,i} - \mu_1) + \rho\sigma_1\sqrt{1-\rho^2}z_{2,i} + \sigma_1\sqrt{1-\rho^2}z_{1,i+1}. \end{aligned} \tag{1.4}$$

Sea $\psi = \rho^2$ y $\sigma_1^{*2} = \sigma_1^2(1-\rho^4)$. Sea $\{z_i^*,\ i \geq 0\}$ que denota una sucesión de variables aleatorias i.i.d. $N(0,1)$. Como $z_{1,i}$ y $z_{2,i+1}$ son independientes e idénticamente distribuidas $N(0,1)$, entonces podemos reescribir (1.4) como

$$\theta_{1,0} = \mu_1 + \sigma_1 z_0^*, \tag{1.5}$$

$$\theta_{1,i+1} = \mu_1 + \psi(\theta_{1,i} - \mu_1) + \sigma_1^* z_{i+1}^* \qquad \forall i \geq 0. \tag{1.6}$$

Así, $\{\theta_{1,i},\ i \geq 0\}$ es un proceso AR(1) con rezago 1 y autocorrelación $\psi = \rho^2$. Similarmente, $\{\theta_{2,i},\ i \geq 0\}$ es también un proceso AR(1) con rezago 1 y autocorrelación $\psi = \rho^2$. La única diferencia es que usamos $\sigma_2^* = \sigma_2\sqrt{1-\rho^4}$ en vez de σ_1^* en (1.6), y usamos μ_2 y σ_2 en vez de μ_1 y σ_1 en (1.5). ■

Convergencia

Supongamos que se desea generar una muestra de tamaño N de la distribución $p(\theta|x)$. Si para cada uno de N valores iniciales $\theta_1^{(0)}, \ldots, \theta_N^{(0)}$ corremos alguno de los algoritmos discutidos en esta sección, entonces, de acuerdo con la proposición 1.1.1(i), después de un cierto número de iteraciones T suficientemente grande los valores $\theta_1^{(T)}, \ldots, \theta_N^T$ pueden considerarse como una muestra de tamaño N de la distribución final de θ. Alternativamente podemos generar una sola cadena y tomar los valores $\theta^{(T+K)}, \theta^{(T+2K)}, \ldots, \theta^{(T+NK)}$ como una muestra de $p(\theta|x)$, donde K se elige de manera que la correlación entre las observaciones sea pequeña.

En general no es fácil determinar en qué momento la(s) cadena(s) ha(n) convergido. Un método empírico comúnmente utilizado, basado en la proposición 1.1.1(ii), consiste en graficar los promedios ergódicos de algunas funciones de θ contra el número de iteraciones y elegir el valor T a partir del cual las gráficas se estabilizan. En este caso es frecuente omitir los primeros valores de la(s) cadena(s) al calcular los promedios ergódicos. La idea de este *periodo de calentamiento* es permitir que la(s) cadena(s) salga(n) de una primera fase de inestabilidad. En el caso particular del muestreo de Gibbs la velocidad de convergencia depende fuertemente de la correlación entre los componentes del vector θ bajo la distribución final $p(\theta|x)$: entre más alta sea la correlación más lenta será la convergencia.

1.2. WinBUGS

WinBUGS es un *software* de libre acceso en internet que se puede obtener de la página `http://www.mrc-bsu.cam.ac.uk/bugs/`. Fue diseñado por Spiegelhalter, Thomas y Best y es parte del proyecto BUGS (*Bayesian Inference Using Gibbs Sampling*) que desarrollaron estos investigadores para

el análisis Bayesiano de modelos estadísticos a través de métodos de Monte Carlo vía cadenas de Markov.

WinBUGS permite que métodos complejos de simulación sean sencillos para los usuarios de la estadística Bayesiana aplicada en diversas disciplinas.

El *software* ofrece una interfaz con el usuario basada en cuadros de diálogo y comandos a través de los cuales se analiza el modelo, por lo que el ambiente de WinBUGS se vuelve más amigable. Además, también es posible realizar una interfaz con R[1] o S-Plus, los cuales manejan una sintaxis muy similar a la que usa WinBUGS, aunque WinBUGS no cuenta con tantos comandos como R o S-Plus.

La manera de operar de WinBUGS está basada en el muestreo de Gibbs (ver sección 1.1.4); es decir, dada una función de verosimilitud y una distribución inicial, el propósito es muestrear valores de los parámetros del modelo a partir de la distribución final. De estos valores muestreados, es posible obtener estimadores de los parámetros y hacer todo tipo de inferencias sobre éstos.

El desarrollo y mejoramiento de las técnicas de Monte Carlo basadas en cadenas de Markov ha hecho posible construir la distribución final de modelos Bayesianos muy complejos. Sin embargo, generar una muestra a través del algoritmo de Gibbs para obtener la distribución final de un modelo Bayesiano puede ser una tarea muy complicada, sobre todo para el caso en el que se introducen muchos parámetros en el modelo.

WinBUGS proporciona herramientas simples para llevar a cabo la simulación Monte Carlo; a través de WinBUGS es posible implementar el muestreo de Gibbs para una gran variedad de modelos.

El *software* tiene ciertas limitaciones para detectar la convergencia, resumir las muestras y realizar diagnósticos sobre el ajuste de los modelos. Sin embargo, es posible usar los paquetes `boa` o `coda` de R junto con WinBUGS para realizar un mejor análisis de las muestras simuladas.

WinBUGS intenta usar el método de muestreo más apropiado para cada parámetro del modelo estocástico de acuerdo con la tabla 1.1.

Ejemplo

El siguiente ejemplo se obtuvo del libro de Peter Congdon (2001), *Bayesian Statistical Modelling*. Hay preocupación por el incremento de la incidencia de cáncer de seno en las mujeres, sin embargo, en el Reino Unido el número de muertes por esta enfermedad es más o menos constante. Ambas tendencias implican una disminución de "casos fatales", es decir, una disminución de la tasa de muerte entre los nuevos casos de la enfermedad.

[1] *Software* de libre acceso en la página de internet `http://www.r-project.org`.

Distribución Objetivo	Método de muestreo
Discreta	Inversión de una función de distribución acumulada
Forma cerrada	Muestreo directo usando algoritmos estándar
Log-cóncava	Muestreo de rechazo adaptativo
Rango restringido	Muestreo *slice*
Rango sin restricción	Metropolis-Hastings (sección 1.1.4)

Tabla 1.1: Métodos de muestreo de WinBUGS.

Suponga que estamos interesados en el intervalo de mayor densidad de esta tasa, en una situación donde tenemos $k = 11$ tipos de cáncer.

Sea $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ el número de muertes en cada uno de los 11 tipos de cáncer (1 es cáncer de seno). El vector aleatorio $\mathbf{n}$ tiene una distribución $Multinomial(n_1, \ldots, n_k|\pi, N)$, donde $N = \sum_{i=1}^{k} n_i$ es el número total de muertes por cáncer.

Para elegir la distribución inicial de π se podría tomar el patrón de muertes por cáncer de mujeres en el Reino Unido en años anteriores, o de algún otro país semejante a él.

En este ejemplo se considera una distribución inicial conjugada no informativa (ver sección 1.1.1) para π, la distribución $Dirichlet(\pi|1, 1, \ldots, 1)$.

El programa realizado en WinBUGS es el siguiente:

Modelo

```
model { # Modelo Multinomial
for(i in 1:k) {
c[i] <- 1
}
pi[1:k] ~ ddirch(c[]) # Inicial
n[1:k] ~ dmulti(pi[],N) # Verosimilitud
}
```

Datos

```
list(n = c(14080, 12990, 6440, 4350, 3420, 3190, 2600, 2420,
1820, 1760, 23610),
k = 11, # Tipos de Cáncer
N = 76680) # Total de Muertes por Cáncer
```

La distribución multinomial puede expresarse, equivalentemente, como la distribución conjunta de k variables aleatorias independientes con distribución Poisson, de tal manera que $n_i \sim Poisson(n_i|\mu_i)$ con $i = 1, \ldots, k$, sujetas a la condición $\sum_{j=1}^{k} n_j = N$ (ver sección 1.1.1). Las probabilidades

de la distribución multinomial se obtienen con:

$$\pi_j = \frac{\mu_j}{\sum_{i=1}^{k} \mu_i}.$$

Para este caso las distribuciones iniciales conjugadas para los parámetros μ_i son gamma, y utilizaremos distribuciones iniciales que son no informativas, $Gamma(\mu_i|1,1)$.

El programa utilizado en WinBUGS es el siguiente:

Modelo

```
model{ # Modelo Poisson
for(i in 1:k) {
mu[i] ~ dgamma(1,1) # Inicial
n[i] ~ dpois(mu[i]) # Verosimilitud
pi[i] <- mu[i]/mu.sum # Parámetro Multinomial
}
mu.sum <- sum(mu[ ])
}
```

Datos

```
list(n = c(14080, 12990, 6440, 4350, 3420, 3190, 2600, 2420,
1820, 1760, 23610),
k = 11) # Tipos de Cáncer
```

Las estadísticas básicas que se obtienen para la distribución final del vector de parámetros π son las mismas en ambos modelos y se presentan en la tabla 1.2.

■

1.3. Suficiencia y Familias Exponenciales

En esta sección se presenta una breve introducción al concepto de suficiencia y familias exponenciales, así como sus principales propiedades (ver Bernardo y Smith , 1994). Esto es de gran utilidad para el estudio de datos categóricos debido a que las distribuciones binomial, Poisson y multinomial pertenecen a la familia exponencial.

1.3.1. Estadística Suficiente

Una estadística suficiente es una función de los datos que resume toda la información de la muestra disponible referente al parámetro θ.

	mean	sd	MC error	2.5 %	median	97.5 %
π_1	0.1836	0.00139	$1.542E-5$	0.1809	0.1836	0.1863
π_2	0.1694	0.00135	$1.285E-5$	0.1667	0.1694	0.1721
π_3	0.0840	$9.939E-4$	$8.783E-6$	0.0820	0.0839	0.0859
π_4	0.0567	$8.385E-4$	$8.441E-6$	0.0551	0.0567	0.0584
π_5	0.0446	$7.422E-4$	$6.667E-6$	0.0431	0.0445	0.0460
π_6	0.0416	$7.225E-4$	$7.654E-6$	0.0402	0.0416	0.0430
π_7	0.0339	$6.555E-4$	$7.594E-6$	0.0326	0.0339	0.0351
π_8	0.0315	$6.333E-4$	$6.818E-6$	0.0303	0.0315	0.0328
π_9	0.0237	$5.534E-4$	$4.943E-6$	0.0226	0.0237	0.0248
π_{10}	0.0229	$5.457E-4$	$5.224E-6$	0.0219	0.0229	0.0240
π_{11}	0.3079	0.00168	$1.604E-5$	0.3047	0.3079	0.3113

Tabla 1.2: Estadísticas básicas de la distribución final de π.

Estadística

Comenzaremos con una definición formal, que nos permite discutir el proceso de resumir la información contenida en una sucesión, o *muestra*, de variables aleatorias, $X_1, \ldots, X_m$.

Definición 1.3.1 (Estadística)**.** *Dados los valores $x_1, \ldots, x_m$, de las cantidades aleatorias $X_1, \ldots, X_m$, un vector aleatorio $\mathbf{t}_m(x_1, \ldots, x_m)$, con $\mathbf{t}_m : \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \to \Re^{k(m)}$ $(k(m) \leq m)$, es una estadística $k(m)$-dimensional.*

Para reducir la dimensión de los datos claramente necesitamos que $k(m) < m$; como en los casos anteriores la dimensión de la estadística es fija e independiente del tamaño de muestra, es decir, $k(m) = k$.

En la siguiente sección estudiaremos estadísticas que nos permiten resumir la mayor información posible.

Suficiencia Predictiva y Suficiencia Paramétrica

La densidad de las observaciones futuras, $x_{m+1}, \ldots, x_n$, condicional en una muestra aleatoria observada, $x_1, \ldots, x_m$, es $p(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m)$. La siguiente definición describe una forma posible de reducir los datos e incorporarlos en la estructura de esta densidad condicional.

Definición 1.3.2 (Suficiencia predictiva)**.** *Dada una sucesión de cantidades aleatorias $x_1, x_2, \ldots$, con medida de probabilidad P, donde x_i toma valores en $\mathcal{X}_i$, $i = 1, 2, \ldots$, la sucesión de estadísticas $\mathbf{t}_1, \mathbf{t}_2, \ldots$, con $\mathbf{t}_j$ definida sobre*

$\mathcal{X}_1 \times \cdots \times \mathcal{X}_j$, se dice que es suficiente predictiva para la sucesión $x_1, x_2, \ldots$ si, para toda $m \geq 1$, $r \geq 1$ y $\{i_1, \ldots, i_r\} \bigcap \{1, \ldots, m\} = \emptyset$, se tiene que

$$p(x_{i_1}, \ldots, x_{i_r} | x_1, \ldots, x_m) = p(x_{i_1}, \ldots, x_{i_r} | \mathbf{t}_m)$$

donde $p(\cdot|\cdot)$ es la densidad condicional inducida por P.

La definición captura la idea de que, dada $\mathbf{t}_m = \mathbf{t}_m(x_1, \ldots, x_m)$, los valores individuales de $x_1, \ldots, x_m$ no contribuyen con más información para las probabilidades de los eventos futuros. Las observaciones futuras $(x_{i_1}, \ldots, x_{i_r})$ y los valores observados $(x_1, \ldots, x_m)$ son condicionalmente independientes dada $\mathbf{t}_m$.

Suponiendo que las variables son intercambiables, otra forma de definir la estadística $\mathbf{t}_m = \mathbf{t}_m(x_1, \ldots, x_m)$ como suficiente es la siguiente.

Definición 1.3.3 (Suficiencia paramétrica)**.** *Si $x_1, x_2, \ldots$ es una sucesión infinita intercambiable de cantidades aleatorias, donde x_i toma valores en $\mathcal{X}_i$, $i = 1, 2, \ldots$, la sucesión de estadísticas $\mathbf{t}_1, \mathbf{t}_2, \ldots$, con $\mathbf{t}_j$ definida en $\mathcal{X}_1 \times \cdots \times \mathcal{X}_j$, es suficiente paramétrica para $x_1, x_2, \ldots$ si, para cualquier $n \geq 1$,*

$$dQ(\theta | x_1, \ldots, x_n) = dQ(\theta | \mathbf{t}_n),$$

para cualquier $dQ(\theta)$ que defina un modelo de probabilidad predictiva intercambiable vía la representación

$$p(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) dQ(\theta).$$

A partir de estas definiciones se pueden establecer las siguientes proposiciones.

Proposición 1.3.1 (Equivalencia de suficiencias predictiva y paramétrica)**.** *Dada una sucesión infinita intercambiable de cantidades aleatorias $x_1, x_2, \ldots$, donde x_i toma valores en $\mathcal{X}_i$, $i = 1, 2, \ldots$, la sucesión de estadísticas $\mathbf{t}_1, \mathbf{t}_2, \ldots$ con $\mathbf{t}_j$ definida en $\mathcal{X}_1 \times \cdots \times \mathcal{X}_j$ es suficiente predictiva si, y sólo si, es suficiente paramétrica.*

Proposición 1.3.2 (Criterio de factorización de Neyman-Fisher)**.** *La sucesión de estadísticas $\mathbf{t}_1, \mathbf{t}_2, \ldots$ es suficiente paramétrica para $x_1, x_2, \ldots$ (intercambiable infinitamente con representación paramétrica infinita) si y sólo si, para cualquier $m \geq 1$, la densidad conjunta para $x_1, \ldots, x_m$ dada θ tiene la forma*

$$p(x_1, \ldots, x_m | \theta) = h_m(\mathbf{t}_m, \theta) g(x_1, \ldots, x_m),$$

para algunas funciones $h_m \geq 0$, $g > 0$.

Proposición 1.3.3 (Suficiencia e independencia condicional)**.** *La sucesión* $\mathbf{t}_1, \mathbf{t}_2, \ldots$ *es suficiente paramétrica para* $x_1, x_2, \ldots$ *(infinitamente intercambiable) si, y sólo si, para cualquier* $m \geq 1$*, la densidad* $p(x_1, \ldots, x_m|\theta, \mathbf{t}_m)$ *es independiente de* θ*.*

De la definición general de estadística suficiente (paramétrica o predictiva), $\mathbf{t}_n(x_1, \ldots, x_n) = (x_1, \ldots, x_n)$ siempre es una estadística suficiente. Dado que podemos encontrar varias estadísticas suficientes para el mismo modelo paramétrico, cabe preguntarse si existe una estadística que sea suficiente y además resuma la información de manera óptima.

Definición 1.3.4 (Estadística suficiente minimal)**.** *Si* $x_1, x_2, \ldots$*, es una sucesión intercambiable infinita de cantidades aleatorias, donde* x_i *toma valores en* $\mathcal{X}_i$*, la sucesión de estadísticas* $\mathbf{t}_1, \mathbf{t}_2, \ldots$*, con* $\mathbf{t}_j$ *definida en* $\mathcal{X}_1 \times \cdots \times \mathcal{X}_j$*, es suficiente minimal para* $x_1, x_2, \ldots$ *si dada cualquier otra sucesión de estadísticas suficientes,* $\mathbf{s}_1, \mathbf{s}_2, \ldots$*, existen funciones* $g_1(\cdot), g_2(\cdot), \ldots$ *tales que* $\mathbf{t}_i = g_i(\mathbf{s}_i)$*,* $i = 1, 2, \ldots$.

Intuitivamente hablando, una vez que se conoce el valor de t, no hay otras cantidades calculadas que den nueva información acerca de θ si el modelo muestral $p(x) = p(x|\theta)$ es verdadero. El vector completo $\mathbf{x}$ siempre es suficiente para θ. Una estadística suficiente minimal siempre posee la dimensión más pequeña posible, entre las diferentes estadísticas suficientes.

Ejemplo. *Modelo Bernoulli.*

Si $x_1, x_2, \ldots$ es una sucesión infinitamente intercambiable con valores en 0 y 1, tenemos la representación general

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_0^1 p(x_1, \ldots, x_n|\theta) dQ(\theta) \\ &= \int_0^1 \prod_{0=1}^{n} Br(x_i|\theta) dQ(\theta) \\ &= \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n} dQ(\theta), \end{aligned}$$

donde $s_n = x_1 + \cdots + x_n$. Notemos que podemos escribir

$$p(x_1, \ldots, x_n|\theta) = h_n(s_n, \theta) g(x_1, \ldots, x_n),$$

con

$$h_n(s_n, \theta) = \theta^{s_n}(1-\theta)^{n-s_n}, \qquad g(x_1, \ldots, x_n) = 1,$$

de esta manera, por el criterio de factorización de Neyman-Fisher, la sucesión $s_1, s_2, \ldots$ es suficiente predictiva y paramétrica para $x_1, x_2, \ldots$ y además es una estadística suficiente minimal. ■

Definición 1.3.5 (Estadística ancilar)**.** *Una estadística $a(\mathbf{x})$ se dice* ***ancilar*** *con respecto a θ en un modelo parámetrico $p(\mathbf{x}|\theta)$, si $p(a(\mathbf{x})|\theta) = p(a(\mathbf{x}))$ para todos los valores de θ.*

Ejemplo. *Suficiencia y ancilaridad.*

Sea (X_1, X_2, X_3) un vector aleatorio con distribución multinomial con tres categorías, Multinomial$(x_1, x_2, x_3|\theta_1, \theta_2, \theta_3, n)$, con función de masa de probabilidad

$$p(x_1, x_2, x_3|\theta_1, \theta_2, \theta_3, n) = \frac{n!}{x_1!x_2!x_3!}\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}$$

donde $0 \leq x_i \leq n$, $0 < \theta_i < 1$, para $i = 1, 2, 3$, con $n = \sum_{i=1}^{3} x_i$ y $1 = \sum_{i=1}^{3} \theta_i$.

Dado que podemos reescribir una variable y un parámetro en términos de los otros por ser linealmente dependientes, sean $x_3 = n - x_1 - x_2$ y $\theta_3 = 1 - \theta_1 - \theta_2$. Entonces la función de masa de probabilidad es

$$p(x_1, x_2|\theta_1, \theta_2, n) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!}\theta_1^{x_1}\theta_2^{x_2}(1 - \theta_1 - \theta_2)^{n-x_1-x_2}.$$

Sea $\varphi_2 = \frac{\theta_2}{1-\theta_1}$. Entonces

$$\begin{aligned} p(x_1, x_2|\theta_1, \varphi_2, n) &= \binom{n}{x_1}\theta_1^{x_1}(1 - \theta_1)^{n-x_1}\binom{n - x_1}{x_2}\varphi_2^{x_2}(1 - \varphi_2)^{(n-x_1)-x_2} \\ &= Bin(x_1|\theta_1, n)Bin(x_2|\varphi_2, n - x_1). \end{aligned}$$

Por lo tanto x_1 es una estadística suficiente para θ_1 y una estadística ancilar para θ_2. En otras palabras, x_1 es totalmente informativa acerca de θ_1, mientras que x_2 no contiene información sobre θ_1.

Por otro lado, x_1 no es directamente informativa sobre φ_2, simplemente afecta la distribución condicional de x_2 dada x_1 (que sólo depende de φ_2). Por lo tanto, las inferencias sobre φ_2 sólo se basan en dicha distribución condicional. Las inferencias hacen uso del valor observado de x_1, pero ignoran el mecanismo aleatorio que generó ese valor. Se obtendrían las mismas inferencias sobre φ_2 si hubiéramos fijado el valor de x_1 desde el principio. ■

1.3.2. Familias Exponenciales

En la sección anterior se estudiaron las definiciones de estadística predictiva y paramétrica en forma general; en esta sección estudiaremos estadísticas suficientes de dimensiones fijas. Para más detalles ver Bernardo y Smith (1994).

Consideremos la propiedad de intercambiabilidad y la representación con respecto a $dQ(\theta)$ en una forma paramétrica específica

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta), \qquad (x_1, \ldots, x_n) \in \mathcal{X}^n \subseteq \Re^n$$

donde θ es un parámetro de dimensión k. Si $p(x|\theta)$ es tal que $p(x_1, \ldots, x_n|\theta)$ se factoriza en $h_n(\mathbf{t}_n, \theta)g(x_1, \ldots, x_n)$, para algunas funciones h_n y g, la estadística $\mathbf{t}_n = \mathbf{t}_n(x_1, \ldots, x_n)$ sería suficiente. Una clase importante de funciones $p(x|\theta)$ está identificada por la siguiente definición.

Definición 1.3.6 (Familia exponencial con k parámetros)**.** *Una densidad de probabilidad (o función de masa) $p(x|\theta)$, $x \in \mathcal{X}$, con parámetro $\theta \in \Theta \subseteq \Re^k$, pertenece a la familia exponencial con k parámetros si es de la forma*

$$p(x|\theta) = f(x)g(\theta) \exp\left\{\sum_{i=1}^{k} c_i \phi_i(\theta) h_i(x)\right\},$$

donde $h = (h_1, \ldots, h_k)$, $\phi(\theta) = (\phi_1, \ldots, \phi_k)$ y, dadas las funciones f, h, ϕ, y las constantes c_i,

$$\frac{1}{g(\theta)} = \int_X f(x) \exp\left\{\sum_{i=1}^{k} c_i \phi_i(\theta) h_i(x)\right\} dx < \infty.$$

Proposición 1.3.4 (Estadísticas suficientes para familias exponenciales con k parámetros)**.** *Si $x_1, \ldots, x_i \in \mathcal{X}$, es una sucesión intercambiable tal que, dada una familia exponencial regular con k parámetros,*

$$p(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} f(x)g(\theta) \exp\left\{\sum_{i=1}^{k} c_i \phi_i(\theta) h_i(x)\right\} dQ(\theta),$$

para alguna $dQ(\theta)$, entonces

$$\mathbf{t}_n = \mathbf{t}_n(x_1, \ldots, x_n) = \left[\sum_{i=1}^{n} h_1(x_i), \ldots, \sum_{i=1}^{n} h_k(x_i)\right] \qquad n = 1, 2, \ldots,$$

es una sucesión de estadísticas suficientes.

Demostración. Note que

$$\prod_{i=1}^{n} f(x_i)g(\theta)\exp\left\{\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(x_i)\right\}$$
$$= \prod_{i=1}^{n} f(x_i)\cdot[g(\theta)]^n \exp\left\{\sum_{j=1}^{k} c_j\phi_j(\theta)\sum_{i=1}^{n} h_j(x_i)\right\}.$$

El resultado se obtiene aplicando el criterio de factorización de Neyman-Fisher. □

En ocasiones es más conveniente expresar a la familia exponencial en forma canónica, dada por la siguiente definición.

Definición 1.3.7 (Forma Canónica)**.** *La densidad de probabilidad (o función de masa)*

$$p(\mathbf{y}|\psi) = a(\mathbf{y})\exp\left\{\mathbf{y}'\psi - b(\psi)\right\}, \qquad \mathbf{y}\in\mathcal{Y},$$

obtenida de la familia exponencial con k parámetros, vía las transformaciones

$$\mathbf{y} = (y_1,\ldots,y_k), \qquad \psi = (\psi_1,\ldots,\psi_k),$$
$$y_i = h_i(x), \qquad \psi_i = c_i\phi_i(\theta), \qquad i = 1,\ldots,k,$$

se llama la ***forma canónica*** *de la familia exponencial.*

Proposición 1.3.5 (Primeros dos momentos de la familia exponencial)**.** *Para* $\mathbf{y}$ *con densidad de probabilidad (o función de masa) perteneciente a la familia exponencial canónica, se tiene que*

$$E(\mathbf{y}|\psi) = \nabla b(\psi), \qquad V(\mathbf{y}|\psi) = \nabla^2 b(\psi).$$

Demostración. Es fácil verificar que la función característica de $\mathbf{y}$ condicional a ψ está dada por

$$E(\exp\{i\mathbf{u}'\mathbf{y}\}|\psi) = \exp\{b(i\mathbf{u}+\psi) - b(\psi)\},$$

de la cual se obtienen los momentos directamente. □

Proposición 1.3.6 (Suficiencia en la familia exponencial)**.** *Si las cantidades aleatorias* $\mathbf{y}_1,\ldots,\mathbf{y}_n$ *son independientes con densidad de probabilidad (o función de masa) perteneciente a la familia exponencial, entonces*

$$\mathbf{s} = \sum_{i=1}^{n}\mathbf{y}_i$$

es una estadística suficiente y tiene una distribución

$$p(\mathbf{s}|\psi) = a^{(n)}(\mathbf{s}) \exp\{\mathbf{s}'\psi - nb(\psi)\}$$

donde $a^{(n)}$ *es una convolución de orden* n *de* a.

Demostración. La suficiencia es inmediata ya que las variables pertenecen a la familia exponencial. La función característica de $\mathbf{s}$ es

$$\exp\{nb(i\mathbf{u} + \psi) - nb(\psi)\}$$

de tal manera que la distribución de $\mathbf{s}$ es como se pretendía, donde $a^{(n)}$ satisface

$$nb(\psi) = \log \int a^{(n)}(\mathbf{s}) \exp\{\psi'\mathbf{s}\} d\mathbf{s}.$$

Examinando la convolución para $n = 1$, y por inducción, se establece la forma de $a^{(n)}$. □

Ejemplo

Sea x variable aleatoria con distribución $U(x|0, \theta)$, con densidad de probabilidad

$$p(x|\theta) = U(x|0, \theta) = \theta^{-1}, \qquad x \in (0, \theta),\ \theta \in \Re^+.$$

Esta distribución pertenece a la familia exponencial no regular en donde

$$f(x) = 1, \quad g(\theta) = \theta^{-1}, \quad h(x) = 0, \quad \phi(\theta) = \theta, \quad c = 1.$$

Sea $x_1, \ldots, x_n$ una muestra de variables aleatorias independientes con distribución $U(x|0, \theta)$,

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^{-n} I_{(0,\theta)}(\max_{i=1,\ldots,n}\{x_i\}), \qquad (x_1, \ldots, x_n) \in \Re^n.$$

Por el criterio de factorización de Neyman-Fisher

$$t_n = t_n(x_1, \ldots, x_n) = \max_{i=1,\ldots,n}\{x_i\}, \qquad n = 1, 2, \ldots,$$

es una sucesión de estadísticas suficientes en este caso. ■

1.3.3. Cortes

Para un parámetro multidimensional, el interés de la inferencia algunas veces se concentr en un subconjunto apropiado del parámetro. Un procedimiento usado es particionar el vector de parámetros en dos partes mutuamente exclusivas -la porción que es de interés más los parámetros de ruido- y obtener inferencias condicionados sobre una estadística que es no informativa acerca de la porción de interés, si existe tal estadística auxiliar. Una separación similar también puede ser deseable cuando el contexto requiere que las inferencias sean condicionadas sobre ciertas estadísticas.

En general, hay dos tipos de separación de parámetros. La forma débil es cuando dos porciones complementarias de un vector de parámetros son ortogonales, sus estimadores máximo verosímiles son asintóticamente independientes y la varianza asintótica es la misma si el estimador del parámetro de ruido es tratado como fijo (verosimilitud perfil) o, alternativamente, como aleatorio (verosimilitud completa). La forma fuerte de separación es la independencia en verosimilitud, en la cual la verosimilitud puede ser factorizada. La existencia de un corte facilita tal factorización en términos de la partición de parámetros.

Un corte se define formalmente como sigue.

Definición 1.3.8 (Corte)**.** *Suponga que el parámetro puede ser particionado (posiblemente después de una reparametrización) en* (φ, χ)*, y que la estadística suficiente es* $(s^{(1)}, s^{(2)})$*. Si*

(1) los parámetros φ *y* χ *son independientes en variación (i.e., el rango de* (φ, χ) *es el producto cartesiano del rango de* φ *y el rango de* χ*),*

(2) la distribución condicional de los datos dada $s^{(2)}$ *depende sólo de* φ*, y*

(3) la distribución marginal de $s^{(2)}$ *depende sólo de* χ *y no de* φ*,*

entonces se dice que la estadística $s^{(2)}$ *es un corte.*

Cuando existe un corte, la verosimilitud se factoriza como:

$$g\left(\varphi, s^{(1)}|s^{(2)}\right) f\left(\chi, s^{(2)}\right),$$

esto es, $s^{(2)}$ no da información acerca de φ, pero su valor observado puede indicar la naturaleza de la información acerca de φ (ver ejemplo siguiente). La eficiencia se gana debido a la eliminación del parámetro de ruido χ condicionando sobre $s^{(2)}$. En el análisis Bayesiano, si la distribución inicial se factoriza como $p_1(\varphi)p_2(\chi)$, entonces los cortes permiten también la factorización de la distribución final correspondiente.

Ejemplo

Considere una distribución trinomial con función de masa de probabilidad

$$p(x_1, x_2|\theta_1, \theta_2) = \frac{1}{x_1!x_2!(1-x_1-x_2)!}\theta_1^{x_1}\theta_2^{x_2}(1-\theta_1-\theta_2)^{1-x_1-x_2},$$

donde $x_i \in \{0, 1\}$, $0 < \theta_i < 1$ para $i = 1, 2$, y $0 \leq x_1+x_2 \leq 1$, $0 < \theta_1+\theta_2 < 1$.

Sean $\varphi = \frac{\theta_1}{1-\theta_2}$ y $\chi = \theta_2$, entonces con la reparametrización en términos de φ y χ la función de masa de probabilidad es

$$\begin{aligned} p(x_1, x_2|\varphi, \chi) &= \binom{1-x_2}{x_1}\varphi^{x_1}(1-\varphi)^{1-x_2-x_1}\binom{1}{x_2}\chi^{x_2}(1-\chi)^{1-x_2} \\ &= Bin(x_1|\varphi, 1-x_2)Bin(x_2|\chi, 1). \end{aligned}$$

De esta manera el parámetro se particiona en φ y χ, y las correspondientes estadísticas suficientes son $s^{(1)} = x_1$ y $s^{(2)} = x_2$. Entonces, por la definición 1.3.8, x_2 es un corte. ■

Para mayores detalles acerca de las familias exponenciales y cortes ver Barndorff-Nielsen (1978), Efstathiou, Gutiérrez-Peña y Smith (1998), Ip y Wang (2007).

Capítulo 2

Análisis Clásico de Datos Categóricos

En este capítulo se estudian los datos categóricos arreglados en tablas de contingencia desde el enfoque de la estadística Clásica, utilizando principalmente pruebas estadísticas y la modelación de tablas de contingencia usando modelos loglineales. Para un análisis más detallado ver Agresti (1984, 1996, 2002).

En la sección 2.1 se estudian las tablas de contingencia, su notación, sus distribuciones y sus propiedades de asociación. En la sección 2.2 se estudia la paradoja de Simpson. En la sección 2.3 se presenta el análisis estadístico para tablas de contingencia. Finalmente, en la sección 2.4 se analizan los modelos loglineales; estos modelos se utilizan generalmente para modelar estructuras más complejas en tablas de contingencia.

2.1. Tablas de Contingencia

Cuando cada miembro de una muestra se clasifica simultáneamente en dos o más variables categóricas se usan las tablas de contingencia[1] para poder mostrar las frecuencias de las observaciones que ocurren en las diferentes combinaciones de las categorías de las variables. Cada celda en la tabla muestra el número de observaciones que tiene una cierta combinación.

Las tablas para dos variables tienen dos dimensiones, r renglones que representan las r categorías de una variable y c columnas que representan las c categorías de la segunda variable. Las rc celdas de la tabla contienen las frecuencias de las rc combinaciones de las categorías de las variables. Un caso importante de una tabla de contingencia $r \times c$ se presenta cuando $r = c = 2$.

[1]En inglés se conocen como *contingency tables, cross-classification tables* o *cross-tab.*

La tabla de contingencia 2×2 generalmente resulta de la comparación de dos grupos o de dos variables cuyas respuestas son dicotómicas (variables con dos categorías o variables binarias); por ejemplo, una comparación entre hombres y mujeres respecto al número de los que están a favor o en contra del aborto, o una comparación de dos medicamentos con respecto a sus efectos positivos o negativos al tratar una enfermedad.

2.1.1. Notación y Distribuciones

Sea n_{ij} el número de observaciones en la celda correspondiente al renglón i y columna j, y sea p_{ij} la proporción de la muestra total correspondiente a esa celda. De esta manera tenemos que $p_{ij} = n_{ij}/n$, donde $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$ es el tamaño total de la muestra, y $\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$. El conjunto $\{p_{ij} : i = 1, \ldots, r,\ j = 1, \ldots, c\}$ es la *distribución muestral conjunta* de las observaciones. Las *distribuciones muestrales marginales* son los totales por renglón y por columna obtenidos al sumar las columnas de cada renglón y los renglones de cada columna, respectivamente, de las proporciones conjuntas. Éstas estarán denotadas por $\{p_{i+}\}$ para la variable renglón, donde $p_{i+} = \sum_{j=1}^{c} p_{ij}$, y por $\{p_{+j}\}$ para la variable columna, donde $p_{+j} = \sum_{i=1}^{r} p_{ij}$. Note que $p_{i+} = n_{i+}/n$, $p_{+j} = n_{+j}/n$ y $\sum_{i=1}^{r} p_{i+} = \sum_{j=1}^{c} p_{+j} = 1$.

Para la *distribución condicional muestral* del renglón i, denotamos por $p_{j|i}$ a la proporción de observaciones cuya respuesta es la j-ésima categoría de la variable columna. De esta manera $p_{j|i} = n_{ij}/n_{i+}$, donde $n_{i+} = \sum_{j=1}^{c} n_{ij}$ es el número total de observaciones del i-ésimo renglón y $\sum_{j=1}^{c} p_{j|i} = 1$.

Estas notaciones han sido definidas para los datos muestrales. Para las probabilidades poblacionales las definiciones son análogas, $\{\pi_{ij}\}$ es el conjunto de las *probabilidades conjuntas*, $\{\pi_{i+}\}$ es el conjunto de las *probabilidades marginales* de la variable renglón, $\{\pi_{+j}\}$ es el conjunto de las *probabilidades marginales* de la variable columna, y $\{\pi_{j|i}\}$ es el conjunto de las *probabilidades condicionales*.

La tabla 2.1 muestra la notación para las proporciones muestrales y las probabilidades poblacionales.

2.1.2. Independencia de las Variables Categóricas

Para describir la asociación entre dos variables categóricas podemos usar su distribución conjunta, la distribución condicional de la variable columna dada la variable renglón, o la distribución condicional de la variable renglón dada la variable columna. Las probabilidades poblacionales condicionales,

Renglón	Columna 1	Columna 2	Total
1	p_{11}	p_{12}	p_{1+}
	$p_{1\|1}$	$p_{2\|1}$	1
2	p_{21}	p_{22}	p_{2+}
	$p_{1\|2}$	$p_{2\|2}$	1
Total	p_{+1}	p_{+2}	1

Renglón	Columna 1	Columna 2	Total
1	π_{11}	π_{12}	π_{1+}
	$\pi_{1\|1}$	$\pi_{2\|1}$	1
2	π_{21}	π_{22}	π_{2+}
	$\pi_{1\|2}$	$\pi_{2\|2}$	1
Total	π_{+1}	π_{+2}	1

Tabla 2.1: Notación para las proporciones muestrales conjuntas, condicionales y marginales, y las probabilidades poblacionales conjuntas, condicionales y marginales.

conjuntas y marginales se relacionan por medio de

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} \qquad \forall i, j,$$

y satisfacen $\sum_i \sum_j \pi_{ij} = 1$ y $\sum_j \pi_{j|i} = 1$ para $i = 1, \ldots, r$.

Se dice que dos variables son *independientes* si todas las probabilidades conjuntas son iguales al producto de las correspondientes probabilidades marginales, es decir

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{para } i = 1, \ldots, r \text{ y } j = 1, \ldots, c.$$

La independencia de dos variables implica que las distribuciones condicionales dentro de los r renglones son idénticas, es decir $\pi_{j|i} = \pi_{ij}/\pi_{i+} = \pi_{i+}\pi_{+j}/\pi_{i+} = \pi_{+j}$ para $i = 1, \ldots, r$. Si dos variables son independientes, la probabilidad de obtener una respuesta particular j sobre la variable columna es la misma para cada renglón. Debido a esto, algunas veces al hablar de independencia se hace referencia a la *homogeneidad* de las distribuciones condicionales.

Ejemplo

La tabla 2.2 se basa en los datos presentados por Radelet (1981). Es un ejemplo de una tabla de contingencia de 2×2. El artículo de Radelet se refiere a los efectos de la raza sobre la decisión de condenar a la pena de muerte a un individuo que es declarado culpable por homicidio. Las variables consideradas en la tabla 2.2 son "veredicto de pena de muerte", teniendo categorías {sí, no}, y "raza del acusado", teniendo categorías {blanco, negro}. Los 326 sujetos fueron acusados de homicidio en 20 ciudades de Florida durante 1976-1977.

Para estos datos 36 acusados recibieron la pena de muerte, y 290 acusados no la recibieron, así que consideremos A={Raza del acusado} y B={Pena de

	Pena de muerte	
Raza del acusado	Sí	No
Blanco	19	141
Negro	17	149

Tabla 2.2: Veredicto de pena de muerte por raza del acusado.

muerte}. Las distribuciones marginales muestrales son

$$\begin{aligned} p_{+1} &= p(B = \text{sí}) = \frac{36}{326} = 0.1104 \\ p_{+2} &= p(B = \text{no}) = \frac{290}{326} = 0.8896 \\ p_{1+} &= p(\text{A=blanco}) = \frac{160}{326} = 0.4908 \\ p_{2+} &= p(\text{A=negro}) = \frac{166}{326} = 0.5092 \end{aligned}$$

Las probabilidades condicionales muestrales son

$$\begin{aligned} p_{1|1} &= p(B = \text{sí}|A = \text{blanco}) = \frac{19}{160} = 0.1187 \\ p_{2|1} &= p(B = \text{no}|A = \text{blanco}) = \frac{141}{160} = 0.8813 \\ p_{1|2} &= p(B = \text{sí}|A = \text{negro}) = \frac{17}{166} = 0.1024 \\ p_{2|2} &= p(B = \text{no}|A = \text{negro}) = \frac{149}{166} = 0.8976 \end{aligned}$$

Esto nos indica que el 11.88 % de los 160 acusados de raza blanca recibieron la pena de muerte, mientras que el 10.24 % de los 166 acusados negros recibieron la pena de muerte. Por lo tanto hay evidencia no significativa de discriminación para los acusados de raza blanca.

La tabla 2.3 tiene las frecuencias esperadas estimadas bajo el supuesto de independencia dentro de los paréntesis.

■

2.1.3. Asociación Parcial

En la mayoría de las aplicaciones existen varias variables relevantes, y se usa una tabla multidimensional para mostrar sus datos en una forma más completa. Cuando estudiamos la relación entre dos variables X y Y, por

Raza del acusado	Pena de muerte		Total
	Sí	No	
Blanco	19	141	160
	(17.67)	(142.33)	
Negro	17	149	166
	(18.33)	(147.67)	
Total	36	290	326

Tabla 2.3: Frecuencias esperadas estimadas y observadas.

ejemplo, generalmente existen otras variables ("covariables") cuyos efectos queremos controlar por la posible influencia que puedan tener sobre esa relación.

Las secciones cruzadas de una tabla multidimensional muestran la distribución del conteo de XY de las celdas en los niveles de otra variable (o en varias combinaciones de los niveles de otras variables). Estas secciones cruzadas se conocen como *tablas parciales*. En las tablas parciales la otra o las otras variables están controladas en el sentido de que sus valores son constantes. La tabla de dos dimensiones obtenida al agregar las correspondientes cantidades de las tablas parciales se llama *tabla marginal XY*.

Las asociaciones de las tablas parciales se conocen como *asociaciones condicionales*, porque hacen referencia al efecto que existe entre X y Y condicional en algún nivel fijo de otra variable.

La información contenida en las tablas parciales XY puede ser totalmente diferente de la contenida en la correspondiente tabla marginal XY.

Ejemplo

Utilizando nuevamente los datos de pena de muerte de Radelet (1981), estudiamos la relación bivariada entre el veredicto de la pena de muerte y la raza del acusado de la tabla 2.2. Esta tabla es una tabla de dos dimensiones de una tabla de contingencia de tres dimensiones que Radelet presentó en la cual se incluyó la variable "raza de la víctima". Esta tabla de contingencia expandida se presenta en la tabla 2.4. La tabla 2.4 presenta las tablas parciales de acuerdo a la raza de la víctima. La tabla de contingencia dada en la tabla 2.2 es la tabla de contingencia marginal de las variables raza del acusado y pena de muerte y se obtiene de la tabla 2.4 sumando las frecuencias bajo las dos categorías de la raza de la víctima (i.e., 19+0, 132+9, 11+6, 52+97), para cada una de las cuatro combinaciones de la raza del acusado y el veredicto de pena de muerte. ■

Raza de la víctima	Raza del acusado	Pena de muerte	
		Sí	No
Blanco	Blanco	19	132
Blanco	Negro	11	52
Negro	Blanco	0	9
Negro	Negro	6	97

Tabla 2.4: Tabla $2 \times 2 \times 2$ del veredicto de pena de muerte por raza del acusado y raza de la víctima.

Independencia Marginal y Condicional

Vamos a considerar el tratamiento formal para las probabilidades de las celdas. Consideremos una tabla de clasificación cruzada de tres dimensiones (X, Y, Z). Las probabilidades de las celdas serán denotadas por $\{\pi_{ijk}; i = 1, \ldots, r;\ j = 1, \ldots, c;\ k = 1, \ldots, l\}$, donde $\sum_i \sum_j \sum_k \pi_{ijk} = 1$. Primeramente vamos a estudiar la asociación parcial o "condicional"entre X y Y en tablas parciales en las cuales los valores de Z están fijos.

Una tabla $I \times J \times K$ describe la relación entre las variables X y Y, controlando a la variable Z. Si las variables X y Y son independientes en la tabla parcial fijando $Z = k$, entonces X y Y se llaman *condicionalmente independientes en el nivel* k de Z, esto significa que

$$p(Y = j|X = i, Z = k) = p(Y = j|Z = k) \qquad \forall i, j. \tag{2.1}$$

De manera más general, X y Y se dicen *condicionalmente independientes dada* Z cuando son condicionalmente independientes en cada nivel de Z, esto es, cuando se cumple la ecuación (2.1) para todo k. Entonces, dada Z, Y no depende de X.

Suponga que se tiene una tabla de tres variables, con probabilidades conjuntas $\{\pi_{ijk} = p(X = i, Y = j, Z = k)\}$, entonces

$$\pi_{ijk} = p(X = i, Z = k)p(Y = j|X = i, Z = k),$$

la cual bajo la independencia condicional de X y Y, dada Z, es igual a

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k} \qquad \forall i, j, k.$$

La independencia condicional no implica la independencia marginal. Por ejemplo, bajo la independencia condicional

$$\pi_{ij+} = \sum_k \left(\pi_{i+k}\pi_{+jk}/\pi_{++k}\right).$$

Esto no se simplifica a $\pi_{ij+} = \pi_{i++}\pi_{+j+}$, lo cual ocurre bajo la independencia marginal.

Jerarquía de las Estructuras para Tres Variables

En una tabla de contingencia de tres dimensiones las asociaciones parciales pueden tomar varias formas; estas formas pueden describirse en términos de las características estructurales tales como independencia e interacción. Por ejemplo, una tabla puede estar caracterizada por cualquier par de variables que son independientes condicionalmente y por la interacción de tres factores.

Cada una de las estructuras presentadas a continuación será representada como un modelo loglineal para las frecuencias esperadas de las celdas (ver sección 2.4).

Para tablas de tres dimensiones consideraremos una jerarquía de cinco tipos de estructuras, ordenadas en términos de la extensión de la asociación y la interacción de tres factores:

1) Los tres pares de variables son independientes condicionalmente, esto es:

   X es independiente de Y, dado Z.

   X es independiente de Z, dado Y.

   Y es independiente de Z, dado X.

2) Dos de los pares de variables son independientes condicionalmente, por ejemplo:

   X es independiente de Z, dado Y.

   Y es independiente de Z, dado X.

   X y Y son dependientes condicionalmente, dado Z.

   Similarmente, la dependencia condicional podría ser para el par X y Z, o podría ser para Y y Z.

3) Uno de los pares de las variables es independiente condicionalmente, por ejemplo:

   X es independiente de Z, dado Y.

   X y Y son dependientes condicionalmente, dado Z.

   Y y Z son dependientes condicionalmente, dado X.

   Similarmente, la independencia condicional podría ser para el par de variables X y Y, o podría ser para Y y Z.

4) Ninguno de los pares de las variables es independiente condicionalmente, pero no hay interacción de tres factores.

5) Hay interacción de tres factores. De aquí todos los pares de variables son dependientes condicionalmente, pero la asociación entre cada par varía de acuerdo al nivel de la tercera variable.

2.2. Paradoja de Simpson

Considere una situación donde

$$p(A|B) > p(A|B^c)$$

de manera que hay una asociación positiva entre A y B y sea D otro evento. Suponga que, a pesar de la desigualdad anterior, se satisface que

$$p(A|B, D) < p(A|B^c, D)$$

y

$$p(A|B, D^c) < p(A|B^c, D^c)$$

de tal manera que existe una asociación negativa entre A y B, tanto si D ocurre como si D no ocurre. Este fenómeno se conoce como la *Paradoja de Simpson*.

Notemos que

$$p(A|B) = \alpha p(A|B, D) + (1 - \alpha)p(A|B, D^c)$$

y

$$p(A|B^c) = \beta p(A|B^c, D) + (1 - \beta)p(A|B^c, D^c)$$

donde $\alpha = p(D|B)$ y $\beta = p(D|B^c)$. Si B y D fueran independientes, entonces $\alpha = \beta$, y sería imposible obtener la desigualdad inversa descrita anteriormente. Si α y β son distintas, entonces esta desigualdad inversa puede ocurrir.

Note que

$$p(A|D) = \gamma p(A|B, D) + (1 - \gamma)p(A|B^c, D)$$

y

$$p(A|D^c) = \delta p(A|B, D^c) + (1 - \delta)p(A|B^c, D^c)$$

donde $\gamma = p(B|D)$ y $\delta = p(B|D^c)$.

Lema 2.2.1. *Para cualquier* $p(A|B) > p(A|B^c)$*, ambas contenidas en* $(0,1)$*, es posible encontrar* P_1*,* P_2*,* Q_1 *y* Q_2 *dentro del intervalo* $(0,1)$*, tales que* $P_1 < P_2$ *y* $Q_1 < Q_2$ *con*

$$p(A|B) = \alpha P_1 + (1-\alpha)Q_1$$

y

$$p(A|B^c) = \beta P_2 + (1-\beta)Q_2$$

para alguna α *y* β*, ambas en el intervalo* $[0,1]$*.*

Demostración. Si $\alpha = 1$ y $\beta = 0$ entonces $p(A|B) = P_1$ y $p(A|B^c) = Q_2$. Sea P_2 cualquier valor mayor que $p(A|B)$ y Q_1 cualquier valor menor que $p(A|B^c)$. Entonces todas las condiciones descritas en el lema se cumplen.

Una ligera extensión nos dice que α y β necesitan estar dentro del intervalo $(0,1)$. Si $p(A|B)$ y $p(A|B^c)$ caen dentro del intervalo $(0,1)$, entonces se requerirán los valores extremos menos lejanos que $\alpha = 1$ y $\beta = 0$. □

El lema 2.2.1 nos dice que la desigualdad inversa descrita anteriormente puede ocurrir aún si $p(A|B)$ y $p(A|B^c)$ son muy diferentes.

Ejemplo

Utilizando nuevamente los datos de pena de muerte de Radelet (1981) presentados en la tabla 2.4, la tabla marginal de las variables raza del acusado y pena de muerte nos indican que el 11.88 % de los 160 acusados blancos recibieron la pena de muerte, mientras que el 10.24 % de los 166 acusados negros recibieron la pena de muerte. Por lo tanto hay evidencia no significativa de discriminación para los acusados de raza blanca. Sin embargo, cuando utilizamos la tabla completa y consideramos la raza de la víctima ocurre la paradoja de Simpson.

Considerando a los acusados de víctimas de raza blanca, el 12.58 % de los acusados blancos recibieron la pena de muerte, mientras que el 17.46 % de los acusados negros recibieron la pena de muerte. Además, considerando a los acusados de víctimas de raza negra, ningún acusado blanco recibió la pena de muerte, mientras que el 5.8 % de los acusados negros recibieron la pena de muerte.

Las probabilidades se muestran a continuación. Sea A={Raza del acusado}, B={Pena de muerte} y D={Raza de la víctima}, entonces

$$\begin{aligned} p(\text{B} = \text{sí}|\text{A} = \text{b}) = 0.118 &> p(\text{B} = \text{sí}|\text{A} = \text{n}) = 0.102 \\ p(\text{B} = \text{sí}|\text{A} = \text{b}, \text{D} = \text{b}) = 0.125 &< p(\text{B} = \text{sí}|\text{A} = \text{n}, \text{D} = \text{b}) = 0.174 \\ p(\text{B} = \text{sí}|\text{A} = \text{b}, \text{D} = \text{n}) = 0 &< p(\text{B} = \text{sí}|\text{A} = \text{n}, \text{D} = \text{n}) = 0.582 \end{aligned}$$

Por lo tanto hay evidencia de discriminación para los acusados de raza negra. ■

La paradoja de Simpson generalmente ocurre cuando se tienen datos que no están conectados o que no son aleatorios.

La paradoja de Simpson puede usarse para rebatir muchos análisis de tablas de contingencia que se basan en datos que no son aleatorios. Siempre puede existir una variable que lleve a conclusiones opuestas, si la tabla de contingencia fuera dividida de acuerdo a esta variable. Sin embargo, si por ejemplo los datos son aleatorios (si los acusados blancos y los acusados negros hubieran sido elegidos de manera aleatoria de subpoblaciones de acusados blancos y acusados negros) entonces es menos probable que alguna variable afecte las conclusiones del análisis. El problema puede minimizarse si los individuos pueden considerarse subjetivamente como miembros "intercambiables"de una población. En cualquier caso, cualquier conclusión obtenida de una tabla de contingencia, que se basa en datos que no son aleatorios, puede considerarse como subjetiva.

Para más detalles acerca de la paradoja de Simpson consultar Leonard y Hsu (1994) y Leonard (2000).

2.3. Análisis

En el análisis de tablas de contingencia, aparte de poder ver la descripción de los datos muestrales $\{n_{ij}\}$, se hace inferencia acerca de la estructura de la tabla suponiendo probabilidades $\{\pi_{ij}\}$ que son desconocidas. En esta sección se estudian dos estadísticas que se usan para probar hipótesis acerca de las asociaciones entre las variables.

2.3.1. Esquemas de Muestreo

En el análisis de tablas de contingencia de dos dimensiones hay tres esquemas de muestreo que ocurren en la práctica:

- *Esquema 1* (muestreo multinomial o *multinomial sampling*), donde sólo el número total de observaciones n es fijo.
- *Esquema 2* (muestreo multinomial-producto o muestreo estratificado, *product-multinomial sampling* o *stratified sampling*), donde el total por renglones $\{n_{i+}\}$ o el total por columnas $\{n_{+j}\}$ son fijos.
- *Esquema 3*, donde el total por renglones y el total por columnas son fijos.

Los esquemas de muestreo análogos ocurren en tablas de contingencia cuando algunos de los totales marginales están dados por el experimentador. El esquema 3 no es muy común. En el caso de una tabla de contingencia de 2×2 es la estructura que lleva a la prueba exacta de Fisher.

Para el caso de tablas de contingencia de tres dimensiones ($q \times r \times s$) hay seis esquemas de muestreo:

- *Esquema 1*, muestreo multinomial con qrs categorías, probabilidades de celda p_{hij}, y tamaño de muestra n.
- *Esquema 2*, estableciendo los totales $\{n_{h++}\}$ (ó $\{n_{+i+}\}$ ó $\{n_{++j}\}$), y tomando q muestreos multinomiales independientes (ó r ó s), cada uno con rs categorías (ó qs ó qr respectivamente), y con probabilidades $\{p_{hij}/p_{h++}\}$ (ó $\{p_{hij}/p_{+i+}\}$ ó $\{p_{hij}/p_{++j}\}$).
- *Esquema 3*, estableciendo los totales $\{n_{+ij}\}$ (ó $\{n_{h+j}\}$ ó $\{n_{hi+}\}$), y tomando rs (ó qs ó qr) muestreos multinomiales independientes, cada una con q categorías (ó r ó s), y con probabilidades de celda $\{p_{hij}/p_{+ij}\}$ (ó $\{p_{hij}/p_{h+j}\}$ ó $\{p_{hij}/p_{hi+}\}$ respectivamente).
- *Esquema 4*, estableciendo los totales $\{n_{+i+}\}$ y $\{n_{++j}\}$ (ó $\{n_{h++}\}$ y $\{n_{+i+}\}$ ó $\{n_{h++}\}$ y $\{n_{++j}\}$), y tomando muestreos estratificados.
- *Esquema 5*, estableciendo los totales $\{n_{h++}\}$ y $\{n_{+ij}\}$ ($\{n_{+i+}\}$ y $\{n_{h+j}\}$ ó $\{n_{++j}\}$ y $\{n_{hi+}\}$), y tomando muestreos estratificados.
- *Esquema 6*, estableciendo los totales $\{n_{h++}\}$, $\{n_{+i+}\}$ y $\{n_{++j}\}$.

Muestreo Poisson, Binomial y Multinomial

Suponga que una muestra aleatoria de tamaño fijo n se clasifica de acuerdo a dos variables categóricas. La distribución del número de observaciones en las celdas es una distribución multinomial especificada por el tamaño de muestra n y las rc probabilidades poblacionales de las celdas $\{\pi_{ij}\}$. La probabilidad de que el conjunto de celdas contenga $\{n_{ij}\}$ observaciones que sumen n es

$$\left(\frac{n!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \right) \prod_{i=1}^{r} \prod_{j=1}^{c} \pi_{ij}^{n_{ij}}.$$

Este esquema de muestreo se llama *muestreo multinomial completo* (esquema 1).

Suponga que dentro de cada categoría de la variable renglón, una muestra aleatoria independiente se clasifica de acuerdo a la variable columna. Las

celdas del i-ésimo renglón tienen distribución multinomial especificada por el tamaño de muestra n_{i+} y las probabilidades $\{\pi_{j|i}\ j = 1, \ldots, c\}$ y cada una de las celdas de los diferentes renglones son independientes. Las celdas en el renglón i tienen la función de probabilidad

$$\left(\frac{n_{i+}!}{\prod_{j=1}^{c} n_{ij}!}\right) \prod_{j=1}^{c} \pi_{j|i}^{n_{ij}}$$

y el producto de estas probabilidades de cada uno de los r renglones da la función de probabilidad para toda la tabla. Este esquema de muestreo se llama *muestra multinomial independiente* (o *producto*, esquema 2).

Otro esquema de muestreo supone que el número de observaciones de las celdas son variables aleatorias Poisson independientes. Si el número de observaciones n_{ij} de la celda tiene valor esperado m_{ij}, entonces la función de probabilidad para n_{ij} es

$$\frac{e^{-m_{ij}} m_{ij}^{n_{ij}}}{n_{ij}!}$$

para todos los enteros no negativos. El producto de estas probabilidades para las rc celdas da la función de probabilidad para la tabla. Para este esquema el tamaño de la muestra total en la tabla es aleatorio. Puede mostrarse que condicionando sobre el tamaño de la muestra total obtenemos el muestreo multinomial completo. Condicionando sobre los totales marginales obtenemos el muestreo multinomial independiente (ver sección 1.1.1).

Para el muestreo multinomial completo, los estimadores máximo verosímil de $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$ son las proporciones muestrales $\{p_{i+}\}$ y $\{p_{+j}\}$. Bajo el supuesto de independencia de las dos variables tenemos que $\pi_{ij} = \pi_{i+}\pi_{+j}$, y el estimador máximo verosimil de π_{ij} es

$$\hat{\pi}_{ij} = p_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n^2}.$$

Para las muestras multinomial y Poisson, m_{ij} es el valor esperado de la celda con número de observaciones n_{ij}. Si los $\{n_{ij}\}$ tienen distribución multinomial, entonces cada celda tiene una distribución binomial; de esta manera tenemos que $m_{ij} = E(n_{ij}) = n\pi_{ij}$. Las $\{m_{ij}\}$ son las *frecuencias esperadas*.

La estimación de m_{ij} es $\hat{m}_{ij} = n\hat{\pi}_{ij}$. Bajo el supuesto de independencia tenemos que

$$\hat{m}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}.$$

Las $\{\hat{m}_{ij}\}$ son las *frecuencias esperadas estimadas* bajo la hipótesis nula de independencia.

2.3.2. Pruebas de Hipótesis

En 1900 Karl Pearson sugirió la estadística

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

para probar la hipótesis nula:

$$H_o : \text{las variables son independientes}$$

en tablas de dos dimensiones.

Otra estadística de gran utilidad es la aproximación de la razón de verosimilitudes

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right).$$

Cuando H_o es verdadera, ambas X^2 y G^2 tienen distribución asintótica ji-cuadrada (cuando $n \to \infty$) con grados de libertad $df = (r-1)(c-1)$. Las dos estadísticas de prueba son equivalentes asintóticamente; en el sentido de que la diferencia entre las dos es de orden menor a $1/n$ cuando $n \to \infty$. Si el valor de las estadísticas de prueba es grande se concluye que H_o es falsa.

Para que χ^2 y G^2 tengan una buena aproximación a la distribución ji-cuadrada se sugiere que al menos el 80 por ciento de las celdas tenga un $\hat{m}_{ij}$ mayor a 5, y $\hat{m}_{ij}$ sea mayor a 1 en todas las celdas.

Si $\hat{m}_{ij}$ son pequeñas, generalmente se pueden combinar las categorías de las variables para obtener una tabla con frecuencias mayores. Generalmente no se fusionan las categorías a menos que haya una forma natural de hacerlo. Hay alternativas para usar la ji-cuadrada, como la prueba exacta con distribución muestral hipergeométrica para las frecuencias de las celdas; esta distribución se da cuando las marginales son fijas.

Las $\{\hat{m_{ij}}\}$ dependen de los totales marginales de los renglones y las columnas, por ello las estadísticas X^2 y G^2 para probar independencia no cambian bajo permutaciones de renglones y permutaciones de columnas. Esto implica que ambas clasificaciones son tratadas como escalas nominales en estas pruebas. Si se aplican estas estadísticas para probar independencia entre variables ordinales, se ignorará parte de la información que puede ser relevante.

Resumiendo, en una tabla de contingencia de dos dimensiones, la hipótesis nula de independencia estadística entre dos variables tiene la forma

$$H_0 \ : \ \pi_{ij} = \pi_{i+}\pi_{+j},$$

para todo i y j. Las probabilidades marginales especifican las probabilidades conjuntas. Para probar H_0, identificamos a $m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ como la

frecuencia esperada. De aquí, m_{ij} es el valor esperado de n_{ij} bajo el supuesto de independencia. Generalmente $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$ son desconocidas.

Estimamos las frecuencias esperadas sustituyendo las probabilidades desconocidas por las proporciones muestrales, obteniendo

$$\hat{m}_{ij} = np_{i+}p_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

Las $\{\hat{m}_{ij}\}$ son las llamadas *frecuencias esperadas estimadas*. Éstas tienen el mismo total de renglones y columnas que los conteos observados, pero muestran los patrones de independencia.

Para probar la independencia en una tabla de contingencia de $r \times c$, la estadística de Pearson y la estadística de razón de verosimilitudes son iguales a

$$X^2 = \sum \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}, \qquad G^2 = 2\sum n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right),$$

respectivamente.

Para muestras grandes estas estadísticas tienen función de distribución ji-cuadrada con $(r-1)(c-1)$ grados de libertad. Esto significa que bajo H_0, $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$ determinan las probabilidades de las celdas. Hay $r-1$ renglones de probabilidades independientes, debido a que éstos deben sumar 1, y por tanto podemos determinar uno de los renglones por medio de los demás; por ejemplo, el primer renglón está determinado por $\pi_{1+} = 1-(\pi_{2+} + \cdots + \pi_{r+})$. Análogamente, hay $c-1$ columnas de probabilidades independientes, ya que éstas deben sumar 1, también podemos determinar una de las columnas por medio de las demás; por ejemplo, la primera columna está determinada por $\pi_{+1} = 1 - (\pi_{+2} + \cdots + \pi_{+c})$. Así obtenemos un total de $(r-1) + (c-1)$ parámetros. La hipótesis alternativa no especifica las rc probabilidades de celdas. Las probabilidades de celdas deben sumar 1, de tal manera que hay $rc-1$ parámetros independientes. El valor para los grados de libertad es la diferencia entre el número de parámetros bajo la hipótesis alternativa y la hipótesis nula, es decir,

$$(rc-1) - [(r-1) + (c-1)] = rc - r - c + 1 = (r-1)(c-1).$$

Ejemplo

Las estadísticas para los datos de Radelet (1981) para la tabla de contingencia de 2×2 de la tabla 2.3 son $X^2 = 0.22$ y $G^2 = 0.22$, con un grado de libertad. Si comparamos estos valores con el cuantil 95 % de una distribución $\chi^2_{(1)}$, $\chi^2_{(1),0.95} = 3.8415$, estos valores son pequeños y muestran que no hay una asociación significativa entre el veredicto de pena de muerte y la raza

del acusado. Para el caso de la tabla de contingencia de $2 \times 2 \times 2$ de la tabla 2.4 las estadísticas son $X^2 = 122.39$ y $G^2 = 137.92$; si comparamos con el mismo cuantil de una distribuión $\chi^2_{(1)}$ entonces los valores muestran que hay una asociación significativa entre las variables. ■

2.3.3. Residuales

Una prueba estadística y su valor-p describen la evidencia contra la hipótesis nula. Una comparación de celda a celda de las observaciones y las frecuencias esperadas estimadas ayuda a entender mejor la naturaleza de la evidencia. Bajo la hipótesis nula H_0: "las variables son independientes", diferencias grandes entre n_{ij} y $\hat{m}_{ij}$ ocurren en las celdas en donde se tienen frecuencias esperadas m_{ij} grandes.

Los residuales de Pearson están definidos por

$$e_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}} = \frac{n_{ij} - \hat{m}_{ij}}{\hat{m}_{ij}^{1/2}}, \qquad \forall i = 1, \ldots, r,\ j = 1, \ldots, c.$$

Los residuales de Pearson se relacionan con la estadística de Pearson por medio de $\sum_i \sum_j e_{ij}^2 = \chi^2$. Bajo la hipótesis nula H_0: "las variables son independientes", los residuales de Pearson $\{e_{ij}\}$ se distribuyen asintóticamente normal con media cero.

Los residuales de Pearson estandarizados que tienen una distribución asintótica normal estándar resultan de dividir los residuales de Pearson entre su error estándar estimado. Bajo la hipótesis H_0 los residuales de Pearson estandarizados son:

$$\frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}/(1 - p_{i+})(1 - p_{+j})}}, \qquad \forall i = 1, \ldots, r,\ j = 1, \ldots, c.$$

También se conocen como residuales ajustados.

Cuando la hipótesis nula es verdadera, para muestras grandes, cada residual ajustado tiene aproximadamente una distribución normal estándar. Un residual ajustado que excede alrededor de 2 ó 3 en valor absoluto indica que existe carencia de ajuste de H_0 en esa celda.

2.3.4. Devianza

La devianza es la estadística de la razón de verosimilitudes para comparar el modelo de interés M con el modelo más complejo S (en la sección 2.4 se estudiarán los modelos loglineales, los cuales son algunos de los modelos utilizados en el análisis de datos categóricos). Al modelo S se le conoce como

modelo saturado. La devianza es la estadística para probar la hipótesis de que todos los parámetros que aparecen en el modelo saturado pero no en el modelo M son iguales a cero.

La devianza tiene la misma forma que la estadística de razón de verosimilitudes G^2 para el modelo M. Sean L_M y L_S los valores máximos del logaritmo de la verosimilitud para el modelo M y S, respectivamente. La devianza está definida por

$$Devianza = -2[L_M - L_S].$$

Los componentes de la devianza, llamados *residuales de la devianza*, proporcionan un diagnóstico de la falta de ajuste del modelo. Son alternativas para los residuales de Pearson y los residuales ajustados. Similares a los residuales de Pearson, los residuales de la devianza se distribuyen aproximadamente normal.

Para dos modelos, suponga que uno (M_0) es un caso especial del otro (M_1). Suponiendo que el modelo más complejo se cumple, la estadística de razón de verosimilitudes para probar que el modelo más simple (M_0) se cumple es

$$-2[L_0 - L_1] = -2[L_0 - L_S] - \{-2[L_1 - L_S]\} = Devianza_0 - Devianza_1.$$

Uno puede comparar los modelos comparando sus devianzas. Para muestras grandes, esta estadística es aproximadamente ji-cuadrada, con grados de libertad igual a la diferencia entre los grados de libertad de los residuales para los modelos separados. Estos grados de libertad son iguales al número de parámetros adicionales no redundantes que están en M_1 pero no en M_0.

2.4. Modelos Loglineales

Los modelos loglineales generalmente se usan para modelar los conteos de las celdas de las tablas de contingencia. Los modelos especifican cómo los conteos esperados dependen de los niveles de las variables categóricas para cada celda, así como de las asociaciones y las interacciones entre esas variables. El propósito de los modelos loglineales es analizar los patrones de asociación e interacción de las variables. Para más detalle acerca de los modelos loglineales consultar Bishop et al. (1975) y Agresti(1984, 1996, 2002).

2.4.1. Modelos Loglineales para Dos Dimensiones

Considere una tabla de contingencia $r \times c$ que clasifica un muestra con distribución multinomial de n sujetos en dos variables categóricas. Las probabilidades de celda son $\{\pi_{ij}\}$ y las frecuencias esperadas son $\{m_{ij} = n\pi_{ij}\}$.

Los modelos loglineales usan $\{m_{ij}\}$ en lugar de $\{\pi_{ij}\}$. Recordemos que también podemos considerar muestras con distribución Poisson para $r \times c$ conteos de celdas independientes $\{y_{ij}\}$ conociendo los valores de $\{m_{ij} = E(Y_{ij})\}$. En ambos casos los $\{n_{ij}\}$ son los conteos observados.

Ya sea que los datos tengan una distribución multinomial o Poisson, la media es positiva. Si bien los modelos lineales generalizados pueden modelar una media positiva usando la liga identidad, es más común usar el logaritmo de la media. Al igual que cualquier combinación lineal de las covariables, el logaritmo de la media puede tomar cualquier valor real. El logaritmo de la media es el parámetro natural para la distribución Poisson; y la liga logaritmo es la liga canónica para un modelo lineal generalizado cuyos datos se distribuyen Poisson.

Modelo de Independencia

Como ya se mencionó, dos variables son independientes si $\pi_{ij} = \pi_{i+}\pi_{+j}$ para todo i y j. La expresión correspondiente para las frecuencias esperadas $\{m_{ij} = n\pi_{ij}\}$ es $m_{ij} = n\pi_{i+}\pi_{+j}$ para todo i y j. Usando una escala logarítmica la independencia es equivalente a la relación aditiva

$$\begin{aligned} \log m_{ij} &= \log n + \log \pi_{i+} + \log \pi_{+j} \\ &= -\log n + \log m_{i+} + \log m_{+j}. \end{aligned} \tag{2.2}$$

Lo cual nos indica que, si dos variables X y Y son independientes, el logaritmo de la frecuencia esperada para la celda (i, j) es una función aditiva determinada por un efecto del i-ésimo renglón y un efecto de la j-ésima columna.

Sean $\mu_{ij} = \log m_{ij}$,

$$\mu_{i.} = \sum_{j=1}^{c} \frac{\mu_{ij}}{c}, \qquad \mu_{.j} = \sum_{i=1}^{r} \frac{\mu_{ij}}{r}$$

y

$$\mu = \mu_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\mu_{ij}}{rc},$$

la cual denota la media de $\{\log m_{ij}\}$. Entonces, si

$$\begin{aligned} \lambda_i^x &= \mu_{i.} - \mu \\ \lambda_j^y &= \mu_{.j} - \mu \\ \lambda_{ij}^{xy} &= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu \end{aligned}$$

obtenemos que, bajo el modelo de independencia, una formulación alternativa del modelo (2.2) es

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y. \tag{2.3}$$

Este modelo se llama *modelo loglineal de independencia* para una tabla de contingencia de 2 dimensiones.

Como es usual, se requieren condiciones adicionales para los parámetros, tales como $\lambda_i^X = \lambda_i^Y = 0$, ó $\lambda_i^X + \lambda_i^Y = 0$ para alguna i, para poder obtener los parámetros ajustados.

Para este modelo las frecuencias esperadas estimadas por medio del ajuste de máxima verosimilitud son $\{\hat{m}_{ij} = n_{i+}n_{+j}/n\}$. Las pruebas que usan las estadísticas χ^2 y G^2 son también pruebas de bondad de ajuste para este modelo loglineal (ver sección 2.3.2).

Modelo Saturado

Ahora consideremos un modelo loglineal más general para dos variables, que corresponde a diversos tipos de dependencia. Sean $\mu_{ij} = \log m_{ij}$,

$$\mu_{i.} = \sum_{j=1}^{c} \frac{\mu_{ij}}{c}, \qquad \mu_{.j} = \sum_{i=1}^{r} \frac{\mu_{ij}}{r}$$

y

$$\mu = \mu_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\mu_{ij}}{rc},$$

la cual denota la media de $\{\log m_{ij}\}$. Entonces, si

$$\begin{aligned} \lambda_i^x &= \mu_{i.} - \mu \\ \lambda_j^y &= \mu_{.j} - \mu \\ \lambda_{ij}^{xy} &= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu \end{aligned}$$

obtenemos que

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}. \tag{2.4}$$

Este es el modelo más general para dos dimensiones; un modelo de esta forma da un buen ajuste para cualquier conjunto de frecuencias esperadas positivas $\{m_{ij}\}$.

Los $\{\lambda_i^X\}$ y $\{\lambda_j^Y\}$ son desviaciones alrededor de la media, de tal manera que

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0.$$

Así, hay $r-1$ renglones linealmente independientes y $c-1$ columnas linealmente independientes. Los $\{\lambda_i^X\}$ y $\{\lambda_j^Y\}$ se refieren a los números relativos de casos en las categorías de X y Y. Este promedio se calcula sobre una escala logarítmica, ya que $\{\lambda_i^X\}$ se refiere al tamaño relativo de la media aritmética de los $\{\log m_{ij};\ j=1,\ldots,c\}$.

Los $\{\lambda_{ij}^{XY}\}$ satisfacen

$$\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0,$$

de tal manera que $(r-1)(c-1)$ de esos términos son linealmente independientes. El modelo de independencia (2.3) es el caso especial del modelo general (2.4) en el cual todos los $\{\lambda_{ij}^{XY}=0\}$. De esta manera, $\{\lambda_{ij}^{XY}\}$ son *asociaciones paramétricas* que reflejan las desviaciones a partir de la independencia.

El modelo de independencia, $\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y$, tiene $1+(r-1)+(c-1) = r+c-1$ parámetros linealmente independientes. El modelo general tiene los $(r-1)(c-1)$ $\{\lambda_{ij}^{XY}\}$ parámetros de asociación adicionales. El número total de parámetros en el modelo general es igual a $1+(r-1)+(c-1)+(r-1)(c-1) = rc$, que es el número total de celdas en la tabla. Para tablas de cualquier número de dimensiones, el modelo loglineal tiene tantos parámetros como número de celdas en la tabla y se llama *modelo saturado.*

Existen relaciones directas entre los parámetros en los modelos loglineales y los momios. Esta relación es la más simple para una tabla 2×2,

$$\begin{aligned}
\log\theta &= \log\frac{m_{11}m_{22}}{m_{12}m_{21}} = \log m_{11} + \log m_{22} - \log m_{12} - \log m_{21} \\
&= (\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\mu + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\
&\quad -(\mu + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\mu + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{Xy}) \\
&= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}.
\end{aligned}$$

De esta manera, $\{\lambda_{ij}^{XY}\}$ determina la asociación entre X y Y.

Los modelos loglineales son de alguna manera modelos jerárquicos; esto significa que en la mayoría de los casos los modelos incluyen a todos los términos de orden inferior compuestos de las variables contenidas en los términos del modelo. Por ejemplo, cuando un modelo contiene λ_{ij}^{XY}, éste también contiene a los términos λ_i^X y λ_j^Y. Una razón para incluir a los términos de orden inferior es que, de otra manera, la significancia estadística y la interpretación de los términos de orden mayor dependerían en cómo están codificadas las variables, además la diferencia de las devianzas de dos modelos no tendría una distribución ji-cuadrada.

Al igual que en el modelo de independencia, debemos establecer condiciones adicionales sobre los parámetros para poderlos estimar. Para los efectos

principales tendríamos que $\lambda_i^X = \lambda_i^Y = 0$, ó $\lambda_i^X + \lambda_i^Y = 0$ para alguna i. Para los efectos de interacción tenemos varias posibilidades, una de ellas podría ser que $\lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$, y otra sería que $\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$, para todo i y j.

2.4.2. Modelos Loglineales para Tres Dimensiones

La forma general dada en la sección anterior para un modelo loglineal en dos dimensiones se extiende directamente a dimensiones mayores. Sea $\mu_{ijk} = \log m_{ijk}$. Podemos expresar $\log m_{ijk}$ como

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \qquad (2.5)$$

donde

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = \cdots = \sum_k \lambda_{ijk}^{XYZ} = 0.$$

En el modelo (2.5) los términos que tienen doble subíndice pertenecen a las asociaciones parciales, y el término con triple subíndice pertenece a la interacción de los tres factores. Cuando algunos parámetros son iguales a cero en el modelo (2.5), los modelos corresponden a alguna de las estructuras discutidas en la sección 2.3.1, que explicaremos a continuación y que se listan en la tabla 2.5. Por simplicidad cada modelo puede representarse por medio de un símbolo que se lista también en la tabla para los términos de orden mayor para cada variable. El símbolo indica abreviadamente los pares de variables que están asociados. Las variables dependientes condicionalmente aparecen juntas en el símbolo.

Modelos Loglineales	Símbolo
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	(X,Y,Z)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	(XY,Z)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	(XY,YZ)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$	(XY,YZ,XZ)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$	(XYZ)

Tabla 2.5: Algunos Modelos Loglineales para Tablas de Tres Dimensiones.

(X,Y,Z)

Tres variables son completamente independientes si para cada i, j y k, $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$. Sobre una escala logarítmica esto puede representarse

por el modelo

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z,$$

por lo cual cada par de variables es condicionalmente independiente dadas las demás.

Recordemos que si una de las variables es condicionalmente independiente de una o de otras dos variables, entonces la asociación marginal de las otras dos es la misma que su asociación parcial. Para este modelo los tres pares de variables son independientes marginalmente. Por ejemplo, la asociación marginal entre X y Y es la misma asociación parcial entre X y Y, dada Z, porque Z es condicionalmente independiente de X, dada Y, o también porque Z es condicionalmente independiente de Y, dada X. Este modelo es tan simple que rara vez da un buen ajuste en la práctica.

(*XY,Z*) o (*XZ,Y*) o (*YZ,X*)

El símbolo (XY,Z) denota el modelo

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY},$$

para el cual los parámetros $\{\lambda_{ij}^{XY}\}$ pertenecen a la asociación parcial entre X y Y, dada Z. El símbolo para el modelo refleja la dependencia condicional de X y Y. Para este modelo X y Z son independientes condicionalmente (dada Y), y Y y Z son independientes condicionalmente (dada X). Si todos los $\lambda_{ij}^{XY} = 0$, las variables X y Y también son condicionalmente independientes, y el modelo se simplifica a (X,Y,Z).

Las asociaciones marginales son idénticas a las asociaciones parciales. Por ejemplo, la marginal XY es igual a la asociación parcial XY, dada Z, porque Z es independiente de X, dada Y, o porque Z es independiente de Y, dada X. La asociación marginal XZ es la misma que la asociación parcial XZ, dada Y (i.e. X y Z son independientes marginalmente), porque Y es independiente de Z, dada X.

Hay tres modelos separados en este nivel de la jerarquía de estructuras, correspondientes a tres posibles pares de variables que pueden ser condicionalmente dependientes. El símbolo (XZ,Y) denota el modelo tal que X y Z son condicionalmente dependientes, y (YZ,X) denota el modelo tal que Y y Z son condicionalmente dependientes.

(*XY,YZ*) o (*XY,XZ*) o (*XZ,YZ*)

Hay tres modelos separados en los cuales sólo un par de variables es condicionalmente independiente. Por ejemplo, (XY,YZ) denota el modelo

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ},$$

donde $\{\lambda_{ij}^{XY}\}$ y $\{\lambda_{jk}^{YZ}\}$ pertenecen a las asociaciones parciales de XY y YZ, respectivamente. Para este modelo, X y Z (las variables que no aparecen juntas en el símbolo del modelo) son condicionalmente independientes, dada Y.

Para el modelo (XY, YZ), la marginal XY y la marginal YZ muestran las mismas asociaciones parciales correspondientes, porque X y Z son condicionalmente independientes. Sin embargo, la asociación marginal de XZ puede diferir de la asociación parcial de XZ (dada Y), porque Y es condicionalmente dependiente de ambas, X y Z. De esta manera X y Z pueden ser marginalmente dependientes, aún cuando sean son condicionalmente independientes en cada nivel de Y.

(XY, XZ, YZ)

Para el modelo

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}.$$

Como los términos de las asociaciones parciales aparecen para cada par de variables, ningún par es condicionalmente independiente. Este es un caso especial del modelo general en el cual $\lambda_{ijk}^{XYZ} = 0$ para todo i, j, y k, lo que significa que no hay interacción entre los tres factores.

Para este modelo la asociación parcial entre cualquier par de variables es la misma en cada nivel de la tercera variable, pero esa asociación puede diferir de la asociación marginal correspondiente. Por ejemplo, la asociación parcial XY, dada Z, puede diferir de la asociación marginal XY porque Z es condicionalmente dependiente con ambas X y Y.

(XYZ)

El modelo más general para tres variables, el modelo (2.5), presenta la interacción de los tres factores. Cada par de variables puede ser condicionalmente dependiente, y la asociación entre cualquier par puede depender del nivel de la tercera variable. La generalidad de esta estructura es tal que describe el conjunto $S\,[(XYZ)]$ de todos los $\{\pi_{ijk}\}$. El conjunto $S\,[(XY,\ XZ,\ YZ)]$ de $\{\pi_{ijk}\}$ para el cual las interacciones entre los tres factores son $\lambda_{ijk}^{XYZ} = 0$. El conjunto $S\,[(XY,\ YZ)]$ que satisface $(XY,\ YZ)$ es un subconjunto más pequeño de $S\,[(XYZ)]$, que es también un subconjunto de $S\,[(XY,\ XZ,\ YZ)]$. Este conjunto $S\,[(XY,\ YZ)]$ indica que no hay interacción entre los tres factores y la independencia condicional entre X y Z, dada Y; esto es, todos los $\lambda_{ijk}^{XYZ} = 0$ y todos los $\lambda_{ik}^{XZ} = 0$. Por ejemplo, podemos dar la siguiente

jerarquía entre los modelos

$$\begin{aligned} S\left[(X,Y,Z)\right] &\subset S\left[(XY,Z)\right] \\ &\subset S\left[(XY,YZ)\right] \subset S\left[(XY,XZ,YZ)\right] \subset S\left[(XYZ)\right]. \end{aligned}$$

Como $\sum_i \lambda_i^X = 0$ en el modelo (2.5), hay $(r-1)$ parámetros $\{\lambda_i^X\}$ linealmente independientes. Como $\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$, hay $(r-1)(c-1)$ parámetros $\{\lambda_{ij}^{XY}\}$ linealmente independientes. Similarmente, la fórmula se aplica a otros parámetros en los modelos. El número total de parámetros linealmente independientes (incluyendo μ) para el modelo loglineal general (2.5) es

$$\begin{aligned} 1+(r-1) &+ (c-1)+(l-1)+(r-1)(c-1) \\ &+ (c-1)(l-1)+(r-1)(l-1)+(r-1)(c-1)(l-1) = rcl \end{aligned}$$

que es el número total de celdas en la tabla, ver Agresti (2002).

La mayoría de las estadísticas que se usan para probar la bondad de ajuste de un modelo se distribuyen asintóticamente como una χ^2. Los grados de libertad asociados a esta distribución dependen de la estructura de los datos y del número de parámetros independientes en el modelo. Bishop et al. (1975) describen dos formas de calcular los grados de libertad:

a) Contar el número de parámetros independientes iguales a cero. La suma de los parámetros asociados con cada uno de estos términos da el número de grados de libertad.

b) Contar el número de parámetros independientes estimados, y substraer este número del número total de celdas estimadas. Esta diferencia es el número de grados de libertad.

2.4.3. Modelos Loglineales para Dimensiones Mayores

Hemos visto que la estructura de una tabla de contingencia de tres dimensiones es más complicada que la de una tabla de dos dimensiones por la asociación parcial y la interacción de tres factores que deben considerarse. En el caso de los modelos loglineales, una vez que entendemos los distintos modelos para tres dimensiones, es relativamente fácil entender aquellas tablas con mayores dimensiones. La dificultad está en que cuando la dimensión crece, también crece el número de modelos diferentes necesarios para describir todas las posibles asociaciones y las interacciones de orden mayor. Para una tabla con cuatro dimensiones, por ejemplo, hay 112 diferentes modelos

loglineales jerárquicos que son más complejos que el modelo con variables mutuamente independientes.

Para ilustrar estos modelos con dimensiones mayores, consideremos algunos modelos para una tabla de contingencia de cuatro dimensiones con variables W, X, Y y Z. Generalmente el modelo de interés más simple es donde las variables son mutuamente independientes, denotado por (*W, X, Y, Z*). Este modelo rara vez provee un ajuste adecuado. Los modelos que tienen interpretaciones más simples son aquellos que no tienen interacciones de tres factores. Tales modelos están anidados dentro del modelo

$$\log m_{hijk} = \mu + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hi}^{WX} + \lambda_{hj}^{WY} + \lambda_{hk}^{WZ} + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

denotado por (*WX,WY,WZ,XY,XZ,YZ*). Para este modelo cada uno de los seis pares de variables es dependiente condicionalmente dadas las otras dos variables. La independencia condicional corresponde a la ausencia de ciertos términos de asociación de dos factores. Si $\lambda_{ij}^{XY} = 0$ para todo i y j, por ejemplo, entonces X y Y son condicionalmente independientes dentro de cada combinación de los niveles de W y Z.

Hay varios modelos que exhiben algún tipo de interacción con tres factores. En el modelo denotado por (*WXY,WZ,XZ,YZ*), por ejemplo, cada par de variables es condicionalmente dependiente, pero (dentro de cada nivel de Z) la asociación entre W y X o entre W y Y o entre X y Y varía cruzando los niveles de una de las tres variables que queda. El modelo no saturado más complejo (*WXY,WXZ,WYZ,XYZ*) tiene todas las interacciones de tres factores.

El modelo más general (el modelo saturado) para una tabla con cuatro dimensiones es (*WXYZ*) que corresponde al que tiene interacción de cuatro factores. Esto significa, por ejemplo, que la interacción de tres factores de W, X y Y varía a lo largo de los niveles de Z; es decir, la forma en la cual la asociación de X y Y varía cruzando los niveles de W es por sí misma diferente dentro de las diferentes categorías de Z.

Ejemplo

Alcohol, Cigarro y Mariguana. La tabla 2.6 contiene una tabla de contingencia de $2 \times 2 \times 2$ cuyos datos se obtuvieron de la encuesta de 1992 realizada por la *Wright State University School of Medicine* y la *United Health Services* en Dayton, Ohio, ver Agresti (2002). La encuesta preguntó a 2276 estudiantes que se encontraban en su año final de preparatoria en zonas rurales cerca de Deaton, Ohio, si alguna vez habían utilizado alcohol, cigarro, o mariguana. Las variables se denotan por A para alcohol, C para cigarro, y M para mariguana.

		Mariguana	
Alcohol	Cigarro	Sí	No
Sí	Sí	911	538
Sí	No	44	456
No	Sí	3	43
No	No	2	279

Tabla 2.6: Cigarro, mariguana y alcohol.

Para estos datos ajustamos el modelo saturado y el modelo de independencia.

El modelo saturado es

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{jk}^{CM} + \lambda_{ijk}^{ACM}$$

y se denota por (ACM). En la tabla 2.7 se presentan los valores observados, los valores esperados, los residuales de Pearson, los residuales ajustados y la devianza para el modelo saturado.

			Observados		Esperados		Residuales	Residuales	
A	C	M	Conteos	%	Conteos	%	de Pearson	Ajustados	Devianza
Sí	Sí	Sí	911	40.0 %	911	40.0 %	0	0	0
Sí	Sí	No	538	23.6 %	538	23.6 %	0	0	0
Sí	No	Sí	44	2.0 %	44	2.0 %	0	.	0
Sí	No	No	456	20.0 %	456	20.0 %	0	0	0
No	Sí	Sí	3	0.2 %	3	0.2 %	0	0	0
No	Sí	No	43	1.9 %	43	1.9 %	0	.	0
No	No	Sí	2	0.1 %	2	0.1 %	0	0	0
No	No	No	279	12.3 %	279	12.3 %	0	0	0

Tabla 2.7: Modelo Saturado (ACM).

Por ser este el modelo saturado, donde se encuentran todas las posibles asociaciones, tenemos un ajuste perfecto, y por esto los valores de los residuales y la devianza son cero.

El modelo de independencia es

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^C + \lambda_k^M$$

y se denota por (A, C, M). En la tabla 2.8 se presentan los valores observados, los valores esperados, los residuales de Pearson, los residuales ajustados y la devianza para el modelo de independencia.

Este modelo de independencia es el modelo más simple, debido a que supone independencia de todas las variables, y por tanto no presente asociación

			Observados		Esperados		Residuales	Residuales	
A	C	M	Conteos	%	Conteos	%	de Pearson	Ajustados	Devianza
Sí	Sí	Sí	911	40.0 %	540.0	23.7 %	16.0	30.5	14.5
Sí	Sí	No	538	23.6 %	740.2	32.5 %	−7.4	−16.1	−7.8
Sí	No	Sí	44	1.9 %	282.1	12.4 %	−14.2	−21.2	−17.7
Sí	No	No	456	20.0 %	386.7	17.0 %	3.5	5.9	3.4
No	Sí	Sí	3	0.1 %	90.6	4.0 %	−9.2	−11.4	−12.4
No	Sí	No	43	1.9 %	124.2	5.5 %	−7.3	−9.8	−8.4
No	No	Sí	2	0.1 %	47.3	2.1 %	−6.6	−7.4	−8.8
No	No	No	279	12.3 %	64.9	2.9 %	26.6	31.2	19.6

Tabla 2.8: Modelo de Independencia (A, C, M).

de dos o tres variables; no hay un buen ajuste de las variables y esto se ve reflejado en los valores de los residuales y de la devianza, ya que estos valores son grandes y, como se había comentado anteriormente (ver secciones 2.3.4 y 2.3.3), esto implica que existe asociación entre algunas variables.

La tabla 2.4.3 muestra las estadísticas de bondad de ajuste para varios modelos ajustados a estos datos, el modelo, distinto al modelo saturado, que mejor se ajusta es (AC, AM, CM).

Modelo	G^2	X^2	gl	valor-p
(A,C,M)	1286.0	1411.4	4	< 0.001
(A,CM)	534.2	505.6	3	< 0.001
(C,AM)	939.6	824.2	3	< 0.001
(M,AC)	843.8	704.9	3	< 0.001
(AC,AM)	497.4	443.8	2	< 0.001
(AC,CM)	92.0	80.8	2	< 0.001
(AM,CM)	187.8	177.6	2	< 0.001
(AM,CM)	0.4	0.4	1	0.54
(ACM)	0.0	0.0	0	—

Tabla 2.9: Pruebas de bondad de ajuste para modelos loglineales de los datos de la tabla 2.6.

■

2.4.4. Modelos Loglineales Generalizados

Sean $\mathbf{n} = (n_1, \ldots, n_N)'$ y $\mathbf{m} = (m_1, \ldots, m_N)'$ vectores columna de los valores observados y los valores esperados para una tabla de contingencia de N celdas. Para simplificar los cálculos sólo usamos un subíndice, pero la tabla puede ser multidimensional. Los modelos loglineales para valores

positivos con distribución Poisson tienen la forma de

$$\log \mathbf{m} = \mathbf{X}\beta \tag{2.6}$$

para la matriz de diseño $\mathbf{X}$ y el vector columna de parámetros β.

Como ejemplo, ilustremos lo anterior con el modelo de independencia para una tabla de contingencia de 2×2, $\log m_{ij} = \mu + \lambda_i^x + \lambda_j^y$. Las condiciones adicionales sobre los parámetros pueden ser $\lambda_1^x = \lambda_1^y = 0$. Entonces podemos representar al modelo como

$$\begin{bmatrix} \log m_{11} \\ \log m_{12} \\ \log m_{21} \\ \log m_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \lambda_2^x \\ \lambda_2^y \end{bmatrix}.$$

Una generalización del modelo (2.6) permite muchos modelos adicionales. Este *modelo loglineal generalizado* es

$$\mathbf{C} \log(\mathbf{Am}) = \mathbf{X}\beta \tag{2.7}$$

para algunas matrices $\mathbf{C}$ y $\mathbf{A}$. El modelo loglineal ordinario (2.6) es un caso particular del modelo loglineal generalizado (2.7) cuando las matrices $\mathbf{C}$ y $\mathbf{A}$ son matrices identidad. Para mayores detalles, ver Agresti (2002).

2.4.5. Ajuste de Modelos Loglineales

En esta sección obtendremos las estimaciones de los parámetros de los modelos loglineales por dos métodos: una estimación directa y un algoritmo iterativo; ambos presentados en Agresti (1984, 1996, 2002). Para la estimación directa de los parámetros primero obtenemos las estadísticas suficientes y posteriormente las ecuaciones de verosimilitud, considerando una distribución Poisson (los resultados son los mismos para la distribución multinomial, ver sección 1.1.1). El algoritmo iterativo que se plantea aquí es el método de Newton-Raphson.

Estadísticas Suficientes Minimales

Consideremos una tabla de contingencia de tres dimensiones cuyos conteos de celdas $\{Y_{ijk} = n_{ijk}\}$ tienen una distribución Poisson

$$\prod_i \prod_j \prod_k \frac{e^{-m_{ijk}} m_{ijk}^{n_{ijk}}}{n_{ijk}!},$$

donde el producto se refiere a todas las celdas de la tabla. El logaritmo de la verosimilitud es

$$L(\mathbf{m}) = \sum_i \sum_j \sum_k n_{ijk} \log m_{ijk} - \sum_i \sum_j \sum_k m_{ijk}. \tag{2.8}$$

Por el modelo loglineal general (2.5), esto se puede expresar como

$$\begin{aligned} L(\mathbf{m}) &= n\mu + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\ &+ \sum_i \sum_j n_{ij+}\lambda_{ij}^{XY} + \sum_i \sum_k n_{i+k}\lambda_{ik}^{XZ} + \sum_j \sum_k n_{+jk}\lambda_{jk}^{YZ} \\ &+ \sum_i \sum_j \sum_k n_{ijk}\lambda_{ijk}^{XYZ} - \sum_i \sum_j \sum_k \exp(\mu + \cdots + \lambda_{ijk}^{XYZ}). \end{aligned} \tag{2.9}$$

Como la distribución Poisson pertenece a la familia exponencial, los coeficientes de los parámetros son estadísticas suficientes. Para este modelo saturado, los conteos observados $\{n_{ijk}\}$ son coeficientes de $\{\lambda_{ijk}\}$, de tal manera que no hay una reducción de los datos. Para otros modelos no saturados, algunos parámetros son cero y la ecuación (2.9) se simplifica. La tabla 2.10 lista las estadísticas suficientes minimales para varios modelos loglineales. Cada uno es coeficiente de los términos de mayor jerarquía en los cuales aparece una variable.

Modelo	Estadísticas Suficientes Minimal
(X, Y, Z)	$\{n_{i++}\}$, $\{n_{+j+}\}$, $\{n_{++k}\}$
(XY, Z)	$\{n_{ij+}\}$, $\{n_{++k}\}$
(XY, YZ)	$\{n_{ij+}\}$, $\{n_{+jk}\}$
(XY, XZ, YZ)	$\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$

Tabla 2.10: Estadísticas suficientes minimales para algunos modelos loglineales.

Ecuaciones de Verosimilitud para Modelos Loglineales

Los valores ajustados para un modelo son soluciones para las ecuaciones de verosimilitud. Obtenemos las ecuaciones de verosimilitud usando la representación general (2.6) para un modelo loglineal. Para un vector de conteos $\mathbf{n}$ con frecuencias esperadas $\mathbf{m} = E(\mathbf{n})$, el modelo es $\log \mathbf{m} = \mathbf{X}\beta$, para el cual $\log(m_i) = \sum_j x_{ij}\beta_j$ para todo i.

Extendiendo (2.8) para conteos con distribución Poisson el logaritmo de la verosimilitud es

$$\begin{aligned} L(\mathbf{m}) &= \sum_i n_i \log m_i - \sum_i m_i \\ &= \sum_i n_i \left(\sum_j x_{ij}\beta_j \right) - \sum_i \exp\left(\sum_j x_{ij}\beta_j \right). \end{aligned} \tag{2.10}$$

La estadística suficiente para β_j es su coeficiente $\sum_i n_i x_{ij}$. Puesto que

$$\begin{aligned} \frac{\partial}{\partial \beta_j}\left[\exp\left(\sum_j x_{ij}\beta_j \right) \right] &= n_{ij} \exp\left(\sum_j x_{ij}\beta_j \right) = x_{ij}\mu_i, \\ \frac{\partial L(\mathbf{m})}{\partial \beta_j} &= \sum_i n_i x_{ij}, \qquad j = 1, \ldots, p. \end{aligned}$$

Las ecuaciones de verosimilitud se obtienen al igualar estas expresiones a cero y tienen la forma

$$\mathbf{X}'\mathbf{n} = \mathbf{X}'\hat{\mathbf{m}}.$$

Estas ecuaciones igualan las estadísticas suficientes a sus valores esperados.

Método de Newton-Raphson

Sea $f : \Re^d \to \Re$ una función positiva, doblemente diferenciable, con al menos un máximo local. El método de Newton-Raphson proporciona un algoritmo para encontrar el máximo de f. Es un método iterativo basado en una aproximación de Taylor de segundo orden a $G(x) = \log f(x)$.

Algoritmo:

1. Elegir un valor inicial x^0.
2. Para $t = 1, 2, \ldots$
   (a) Calcular $G'(x^{t-1})$ y $G''(x^{t-1})$. (El algoritmo se basa en una aproximación cuadrática a $G(x)$ centrada en x^{t-1}.)
   (b) Calcular el nuevo valor, x^t, como
   $$x^t = x^{t-1} - [G''(x^{t-1})]^{-1} G'(x^{t-1}).$$
   (De hecho, x^t maximiza la aproximación cuadrática mencionada en (a).)

La elección del valor inicial, x^0, es importante. No hay garantía de que el algoritmo converja para todos los valores iniciales, particularmente en regiones donde $-G''$ no es definida positiva. Por otro lado, el algoritmo converge rápidamente si el valor inicial está cerca del máximo.

Identificamos a $L(\beta)$ como el logaritmo de la verosimilitud de modelos loglineales cuyos conteos tienen una distribución Poisson.

A partir de la ecuación (2.10), considere

$$L(\beta) = \sum_i n_i \left(\sum_h x_{ih}\beta_h \right) - \sum_i \exp\left(\sum_h x_{ih}\beta_h \right).$$

Entonces

$$\begin{aligned} u_j &= \frac{\partial L(\beta)}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i m_i x_{ij}, \\ h_{jk} &= \frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_i m_i x_{ij} x_{ik}, \end{aligned}$$

así que

$$u_j^{(t)} = \sum_i (n_i - m_i^{(t)}) x_{ij} \qquad \text{y} \qquad h_{jk}^{(t)} = -\sum_i m_i^{(t)} x_{ij} x_{ik}.$$

La t-ésima aproximación $\mathbf{m}^{(t)}$ para $\hat{\mathbf{m}}$ se obtiene de $\beta^{(t)}$ a través de $\mathbf{m}^{(t)} = \exp(\mathbf{X}\beta^{(t)})$. Esto genera el siguiente valor $\beta^{(t)}$ dado por

$$\beta^{(t+1)} = \beta^{(t)} + \left[\mathbf{X}' \text{diag}(\mathbf{m}^{(t)}) \mathbf{X} \right]^{-1} \mathbf{X}'(\mathbf{n} - \mathbf{m}^{(t)}).$$

Esto produce $\mathbf{m}^{(t+1)}$, y así sucesivamente.

Capítulo 3

Análisis Bayesiano

En este capítulo estudiaremos algunos de los métodos que se utilizan para analizar variables categóricas usando métodos de estadística Bayesiana.

En la sección 3.1 se presenta la notación generalizada de tablas de contingencia que se usará en este capítulo para el análisis de datos categóricos. En la sección 3.2 se presentan tres pruebas Bayesianas para analizar tablas de contingencia. En la sección 3.3 estudiamos los modelos loglineales desde el punto de vista Bayesiano, y se presentan una serie de ejemplos usando el *software* WinBUGS. En la sección 3.4 se describen algunos modelos alternativos utilizados para el análisis de datos categóricos. Finalmente, en la sección 3.5 se presenta el criterio de información de la devianza (DIC) como un indicador para la selección de modelos.

3.1. Introducción

Considere una tabla de contingencia de dos dimensiones con r renglones y c columnas. Como en el capítulo anterior, sean n_{ij} y π_{ij} el número de observaciones y las probabilidades de celdas correspondientes al renglón i y la columna j ($i = 1, \ldots, r;\ j = 1, \ldots, c$), respectivamente. Sea $n_{i+} = \sum_j n_{ij}$, $n_{+j} = \sum_i n_{ij}$ y $N = \sum_i \sum_j n_{ij}$.

Sea $m = rc$ el número total de celdas en la tabla de contingencia $r \times c$. Algunas veces encontraremos que es más conveniente ordenar las observaciones y las probabilidades en un vector de orden $1 \times m$. Sea $\tilde{\pi}_l$ y $\tilde{n}_l$ respectivamente, la probabilidad y el número de observaciones de la celda l ($l = 1, \ldots, m$) y denotemos por π tanto a $(\tilde{\pi}_1, \ldots, \tilde{\pi}_m)$ como a $(\pi_{11}, \ldots, \pi_{rc})$ arregladas en el mismo orden lexicográfico. Similarmente, denotemos por $\mathbf{n}$ a $(\tilde{n}_1, \ldots, \tilde{n}_m)$ y a $(n_{11}, \ldots, n_{rc})$.

Bajo el esquema de muestreo multinomial (ver sección 2.3.1), el vector

de conteos de las observaciones $\mathbf{n}$, se considera como una observación de una distribución multinomial $(m-1)$-dimensional con índice $N = \sum_l \tilde{n}_l$ y vector de parámetros desconocido π, cuya función de probabilidad es

$$f(\mathbf{n}|\pi, N) = \frac{N!}{\prod_l \tilde{n}_l!} \prod_l \tilde{\pi}_l^{\tilde{n}_l},$$

donde $\tilde{\pi}_l > 0$ y $\sum_l \tilde{\pi}_l = 1$.

3.2. Tablas de Contingencia

Al analizar las tablas de contingencia es importante estudiar la relación de las variables contenidas en la tabla. En esta sección se presentan tres pruebas, la primera para analizar la independencia a través de una prueba de significancia análoga a la devianza propuesta por Gutiérrez Peña (2005) y la segunda a través de los factores de Bayes propuesta por Good (1976). Finalmente, una prueba propuesta por Lindley (1964) en donde se obtiene una distribución aproximada de los parámetros y una prueba de significancia semejante al análisis de varianza clásico.

3.2.1. Pruebas de Independencia

Cuando las pruebas de hipótesis se realizan con respecto a las probabilidades de las celdas o a las frecuencias en una tabla de contingencia, la hipótesis nula impone ciertas restricciones sobre el espacio de posibles valores de π. En otras palabras, bajo la hipótesis nula, las probabilidades de celda están dadas por $\tilde{\pi}_l^0 = h_l(\pi)$ para algunas funciones $h_l(\cdot)$, $l = 1, \ldots, m$. Consideremos, por ejemplo, una tabla de contingencia $r \times c$ y un modelo nulo bajo el cual las dos variables son independientes. En este caso, es conveniente usar la notación con doble índice para referirnos a las probabilidades de celdas o conteos individuales. Entonces la hipótesis nula es

$$H_0 : \pi_{ij}^0 = h_{ij}(\pi) \equiv \pi_{i+}\pi_{+j},$$

donde $\pi_{i+} = \sum_j \pi_{ij}$ y $\pi_{+j} = \sum_i \pi_{ij}$ $(i = 1, \ldots, r;\ j = 1, \ldots, c)$.

Una Prueba de Significancia

Como sabemos que la distribución final de π es una distribución Dirichlet (ver sección 1.1.1), podemos, en principio, calcular la probabilidad final de cualquier evento que involucre a las probabilidades de celdas π. En particular,

la distribución final de π induce una distribución final sobre el vector $\pi^0 = (\pi^0_{11}, \ldots, \pi^0_{rc})'$ de probabilidades de celda bajo la hipótesis nula.

El modelo nulo de independencia puede probarse sobre la base de la distribución final de

$$\delta = \delta(\pi) \equiv \sum_l \log\left(\frac{\tilde{\pi}_l}{\tilde{\pi}^0_l}\right)\log(\tilde{\pi}_l).$$

Esta cantidad puede considerarse como una versión Bayesiana de la devianza; es siempre no negativa y es cero si y sólo si el modelo nulo y el modelo saturado son el mismo, es decir, si y sólo si $\pi^0_l = \pi_l$ para toda l.

La distribución final marginal de δ no es fácil de obtener en forma analítica, pero puede obtenerse a partir de la distribución de π usando técnicas de Monte Carlo. En este caso, podemos generar una muestra $\{\pi^{(1)}, \ldots, \pi^{(M)}\}$ de tamaño M de la distribución final (Dirichlet) de π. Después calculamos $\delta^{(k)} = \delta(\pi^{(k)})$ para cada $k = 1, \ldots, M$. Los valores resultantes $\{\delta^{(1)}, \ldots, \delta^{(M)}\}$ constituyen una muestra de la distribución final marginal de δ. La precisión de las técnicas de Monte Carlo es mayor conforme el valor de M aumenta.

Las distribuciones finales de δ concentradas alrededor del cero apoyan el modelo nulo, mientras que las distribuciones finales localizadas lejos del cero conducen a rechazar el modelo nulo.

Podemos probar la hipótesis nula de independencia por medio de una "prueba de significancia Bayesiana": rechazar la hipótesis nula si el intervalo de mayor densidad (digamos del 95 %), relativo a la densidad final para δ, no contiene el valor cero (ver sección 3.2.2).

Factores de Bayes

Otra forma de hacer pruebas de hipótesis de independencia en una tabla de contingencia es usando el factor de Bayes, ver Good (1967) y Good (1976).

Para probar independencia por medio de métodos Bayesianos es necesario suponer densidades iniciales para las probabilidades (π_{ij}). Una vez que se eligen las distribuciones iniciales es posible calcular $p((n_{ij})|H_1)$ y $p((n_{ij})|H_0)$ y el cociente de estas probabilidades es el *factor de Bayes en contra de* H_0. Este factor puede verse como un cociente de densidades bajo el modelo establecido.

Cuando suponemos la hipótesis nula H_0 (independencia) adoptamos las densidades Dirichlet $D^*(r, 1)$ y $D^*(s, 1)$ (ver apéndice A) como las densidades iniciales de $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$. Para la hipótesis alternativa H_1 suponemos la densidad Dirichlet $D^*(rs, 1)$ como la densidad inicial de $\{\pi_{ij}\}$.

Para cada uno de los esquemas de muestreo para una tabla de contingencia de dos dimensiones (ver sección 2.3.1) tenemos distintos factores de

Bayes:

- Esquema 1, el total N es fijo; el factor de Bayes es F_1, dado por

$$F_1 = \frac{p\left((n_{ij})|H_1\right)}{p\left((n_{ij})|H_0\right)}.$$

- Esquema 2, el total por renglones $\{n_{i+}\}$ o el total por columnas $\{n_{+j}\}$ es fijo; el factor de Bayes es F_2, dado por

$$F_2 = \frac{p\left((n_{ij})|(n_{i+}), H_1\right)}{p\left((n_{ij})|(n_{i+}), H_0\right)},$$

  para el total por renglones, o

$$F_{(2)} = \frac{p\left((n_{ij})|(n_{+j}), H_1\right)}{p\left((n_{ij})|(n_{+j}), H_0\right)}$$

  para el total por columnas.

- Esquema 3, el total por renglones y total por columnas son fijos; el factor de Bayes es F_3, dado por

$$F_3 = \frac{p\left((n_{ij})|(n_{i+}), (n_{+j}), H_1\right)}{p\left((n_{ij})|(n_{i+}), (n_{+j}), H_0\right)}.$$

Cuando la hipótesis de independencia es falsa entonces $\log(F)/N$ tiende en probabilidad a un límite, donde F es el factor de Bayes correspondiente a cualquiera de los tres esquemas. Este límite es una medida de asociación natural que podría llamarse "el peso de la evidencia en contra de la hipótesis de independencia". Este límite fue llamado W_1 por Good (1976) y se puede interpretar como la "información mutua esperada entre renglones y columnas". Bajo cualquiera de los tres esquemas de muestreo

$$W_1 = \sum \pi_{ij} \log \left\{ \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}} \right\}.$$

La medida de asociación correspondiente para tres dimensiones es

$$\sum \pi_{hij} \log \left\{ \frac{\pi_{hij}}{\pi_{h++}\pi_{+i+}\pi_{++j}} \right\},$$

y de manera análoga podemos obtener esta medida de asociación para mayores dimensiones.

$\log_{10}(F)$	F	Evidencia contra H_0
0 a 1/2	1 a 3.2	No hay evidencia
1/2 a 1	3.2 a 10	Hay evidencia substancial
1 a 2	10 a 100	Hay gran evidencia
>2	>100	Hay evidencia decisiva

Tabla 3.1: Interpretación del factor de Bayes.

$2\log_e(F)$	F	Evidencia contra H_0
0 a 2	1 a 3	No hay evidencia
2 a 6	3 a 20	Hay evidencia positiva
6 a 10	20 a 150	Hay gran evidencia
>10	>150	Hay demasiada evidencia

Tabla 3.2: Interpretación del factor de Bayes.

Kass y Raftery (1995) mencionan que si el factor de Bayes es "grande" entonces se rechaza la hipótesis nula de independencia, sin embargo, para poder comparar el factor de Bayes de una mejor manera, pueden considerarse los criterios de la tabla 3.1. Sin embargo, consideran que los criterios presentados en la tabla 3.2 también pueden utilizarse para realizar una prueba de hipótesis, ya que además éstos están definidos en la misma escala que la devianza (sección 2.3.4) y la estadística de prueba de razón de verosimilitudes (sección 2.3.2).

3.2.2. Otras Pruebas

Como ya se ha mencionado, si las variables n_i $(i = 1, \ldots, k)$ son variables aleatorias independientes con distribución Poisson con medias μ_i, entonces la distribución condicional de estas variables dada $N = \sum_i n_i$ tiene una distribución multinomial con parámetros $\pi_i = \mu_i / \sum_i \mu_i = \mu_i/\theta$ con $\theta = \sum_i \mu_i$. Además la distribución final de π es proporcional a $\prod_i \pi_i^{n_i - 1}$ cuando consideramos una distribución inicial no informativa (ver sección 1.1.1).

Lo que nos interesa es encontrar una aproximación de $\prod_i \pi_i^{n_i-1}$. Para encontrarla consideremos la distribución Poisson. La distribución final de las μ_i, dadas las n_i (variables aleatorias Poisson) es proporcional a

$$\prod_i (e^{-\mu_i} \mu_i^{n_i - 1}),$$

cuando se supone una distribución inicial no informativa. Las μ_i son indepen-

dientes y tienen una distribución gamma (ver sección 1.1.1). Si las variables tienen esta distribución, entonces su logaritmo se distribuye aproximadamente normal para todos los n_i que no sean pequeños. La media de los $\log \mu_i$ es aproximadamente $\log n_i$ y la varianza es aproximadamente n_i^{-1}.

Sea $a_1, \ldots, a_k$ un conjunto de constantes con $\sum_i a_i = 0$. A una combinación lineal cuyos coeficientes suman cero generalmente se le llama *contraste*. Considere un contraste en los $\log \mu_i$, y note que

$$\sum_i a_i \log \mu_i = \sum_i a_i \log(\theta \pi_i) = \sum_i a_i \log \pi_i.$$

Esto implica que los contrastes de los $\log \mu_i$ son iguales a los contrastes de los $\log \pi_i$. Pero los contrastes de los $\log \mu_i$ se distribuyen aproximadamente normal y el conjunto de ellos se distribuyen normal conjuntamente. Lo mismo debe aplicarse a los $\log \pi_i$ cuando se considera que los n_{ij} tienen una distribución multinomial. De estos resultados tenemos el siguiente teorema.

Teorema 3.2.1. *Si las variables aleatorias $n_1, \ldots, n_k$ tienen una distribución multinomial con parámetros $\pi_1, \ldots, \pi_k$; y si la distribución inicial de los parámetros π_i tienen densidad proporcional a $\prod_{i=1}^k \pi_i^{-1}$ en la región $\pi_i \geq 0$, $\sum_i \pi_i = 1$; entonces, si las constantes a_{pi} ($p = 1, \ldots, m$; $i = 1, \ldots, k$; $m < k$) satisfacen $\sum_i a_{pi} = 0$, la distribución final conjunta de $\sum_i a_{pi} \log \pi_i$ ($p = 1, \ldots, m$) es aproximadamente normal con medias*

$$\sum_i a_{pi} \log n_i$$

y covarianzas (varianzas cuando $p = q$)

$$\sum_i a_{pi} a_{qi} n_i^{-1}.$$

Las expresiones para la media y la covarianza de los contrastes se obtienen debido a la independencia de los $\log \mu_i$ y a la forma aproximada de sus medias y varianzas. En el caso binomial, $k = 2$, los contrastes son múltiplos de $\log \pi_1 - \log \pi_2 = \log\{\pi/(1-\pi)\}$, y éstos son los logaritmos de los momios. Consecuentemente, el resultado para las variables que tienen una distribución binomial es un caso particular del teorema 3.2.1. En vista de la generalización, llamaremos a un contraste de los $\log \pi_i$ un "contraste de los logaritmos de los momios".

De esta manera podremos obtener una aproximación de la distribución final del logaritmo de los momios, a través de los contrastes de los $\log \pi_i$ y obtener su distribución final aproximada dada por el teorema 3.2.1. Muchos

de los parámetros de interés en el análisis de datos multinomiales, particularmente aquellos arreglados en tablas de contingencia, pueden expresarse como logaritmos de momios, de tal manera que esta aproximación de la distribución final será de gran utilidad.

A continuación obtendremos una "prueba de significancia" Bayesiana basada en intervalos de mayor densidad de la distribución final del parámetro. Este es un resultado análogo al análisis de varianza clásico y simplifica el análisis (Lindley, 1964).

Vamos a elegir intervalos (únicos) que tienen la propiedad de que ningún valor dentro del intervalo tiene una menor densidad que cualquier valor fuera del intervalo. Estos son los intervalos de longitud más corta. Estos intervalos serán equivalentes a una prueba de significancia clásica.

Aplicaremos estas ideas a las distribuciones finales de los logaritmos de los momios. Sean $\phi_1, \ldots, \phi_s$ los logaritmos de momios linealmente independientes con medias m_i y covarianzas v_{ij}. (Las medias y las covarianzas fueron dadas en el teorema 3.2.1.) La densidad conjunta de los ϕ_i es constante sobre las elipsoides

$$\sum_{i=1}^{s}\sum_{j=1}^{s}(\phi_i - m_i)v^{ij}(\phi_j - m_j) = c, \tag{3.1}$$

donde v^{ij} son los elementos de la inversa de la matriz de varianzas-covarianzas, y c es cualquier constante positiva. El lado izquierdo de la ecuación (3.1) tiene una distribución aproximada χ^2 con s grados de libertad. Si $\chi^2_{1-\alpha}$ es el cuantil $100(1-\alpha)\,\%$ de esta distribución, la probabilidad final de

$$\sum_{i=1}^{s}\sum_{j=1}^{s}(\phi_i - m_i)v^{ij}(\phi_j - m_j) \leq \chi^2_{1-\alpha}, \tag{3.2}$$

es $(1-\alpha)$. Si los valores $\phi_i = \phi_i^*$ son de interés, serían inverosímiles si no satisfacen (3.2) con $\phi_i = \phi_i^*$ $(i = 1, \ldots, s)$. Esto nos da una prueba de significancia de la hipótesis $H_0 : \phi_i = \phi_i^*$. Si realizamos la prueba de significancia de la hipótesis nula $\phi_i = 0$ $(i = 1, \ldots, s)$ la estadística se reduce a

$$\sum_{i=1}^{s}\sum_{j=1}^{s} m_i v^{ij} m_j \tag{3.3}$$

y la podemos comparar con el cuantil apropiado de una distribución χ^2.

En las ecuaciones (3.2) y (3.3) se considera que $\{m_i\}$ y $\{v^{ij}\}$ son estadísticas y $\{\phi_i\}$ son los parámetros, o sea, que los $\{\phi_i\}$ son las variables aleatorias. El argumento usual es considerar que las $\{m_i\}$ son variables aleatorias que

se distribuyen aproximadamente normal con medias $\{\phi_i\}$ y con covarianzas $\{v_{ij}\}$ (conocidas). En particular la comparación de la ecuación (3.3) con la distribución χ^2 es básica para el análisis de varianza. Consecuentemente al analizar el logaritmo de los momios por los métodos Bayesianos descritos, tenemos a nuestra disposición los métodos de análisis de varianza.

Los métodos basados en las ecuaciones (3.2) y (3.3) son invariantes bajo transformaciones lineales de los logaritmos de los momios. Esto se debe a que las formas cuadráticas son invariantes. Esto nos da la posibilidad de reemplazar cualquier conjunto particular de logaritmos de momios por transformaciones lineales de tal manera que puedan simplificar el análisis.

Considere el caso donde n_i $(i = 1, \ldots, k)$ dado $N = \sum_{i=1}^{k} n_i$ se distribuyen multinomial y sea la hipótesis nula $H_0 : \pi_i = \pi_i^*$ $(i = 1, \ldots, k)$. Vamos a encontrar una prueba de significancia de la hipótesis nula usando los resultados de la distribución final de los logaritmos de los momios y el análisis Bayesiano descrito anteriormente. La densidad final de $\log \mu_i$ (considerando la distribución Poisson) es proporcional a

$$\exp\left\{-\frac{1}{2}\sum_{i=1}^{k}(\log\mu_i - \log n_i)^2 n_i\right\}, \tag{3.4}$$

usando la aproximación del teorema 3.2.1. La ecuación (3.4) puede escribirse de la siguiente manera: sea $u_i = \log\mu_i - \log n_i$, cuya distribución es aproximadamente normal con medio cero y varianza n_i^{-1}; entonces la ecuación (3.4) es

$$\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{k} n_i(u_i - \bar{u})^2 + N\bar{u}^2\right]\right\}$$

con $\bar{u} = \sum n_i u_i / \sum n_i$. Hacemos una transformación lineal ortogonal de las u_i a variables nuevas (una de las cuales es un múltiplo de $\bar{u}$). De esta manera la distribución final de los logaritmos de momios es proporcional a $\exp\{-\frac{1}{2}\sum n_i(u_i - \bar{u})^2\}$ y la probabilidad final de $\sum n_i(u_i - \bar{u})^2 \leq \chi^2_{1-\alpha}$ es $(1-\alpha)$, dada una distribución χ^2 con $(k-1)$ grados de libertad. La prueba de hipótesis nula será significativa al nivel $(1-\alpha)$ si los valores de u_i bajo la hipótesis nula no pertenece al conjunto definido arriba.

Consideremos el caso de los datos en una tabla de contingencia de dos dimensiones. Sea A_i la clase correspondiente al renglón i $(i = 1, \ldots, r)$ y sea B_j la clase correspondiente a la columna j $(j = 1, \ldots, c)$; estas clases son excluyentes y exhaustivas para las variables renglón y columna, respectivamente. Sea π_{ij} la probabilidad de pertenecer a la clase A_i y B_j. Si N muestras se clasifican independientemente, los números n_{ij} de muestras pertenecientes a las clases A_i y B_j se distribuyen binomial con parámetros N

y π_{ij}. En tablas de contingencia algunas veces es conveniente considerar las probabilidades de otra manera. Por ejemplo, las probabilidades de las clases A_i, π_{i+}, y las probabilidades condicionales de las clases B_j dada A_i, π_{ij}/π_{i+}. Las inferencias acerca de estas probabilidades necesitan la distribución inicial de los parámetros, y para esto es el siguiente teorema.

Teorema 3.2.2. *Si la distribución inicial de* π_{ij} $(i = 1, \ldots, r;\ j = 1, \ldots, c)$ *es proporcional a* $\prod_i \prod_j \pi_{ij}^{-1}$, *entonces la distribución inicial de* π_{i+} *y* $\phi_{ij} = \pi_{ij}/\pi_{i+}$ $(i = 1, \ldots, r;\ j = 1, \ldots, c)$ *es proporcional a*

$$\prod_{i=1}^{r} \pi_{i+}^{-1} \prod_{i=1}^{r} \prod_{j=1}^{c} \phi_{ij}^{-1}.$$

El teorema establece que la distribución inicial de las rc clases es consistente con la misma distribución inicial sobre el número reducido r de clases.

Considere una tabla de contingencia de 2×2. Una parametrización conveniente es a través de $\pi_{1+} = \pi_{11} + \pi_{12}$, la probabilidad de la clase A_1, y ϕ_{11} y ϕ_{21}, las probabilidades de la clasificación de B_1 dada A_1 y A_2 respectivamente. Combinando los resultados de los dos teoremas anteriores veremos que las distribuciones finales son aproximadamente las siguientes:

(a) $\log\{\pi_{1+}/\pi_{2+}\}$ es normal con media $\log\{n_{1+}/n_{2+}\}$ y varianza $n_{1+}^{-1} + n_{2+}^{-1}$;

(b) $\log\{\phi_{11}/\phi_{12}\}$ es normal con media $\log\{n_{11}/n_{12}\}$ y varianza $n_{11}^{-1} + n_{12}^{-1}$;

(c) $\log\{\phi_{21}/\phi_{22}\}$ es normal con media $\log\{n_{21}/n_{22}\}$ y varianza $n_{21}^{-1} + n_{22}^{-1}$;

y estas tres variables son independientes. Estos resultados se obtienen porque en cada caso estamos tratando con una situación binomial. Los resultados se obtienen de manera análoga al intercambiar las clases A y B.

Si las clasificaciones de renglones y columnas son independientes $\pi_{ij} = \pi_{i+}\pi_{+j}$; además $\phi_{11} = p(B_1|A_1) = \pi_{11}/(\pi_{11} + \pi_{12}) = \pi_{21}/(\pi_{21} + \pi_{22}) = p(B_1|A_2) = \phi_{12}$ y puede escribirse como

$$\frac{\pi_{11}}{\pi_{12}} = \frac{\pi_{21}}{\pi_{22}} \qquad \text{ó} \qquad \frac{\phi_{11}}{\phi_{12}} = \frac{\phi_{21}}{\phi_{22}},$$

es decir, los momios para la clasificación de B son los mismos dentro de A_1 y A_2. Un posible parámetro a considerar es el logaritmo de momios

$$\begin{aligned} \phi &= \log \pi_{11} - \log \pi_{21} - \log \pi_{12} + \log \pi_{22} \\ &= \log \phi_{11} - \log \phi_{21} - \log \phi_{12} + \log \phi_{22}. \end{aligned}$$

Por (b) y (c) de la combinación de los teoremas, este parámetro se distribuye aproximadamente normal con media

$$\log n_{11} - \log n_{21} - \log n_{12} + \log n_{22}$$

y varianza

$$n_{11}^{-1} + n_{21}^{-1} + n_{12}^{-1} + n_{22}^{-1}.$$

La hipótesis nula de independencia es $\phi = 0$ y puede probarse haciendo uso de que

$$\frac{(\log n_{11} - \log n_{21} - \log n_{12} + \log n_{22})^2}{n_{11}^{-1} + n_{21}^{-1} + n_{12}^{-1} + n_{22}^{-1}}$$

tiene una distribución aproximadamente χ^2 con un grado de libertad. Ésta coincide con la distribución asintótica clásica.

Cuando tenemos una tabla de contingencia de orden $r \times c$, donde r y c son mayores a 2, puede ser difícil generalizar este análisis. La dificultad se presenta al intentar encontrar logaritmos de momios que sean independientes y que reflejen la hipótesis de que las dos clasificaciones son independientes. Por ejemplo, considere una prueba de hipótesis de independencia entre las clasificaciones en una tabla de contingencia de 3×3. Cuatro logaritmos de momios que al anularse serían equivalentes a la hipótesis de independencia serían

$$\begin{aligned}
&\log \pi_{11} - \log \pi_{12} - \log \pi_{21} + \log \pi_{22} \\
&\log \pi_{11} - \log \pi_{13} - \log \pi_{21} + \log \pi_{23} \\
&\log \pi_{21} - \log \pi_{22} - \log \pi_{31} + \log \pi_{32} \\
&\log \pi_{21} - \log \pi_{23} - \log \pi_{31} + \log \pi_{33}.
\end{aligned} \tag{3.5}$$

Sin embargo, estos contrastes están correlacionados; por ejemplo la covarianza entre el primero y segundo es $n_{11}^{-1} + n_{21}^{-1}$. La manera directa para analizar esto sería determinar la matriz de dispersión, A, de los logaritmos de los momios de (3.5) y los valores muestrales de los mismos contrastes. Si $\mathbf{n}$ es el vector columna de los valores muestrales, entonces la forma cuadrática relevante es $\mathbf{n}' A^{-1} \mathbf{n}$ que tiene una distribución aproximada χ^2 con 4 grados de libertad.

3.3. Modelos Loglineales

En muchas áreas de investigación, especialmente en ciencias sociales, la información disponible para los modelos estadísticos se obtiene de las muestras

registradas o de estadísticas oficiales, y algunas veces son datos de naturaleza no métrica, es decir, las variables observadas son variables categóricas o son variables que se encuentran agrupadas a pesar de la métrica que originalmente tienen. Como ya hemos visto, en estos casos el interés es estudiar modelos cuyos datos son conteos acumulados arreglados en categorías cruzadas formadas por dos o más de estas variables categóricas o variables cualitativas.

Generalmente adoptamos una transformación logarítmica de la media de la distribución Poisson para modelar tales datos, con el fin de asegurar que los conteos predictivos sean positivos. Aunque los datos sean frecuencias de distribuciones multinomiales, usamos la equivalencia de la distribución multinomial con la distribución condicional de las variables Poisson dada su suma (ver sección 1.1.1).

Algunas veces existe una variable que claramente es una variable respuesta en las tablas de contingencia (o más de una variable respuesta) y el resto de las variables permanecen como variables predictoras de la variable respuesta. En otras situaciones no existe una distinción clara entre las variables predictoras y la variable respuesta, y el interés será estudiar la estructura de la tabla de contingencia considerando a todas las variables como variables respuesta.

Los modelos loglineales también pueden utilizarse cuando hay una respuesta categórica y una mezcla de variables predictoras categóricas y continuas. En estas situaciones la variable respuesta es una variable de conteo y sus variables predictoras usan un modelo lineal generalizado con función liga logarítmica, aunque en algunos casos puede ser igualmente apropiado hacerlo sin adoptar una transformación logarítmica.

En las siguientes secciones se presentan algunos ejemplos que hacen uso del *software* WinBUGS (ver sección 1.2). Los programas correspondientes se presentan en el apéndice C.

3.3.1. Tablas de Contingencia de Dos Dimensiones

Suponga que tenemos una tabla de contingencia de dos dimensiones con r categorías en los renglones y c categorías en las columnas. Sea n_{ij} los conteos de la celda (i, j) correspondiente a la categoría i de los renglones y a la categoría j de las columnas, y sea $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$ el tamaño total de la muestra.

Podemos definir la estructura del modelo de los rc conteos sin necesidad de suponer que una de las variables es la variable respuesta y la otra la variable predictora. Para los datos que son conteos, generalmente se considera un

modelo Poisson, de tal manera que

$$n_{ij} \sim Poisson(\mu_{ij})$$

donde μ_{ij} es la media de la distribución Poisson. El modelo loglineal que se ajusta a este tipo de datos usa una transformación logarítmica para la media μ_{ij} y en una tabla de dos dimensiones se especifican los cuatro tipos de influencia que existen: los niveles de los conteos, los efectos de los renglones (α_i), los efectos de las columnas (β_j), y el efecto de cada una de las combinaciones de las celdas (i, j), llamado el efecto de interacción (γ_{ij}). Un modelo alternativo sería considerar que los datos n_{ij} son observaciones con distribución multinomial con probabilidades de celda π_{ij}

$$n_{ij} \sim Multinomial(n, \pi_{ij}).$$

Los resultados bajo este supuesto son los mismos que bajo el modelo Poisson (ver sección 1.1.1).

Los efectos de los renglones (columnas) expresan la frecuencia relativa que existe en cada categoría de los renglones (categoría de las columnas): por ejemplo, un efecto de renglón (columna) grande α_i (β_j) dará una mayor frecuencia de ocurrencia de la categoría correspondiente. Estos efectos se llaman "efectos principales". Si el mejor modelo ajustado sólo contiene efectos principales entonces las variables son efectivamente independientes ya que no hay efectos de interacción entre éstas. Los efectos de interacción describen el grado de asociación que hay entre las categorías de la variable renglón y las categorías de la variable columna, estos efectos de interacción y los patrones de asociación que presentan las variables son el mayor foco de interés en el estudio de tablas de contingencia y de modelos loglineales. Un modelo que incluye todas las posibles interacciones y los efectos principales representa todas las características posibles de los datos (es decir, tienen un mejor ajuste) y se le llama modelo saturado.

El modelo loglineal saturado para una tabla de contingencia de dos dimensiones incorpora los efectos principales y los efectos de interacción:

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}. \tag{3.6}$$

El número de parámetros en el modelo saturado es $1 + r + c + rc$, mayor al número de celdas rc.

Para estimar los parámetros se deben establecer ciertas condiciones de tal manera que el número de parámetros en el modelo no exceda el número de celdas en la correspondiente tabla de contingencia. Existen dos formas de establecer estas condiciones.

El primer sistema para establecer las condiciones sobre los parámetros del modelo se llama *corner*. Este sistema fija un efecto principal de la variable renglón y un efecto principal de la variable columna, iguales a un término constante, tal como $\alpha_1 = \beta_1 = 0$. También fija el primer renglón y la primer columna de los parámetros de interacción, tal como $\gamma_{1j} = 0$ para toda j, y $\gamma_{i1} = 0$ para toda i.

El segundo sistema para establecer las condiciones sobre los parámetros del modelo se llama *centred*. Este sistema establece que los efectos principales de la variable renglón y los efectos principales de la variable columna sumen cero, de tal manera que cada parámetro es combinación lineal del resto de los parámetros para cada una de las variables, es decir $\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{c} \beta_j = 0$. También establece que los parámetros de interacción en cada renglón y en cada columna sumen cero, $\sum_{i=1}^{r} \gamma_{ij} = \sum_{j=1}^{c} \gamma_{ij} = 0$.

Por ejemplo, para una tabla de contingencia de 2×2. Bajo el sistema *corner* tenemos que $\alpha_1 = \beta_1 = 0$ y $\gamma_{11} = \gamma_{12} = \gamma_{21} = 0$. Por otro lado, bajo el sistema *centred* tenemos que $\alpha_1 = -\alpha_2$, $\beta_1 = -\beta_2$ y $\gamma_{11} + \gamma_{21} = \gamma_{12} + \gamma_{22} = \gamma_{11} + \gamma_{12} = \gamma_{21} + \gamma_{22} = 0$, de esto obtenemos que $\gamma_{11} = \gamma_{22} = -\gamma_{12} = -\gamma_{21}$.

Sujetas a fluctuaciones de muestreo, las predicciones $\{\mu_{ij}\}$ bajo un modelo saturado (3.6) reproducirán más o menos los conteos observados $\{n_{ij}\}$, dependiendo del análisis que se lleve a cabo. El análisis Bayesiano proveerá la distribución completa de cada μ_{ij}, y puede tener ligeras desviaciones de los conteos observados a partir del ajuste de las medias finales; por contraste el análisis de máxima verosimilitud produce una igualdad exacta. Un modelo saturado, o uno muy cercano al modelo saturado con varios conjuntos de interacciones incluidos, puede tener parámetros redundantes ("sobreajuste"), con algunos parámetros mal identificados.

3.3.2. Selección de Modelos y Estrategias

Hay dos grandes estrategias para la selección de modelos loglineales. La primera estrategia consiste en simplificar el modelo saturado y evaluar la bondad de ajuste de los modelos reducidos así obtenidos. Para tablas de dos dimensiones estos modelos simples son relativamente pocos, pero para tablas de mayores dimensiones la elección del modelo puede llegar a ser compleja. Por ejemplo en una tabla de contingencia de cuatro dimensiones $I \times J \times K \times L$ (por ejemplo afiliación política por sexo por edad por clase social) habría cuatro conjuntos de efectos principales, $\{\alpha_{1i}\}$, $\{\alpha_{2j}\}$, $\{\alpha_{3k}\}$, $\{\alpha_{4l}\}$; seis conjuntos correspondientes a la interacción de dos variables $\{\beta_{1ij}\}$, $\{\beta_{2ik}\}$, $\{\beta_{3il}\}$, $\{\beta_{4jk}\}$, $\{\beta_{5jl}\}$, $\{\beta_{6kl}\}$; cuatro conjuntos correspondientes a la interacción entre tres variables, $\{\gamma_{1ijk}\}$, $\{\gamma_{2ijl}\}$, $\{\gamma_{3jkl}\}$, $\{\gamma_{4ikl}\}$; y un término correspondiente a la iteracción entre las cuatro variables $\{\delta_{ijkl}\}$.

Algunas veces el primer paso al explorar un modelo, y en el procedimiento de selección, es examinar los parámetros en el modelo saturado que son claramente identificados y aquellos que son pobremente identificados. Procedimientos más complejos podrían consistir en comparar las medidas de ajuste entre un modelo determinado y el mismo modelo pero con alguna variable omitida. Por ejemplo, en una tabla de cuatro dimensiones podemos excluir el término de cuatro interacciones y dejar los términos de tres interacciones.

Aitkin (1979) propone un procedimiento de prueba simultáneo (STP) para la elección de un modelo para tablas de contingencia complejas.

La segunda estrategia para la selección de modelos loglineales extiende los parámetros que pueden expresarse en una forma que refleje las características substantivas de los datos; por ejemplo, si los renglones y las columnas están ordenados en categorías socioeconómicas o son áreas geográficas, entonces el término de interacción puede expresar efectos sociales o efectos de distancias geográficas.

En el modelo de dos dimensiones (3.6) la simplificación más obvia es suponer que no hay interacción entre la variable renglón y la variable columna. Una ausencia completa de interacción significa que las dos variables son independientes. La bondad de ajuste de varios modelos puede entonces evaluarse por medio de la elección entre el modelo saturado y el modelo de independencia. Estas opciones pueden tomar en cuenta las características substantivas del tipo de tabulación que se está analizando.

El término de interacción puede conservarse pero en una forma "intermedia"simplificada, dando lugar a los modelos de cuasi-independencia (*quasi-independence*). Por ejemplo γ_{ij} podría expresarse como el producto de un efecto de renglón y uno de columna o *scores* (note que éstos son distintos de los efectos principales de renglones y columnas), tal como

$$\gamma_{ij} = \delta_i \epsilon_j$$

así que en lugar de tener $(r-1)(c-1)$ parámetros que describen los efectos de interacción sólo hay $r + c - 2$ parámetros.

3.3.3. Ajuste para Modelos de Dos Dimensiones

Suponga que en una tabla de contingencia de dos dimensiones eliminamos completamente el término de interacción en el modelo saturado (3.6). De esta manera hay $1 + (r-1) + (c-1)$ parámetros para estimar, suponiendo que estamos aplicando un modelo de efectos fijos. El enfoque Bayesiano especifica distribuciones iniciales apropiadas para los efectos fijos incluidos en el modelo loglineal. Suponiendo una función liga logarítmica, estos efectos

podrían estar en la recta real, por lo que podríamos usar como distribuciones iniciales, por ejemplo, la distribución normal o uniforme. Si aplicamos el sistema *corner* entonces esto implica fijar el efecto del primer renglón y de la primera columna igual a cero; si se supone que los efectos principales tienen una distribución inicial normal no informativa (con una varianza grande y con media cero) entonces las distribuciones iniciales son

$$\begin{aligned} \alpha_1 &= 0, \qquad \beta_1 = 0, \\ \alpha_i &\sim N(0, 100), \qquad i = 2, \ldots, r; \\ \beta_j &\sim N(0, 100), \qquad j = 2, \ldots, c. \end{aligned}$$

Ejemplo. *Movilidad social*

Suponga que dos variables en una tabla de clasificación cruzada miden el estatus de los padres y el estatus de los hijos usando el mismo sistema de rango del estatus. Estas clasificaciones cruzadas se llaman tablas de movilidad social, y una asociación positiva entre las dos variables mostrará si los términos de interacción en la diagonal tienden a ser más grandes que el resto de los términos de interacción. Los efectos principales describen la distribución del estatus de los padres e hijos.

En una tabla de movilidad social, el caso en el que no se presentan interacciones entre el origen social i (grupo social parental) y el grupo social j se conoce como el modelo de "movilidad perfecta". Bajo este modelo

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j$$

o en forma multiplicativa

$$\mu_{ij} = a_i b_j$$

donde $a_i = \exp(\alpha_i + 0.5\mu)$ y $b_j = \exp(\beta_j + 0.5\mu)$. El programa C.1 (ver apéndice) simula α_i y β_j de las distribuciones finales de este modelo a los Datos de Movilidad Social Británicos de Glass (1954), como se muestran en la tabla 3.3.

Existe una correspondencia entre el rango de los renglones y el rango de las columnas que algunas veces pueden estar ordenadas con respecto al prestigio, estatus, etc. (Una tabla con variables renglón y columna arregladas de esta manera algunas veces se llama tabla cuadrada.)

Ajustando el modelo de independencia, simulando muestras del criterio de información de la devianza (ver sección 3.5) se obtiene que el $DIC = 976$. El ajuste a lo largo de la diagonal principal no es bueno, pues las predicciones se subestiman. La transición $(1, 1)$ del estatus alto del padre al estatus alto del hijo se predice como $\mu_{11} = 37.6$, y las otras diagonales predecidas en

	Estatus del hijo				
Estatus del padre	1	2	3	4	5
1	50	45	8	18	8
2	28	174	84	154	55
3	11	78	110	223	96
4	14	150	185	714	447
5	0	42	72	320	411

Tabla 3.3: Movilidad social.

promedio son 69.2, 68.0, 617.2 y 246. Un modelo más satisfactorio podría tratar la diagonal principal de una manera diferente del resto de la tabla.

Esta es la base del modelo de "movilidad cuasi-perfecta"(*quasi-perfect mobility* o QPM) de Goodman (1981) en donde

$$\mu_{ij} = a_i b_j \qquad \text{si } i \neq j$$

pero

$$\mu_{ij} = n_{ii} \qquad \text{si } i = j.$$

El modelo loglineal implica $l = 2(r-1) + (c-1) + 2$ parámetros, de tal manera permanecen $rc - l$ grados de libertad, y tiene la forma de

$$\begin{aligned} \log(\mu_{ij}) &= r + s_i + t_j, \qquad i \neq j \\ \log(\mu_{ii}) &= u + v_i. \end{aligned}$$

Ajustando este modelo y simulando muestras del criterio de información de la devianza (ver sección 3.5) se obtiene que el $DIC = 425$, indicando un mejor ajuste para los datos, y así, un mejor modelo. Note que el ajuste promedio a lo largo de la diagonal principal no reproduce de una manera exacta los valores observados, debido a la variabilidad muestral en las simulaciones producidas por WinBUGS. El conjunto completo de ajustes bajo el modelo QPM, y su intervalo de mayor densidad del 95 % se presentan en la tabla 3.4.

El ajuste fuera de la diagonal principal se mejora pero las discrepancias permanecen fijas, por ejemplo en el modelo ajustado de movilidad del estatus de origen 1 a los estatus 2, 3, 4 y 5. Las distancias más grandes en la movilidad están sobreestimadas y las distancias más cortas de la movilidad (del estatus 1 a 2) están subestimadas. Aquí las distancias se refieren a las distancias sociales en el sentido del prestigio entre el estatus origen y el estatus destino. ■

Otro enfoque (el modelo de cuasi-simetría) para estas tablas de transición se considerará a continuación.

	Media	SD	2.5 %	97.5 %
μ_{11}	51.2	7.2	37.6	65.7
μ_{12}	9.7	1.2	7.5	12.2
μ_{13}	11.2	1.3	8.7	14.0
μ_{14}	38.9	4.3	30.9	47.8
μ_{15}	20.2	2.3	16.0	25.1
μ_{21}	6.7	1.0	4.9	8.8
μ_{22}	173.7	13.3	148.4	199.9
μ_{23}	49.8	3.9	42.7	57.6
μ_{24}	173.8	10.8	154.0	195.2
μ_{25}	90.3	6.3	78.5	103.0
μ_{31}	8.7	1.3	6.2	11.4
μ_{32}	56.1	4.2	48.1	64.8
μ_{33}	109.8	10.4	90.6	130.8
μ_{34}	225.9	12.8	201.3	252.2
μ_{35}	117.4	7.6	102.8	132.9
μ_{41}	28.0	4.1	20.7	36.6
μ_{42}	180.9	11.2	159.9	203.9
μ_{43}	208.7	12.2	185.9	233.9
μ_{44}	714.2	26.4	663.2	766.4
μ_{45}	378.2	17.4	345.4	413.3
μ_{51}	10.6	1.6	7.7	13.9
μ_{52}	68.6	5.2	59.0	79.1
μ_{53}	79.2	5.8	68.3	91.0
μ_{54}	276.1	14.7	248.3	305.6
μ_{55}	410.8	20.4	371.3	451.5

Tabla 3.4: Movilidad cuasi-perfecta.

El siguiente ejemplo muestra un modelo en donde se presenta una tabla incompleta.

Ejemplo. *Inter-sib marriage: Una tabla con ceros estructurales.*

La celda en la tabla 3.3 con valor cero se conoce como un muestreo cero y se debe a lo poco frecuente del status 5 al 1. En contraste, algunas tablas de contingencia contienen entradas que son cero por definición. Bishop et al. (1975) analizan datos acerca de personas de distinta raza entre *sibs* (clanes afines por parentesco) entre las personas de Purum en India. Los Purums, son una tribu vieja y aislada que viven en el interior de la India, y están divididos en cinco *sibs*: Marrim, Makan, Kheyang, Thao, y Parpa. Una de sus reglas es la exogamia (matrimonios fuera del parentesco), y se permiten

los matrimonios *inter-sib*. La tabla 3.5 contiene observaciones acerca de 128 matrimonios Purum, los matrimonios prohibidos se denotan por NA. Note que la diagonal presenta ceros estructurales (NA). Hay un matrimonio *intra-sib* pero este era entre compañeros en diferentes *sub-sibs*. La tabla también contiene dos muestras cero, que se deben a que no hubo matrimonio entre los *sibs* Makan y Parpa, así como entre los Kheyang y Thao, aunque estos no fueran prohibidos.

		Sib del esposo				
Sib de la esposa		Marrim	Makan	Parpa	Thao	Kheyang
Marrim	Obs	NA	5	17	NA	6
	Est	NA	10.86	10.71	NA	6.58
Makan	Obs	5	NA	0	16	2
	Est	4.86	NA	6.54	7.58	4
Parpa	Obs	NA	2	NA	10	11
	Est	NA	8.35	NA	9.52	5.05
Thao	Obs	10	NA	NA	NA	9
	Est	10.35	NA	NA	NA	8.65
Kheyang	Obs	6	20	8	0	1
	Est	5.75	7.83	7.71	8.95	4.71

Tabla 3.5: Matrimonios Purum (observaciones y medias finales estimadas).

El modelo aplicado a estos datos es el de independencia pero confinado a permitir enlaces (i.e. una forma de cuasi-independencia). Se define una matriz de orden 5×5 con elementos 1 para permitir enlaces, y con 0 para ceros estructurales. Las estimaciones de la media de la distribución final de la distribución Poisson usando el programa C.2 *Inter Sib Marriage* son similares a los de Bishop et al. (1975), y la estadística de la devianza tiene un mínimo de 75.9, cercana a $G^2 = 76.2$ de Bishop et al. (1975). ■

3.3.4. Modelos de Cuasi-Simetría

El modelo de *cuasi-simetría* reintroduce parámetros de interacción γ_{ij} pero supone que son iguales en las celdas fuera de la diagonal, esto es $\gamma_{ij} = \gamma_{ji}$. Entonces la versión loglineal del modelo es

$$\log(\mu_{ij}) = \delta + \alpha_i + \beta_j + \gamma_{ij} \tag{3.7}$$

con $\gamma_{ij} = \gamma_{ji}$ y condicionando a que sumen cero los α_i y los β_j. La condición sobre el término de interacción se aplica de la misma manera a los renglones

de γ_{ij}, por ejemplo suponiendo que $\sum_i \gamma_{ij} = 0$. El modelo de cuasi-simetría produce valores con simetrías en una tabla cuadrada, de tal manera que $\mu_{ij} \approx \mu_{ji}$. Esta simetría implica que los totales marginales por renglones μ_{i+} son iguales a los totales marginales por columnas μ_{+j}, patrón conocido como "homogeneidad marginal"; sin embargo, puede haber homogeneidad sin que haya simetría.

El modelo de cuasi-simetría puede establecerse en forma multiplicativa como

$$\mu_{ij} = a_i b_j e_{ij}, \qquad i \neq j$$

donde

$$e_{ij} = e_{ji} \qquad \text{y} \qquad \mu_{ii} = a_i.$$

Un tipo particular del modelo de cuasi-simetría es el modelo de parámetros diagonales para las celdas fuera de la diagonal, es decir

$$\mu_{ij} = a_i b_j d_k \tag{3.8}$$

donde $k = i - j$ para $i \neq j$, y k toma valores 1, 2, 3, 4 y -1, -2, -3, -4 en una tabla de 5×5. En el contexto de movilidad social, los parámetros d_k medirían el impacto de la distancia social y el declinamiento esperado en la movilidad cuando k crece en valor absoluto. Generalmente se supone que los efectos descendentes y ascendentes son los mismos, es decir $d_k = d_{-k}$.

Una aplicación epidemiológica del modelo de cuasi-simetría se tiene en el caso de los datos de casos y controles con el mismo número de controles para cada caso (Lovison, 1994). En particular, suponga que hay n pares iguales (un control para cada caso) y una variable con r niveles. Entonces los datos pueden representarse como una tabla de "concordancia" $r \times r$ con n_{ij} el número de pares en los cuales un caso es expuesto al nivel i y un control expuesto al nivel j. Así las variables renglón y columna son las mismas pero observadas sobre dos miembros de un par. Las frecuencias esperadas μ_{ij} pueden modelarse como sigue

$$\mu_{ij} = \frac{n\pi_{ij}\psi_{ij}}{1 + \psi_{ij}},$$

donde π_{ij} es la probabilidad de que un miembro de un par sea expuesto al nivel de riesgo i y el otro al nivel j; y donde ψ_{ij} es la (i, j)-ésima razón de momios de exposición, a saber

$$\psi_{ij} = \frac{P(\text{nivel } i \text{ expuesto}|caso)P(\text{nivel } j \text{ expuesto}|control)}{P(\text{nivel } j \text{ expuesto}|caso)P(\text{nivel } i \text{ expuesto}|control)}.$$

Si los términos ψ_{ij} son constantes bajo las mismas variables, entonces satisfacen la condición

$$\psi_{ij} = \frac{\psi_{ib}}{\psi_{jb}}, \tag{3.9}$$

donde b es el nivel de referencia de los expuestos. Las $r(r-1)/2$ razones de momios pueden expresarse como $(r-1)$ parámetros $\psi_{ib} = \alpha_i$. De tal manera que hay un factor de exposición sobre los decesos y su efecto depende del nivel de exposición. El modelo loglineal equivalente es

$$\begin{aligned} n_{ij} &\sim Poisson(\mu_{ij}) \\ \mu_{ij} &= M + \delta_{ij} + \alpha_i, \qquad i \neq j \\ \mu_{ii} &= M + \gamma_i \end{aligned}$$

donde $\delta_{ij} = \delta_{ji}$, y $\alpha_1 = 0$, $\gamma_1 = 0$. Las hipótesis de no efecto y efecto constante, respectivamente, son $\alpha_i = 0$ y $\alpha_i = \alpha$.

Ejemplo. *Movilidad social (continuación).*

El modelo de cuasi-simetría de la ecuación (3.7) aplicado a los datos de la tabla 3.3 da un promedio de $G^2 = 28.4$ y tiene un mínimo de 12.6 (ver programa C.3). Esta es una mejoría considerable sobre el modelo de movilidad perfecta y el modelo de movilidad cuasi-perfecta.

El mínimo G^2 del ajuste del modelo de "distancia social"de (3.8) (ver programa C.3) es 20.7 con un promedio de 35.5, comparada con un valor de la verosimilitud máxima de 19.1 obtenida por Bishop *et al.* (1975, p. 228). Los parámetros d_k (y sus intervalos de 95 % de densidad) son respectivamente $d_1 = 1$, $d_2 = 0.59$ $(0.53, 0.66)$, $d_3 = 0.26$ $(0.21, 0.32)$ y $d_4 = 0.084$ $(0.035, 0.158)$. Hay el decline esperado con la distancia social, y aproximadamente una progresión geométrica. Note que podemos ajustar este modelo de una mejor forma multiplicativa que el modelo loglineal, así que las distribuciones iniciales sobre parámetros libres se expresan en términos de densidades gammas, $Gamma(0.01, 0.01)$. Las medias ajustadas debajo de los parámetros de la diagonal y los modelos cuasi-simetría están dados en la tabla 3.6.

■

Ejemplo. Observaciones apareadas por grupo de sangre.

Lovison (1994) analizó datos de 301 observaciones apareadas clasificados sobre la variable grupo de sangre con cuatro niveles (grupos O, A, B y AB). No hay categoría de "ausencia de exposición"y en cambio el grupo O es la categoría de referencia. La tabla de contingencia es entonces una tabla de concordancia (tabla 3.7)

	Quasi-symmetry		Parámetros de la diagonal	
	Media	SD	Media	SD
μ_{11}	46.9	6.4	51.6	6.6
μ_{12}	43.3	6.2	35.7	4.8
μ_{13}	11.5	2.7	15.4	2.2
μ_{14}	18.4	3.2	21.4	3.3
μ_{15}	7.1	1.6	5.4	1.9
μ_{21}	30.3	5.4	24.2	3.5
μ_{22}	174.0	12.9	174.1	13.1
μ_{23}	78.2	7.4	85.8	7.1
μ_{24}	155.5	11.2	155.9	11.1
μ_{25}	56.9	6.2	55.3	5.8
μ_{31}	8.5	2.2	11.2	1.7
μ_{32}	83.4	7.7	92.2	7.2
μ_{33}	109.9	10.6	109.9	10.5
μ_{34}	215.2	13.4	208.2	12.6
μ_{35}	101.2	8.6	96.8	8.0
μ_{41}	12.4	2.8	13.8	2.4
μ_{42}	148.7	11.0	148.5	10.7
μ_{43}	193.0	12.4	184.6	11.8
μ_{44}	714.6	27.4	712.9	26.5
μ_{45}	441.7	20.1	449.1	20.4
μ_{51}	3.5	0.9	2.6	1.0
μ_{52}	40.0	4.7	38.7	4.3
μ_{53}	66.8	6.3	63.0	5.8
μ_{54}	324.7	17.1	329.6	17.0
μ_{55}	410.7	20.4	410.3	20.0

Tabla 3.6: Movilidad social.

	Control			
Caso	O	A	B	AB
O	64	18	8	3
A	66	74	14	6
B	4	2	4	2
AB	12	10	12	2

Tabla 3.7: Tabla de concordancia.

Las estimaciones finales de las razones de momios expuestas para los grupos A, B y AB suponiendo el modelo (3.9) se obtienen a partir de la segunda mitad de una corrida de WinBUGS con 10000 iteraciones (tabla 3.8).

	Media	SD	2.5 %	Mediana	97.5 %
ϕ_1	3.68	0.94	2.21	3.55	5.85
ϕ_2	0.58	0.26	0.21	0.54	1.21
ϕ_3	5.41	2.42	2.21	4.90	11.51

Tabla 3.8: Estimaciones finales.

Como podemos ver en la tabla 3.8, las medias de los cocientes de momios exceden a las medianas. Las medianas finales son cercanas a las reportadas por Lovison, $\phi_2 = 3.50$, $\phi_3 = 0.56$ y $\phi = 4.67$ (ver programa C.4). ■

3.4. Otros Modelos

En esta sección presentaremos una revisión selectiva de algunos problemas especializados para los cuales los métodos Bayesianos son muy convenientes.

3.4.1. Datos Faltantes: No Respuesta

Cuando los datos categóricos se recolectan con algunas observaciones incompletas, los datos pueden resumirse en dos tipos de tablas de contingencia: una tabla categorizada completamente y tablas categorizadas parcialmente, tales como tablas marginales. Los mecanismos de no respuesta son llamados *ignorables* para inferencias basadas en verosimilitudes cuando el mecanismo de no respuesta es independiente de la respuesta no observada del sujeto y son llamados *no ignorables* cuando la probabilidad de una no respuesta depende de la respuesta no observada.

Park y Brown (1994) y Forster y Smith (1998) desarrollaron aproximaciones Bayesianas para modelar la no respuesta en problemas de datos categóricos. Específicamente, su trabajo considera tablas de contingencia que contienen datos de clasificación cruzada completamente y parcialmente, donde una de las variables (digamos Y) es una variable respuesta sujeta a una no respuesta no ignorable y las otras variables (colectivamente denotadas por X) son consideradas como covariables y siempre son observadas. Estos autores introducen una variable indicadora R para representar un mecanismo de respuesta dicotómica ($R = 1$ y $R = 0$ indicando respuesta y no respuesta respectivamente). Un modelo de no respuesta se define como un modelo loglineal para el arreglo completo de Y, X, y R. Un modelo de una no respuesta no ignorable es uno que contiene un término de interacción YR.

Park y Brown (1994) muestran que un pequeño cambio o movimiento de las no respuestas puede resultar en un gran cambio en los estimadores máximo verosímiles de las frecuencias de celdas esperadas. La estimación por medio de máxima verosimilitud es problemática porque pueden ocurrir soluciones acotadas, en cuyo caso las estimaciones de los parámetros del modelo no pueden ser determinadas de manera única y pueden ser inestables. Park y Brown (1994) proponen un método Bayesiano que usa distribuciones iniciales que dependen de los datos para proporcionar información acerca de la extensión de la no ignorabilidad. El efecto neto de tales distribuciones iniciales es la introducción de suavizamientos constantes.

3.4.2. Falta de Identificabilidad

Censura

Los modelos estándar para datos categóricos censurados son generalmente no identificables. Para superar este problema, se supone que el mecanismo de censura es ignorable (no informativo) en el sentido de que el parámetro desconocido de la distribución que describe el mecanismo de censura no está relacionado con el parámetro de interés (ver Dickey et al. 1987). Paulino y Pereira (1995) discuten métodos conjugados Bayesianos para datos categóricos en general, con una censura informativa. En particular, están interesados en la estimación Bayesiana de las frecuencias de celdas a través de esperanzas de las distribuciones finales. Walker (1996) considera la maximización de una densidad final, obtenida por medio de un algoritmo EM, para una clase más general de distribuciones iniciales.

Mala Clasificación

Paulino et al. (2003) presentan un análisis Bayesiano completo de datos de una regresión binomial con respuestas posiblemente mal clasificadas. Su enfoque puede extenderse a ajustes multinomiales. Usan un modelo de *misclassification* o *mala clasificación* informativo cuyos parámetros son no identificables. Como en el caso de censura, desde un punto de vista Bayesiano este no es un problema serio debido a que una distribución inicial propia conveniente puede hacer los parámetros identificables. Sin embargo, las inferencias finales sobre los parámetros no identificables pueden depender fuertemente de las distribuciones iniciales aún para tamaños de muestras grandes.

Análisis de Clases Latentes

Un modelo de clases latentes generalmente trae consigo un conjunto de variables observadas llamadas *variables manifiestas* y un conjunto de variables aleatorias no observadas o no observables llamadas variables latentes. Los modelos más usados de este tipo son los modelos latentes condicionalmente independientes, que establecen que todas las variables manifiestas son condicionalmente independientes dadas las variables latentes.

El análisis de clases latentes en tablas de contingencia de dos dimensiones generalmente padece de problemas de identificabilidad. Estos pueden superarse usando técnicas Bayesianas en las cuales ciertas distribuciones iniciales se asignan a los parámetros latentes.

Evans et al. (1989) describen los modelos latentes condicionales en tablas de dos dimensiones de la siguiente manera. Sean X y Y dos variables aleatorias categóricas con rangos que consisten de los enteros del 1 al I y del 1 al J respectivamente. Sea

$$p_{ij} = p(X = i, Y = j), \qquad 1 \leq i \leq I, \quad 1 \leq j \leq J$$

la distribución conjunta de X y Y y sean

$$p_{i\cdot} = \sum_j p_{ij} \quad 1 \leq i \leq I, \qquad p_{\cdot j} = \sum_i p_{ij} \quad 1 \leq j \leq J,$$

las distribuciones marginales de X y Y, respectivamente.

Sea Z otra variable categórica con un rango que consiste de los enteros del 1 al K. Sean

$$\begin{aligned} \theta_k &= p(Z = k) && (1 \leq k \leq K), \\ \alpha_i(k) &= p(X = i | Z = k) && (1 \leq i \leq I, 1 \leq k \leq K), \\ \beta_j(k) &= p(Y = j | Z = k) && (1 \leq j \leq J, 1 \leq k \leq K). \end{aligned}$$

Consideramos a X y Y como variables manifiestas y a Z como una variable latente. Entonces el modelo latente condicionalmente independiente establece que las variables aleatorias X y Y son estocásticamente independientes dada Z. A partir de aquí la probabilidad conjunta p_{ij} puede escribirse como

$$p_{ij} = \sum_{k=1}^{K} \theta_k \alpha_i(k) \beta_j(k) \qquad (1 \leq i \leq I, 1 \leq j \leq J). \tag{3.10}$$

Para que (3.10) sea un modelo no saturado, se requiere que $K < \text{mín}(I, J)$.

Por simplicidad y sin pérdida de generalidad, se utilizarán variables latentes dicotómicas ($K = 2$). Entonces

$$p_{ij} = \theta \alpha_i(1) \beta_j(1) + (1 - \theta) \alpha_i(2) \beta_j(2).$$

Se especifican distribuciones iniciales para θ, $\alpha_i(1)$, $\alpha_i(2)$, $\beta_j(1)$ y $\beta_j(2)$ y se calculan las esperanzas de las distribuciones finales de estos parámetros dadas las frecuencias observadas. Una clase natural de estas distribuciones iniciales es tomar distribuciones Dirichlet marginales mutuamente independientes.

Generalmente es difícil hacer los cálculos de manera analítica, Evans et al. (1989) usan la técnica de muestreo por importancia para obtener las aproximaciones de las esperanzas de las distribuciones finales, las cuales son usadas como estimadores puntuales de los parámetros del modelo.

Ejemplo. *Frecuencia de visita.*

Evans et al. (1989) consideran una tabla de contingencia referente a 132 pacientes de esquizofrenia y la frecuencia con la que éstos ven a sus familiares. Se utiliza una tabla de contingencia de dos dimensiones de orden 3×3, $I = 3$ categorías para la frecuencia de las visitas y $J = 3$ categorías para los grupos de visitas del hospital (tabla 3.9). Evans et al. (1989) estimaron un modelo de clases latentes usando el muestreo por importancia.

El modelo programado en WinBUGS se presenta en el apéndice C.5.

Los resultados se presentan en la tabla 3.10.

■

3.4.3. Categorías Ordenadas

Albert y Chib (1993) desarrollaron métodos Bayesianos para modelar datos de respuesta categórica usando la idea de aumento de datos combinada con técnicas de Monte Carlo vía cadenas de Markov. Por ejemplo, el modelo de regresión probit para datos binarios se supone que tiene una estructura

Frecuencia de visita	Años de estancia en el hospital $[2, 10)$	$[10, 20)$	$[20,)$	Total
Va a casa o es visitado regularmente	43	16	3	62
No va a casa y es visitado menos de una vez al mes	6	11	10	27
No va a casa y no es visitado	9	18	16	43
Total	58	45	29	132

Tabla 3.9: Frecuencias de las visitas para 132 pacientes de esquizofrenia.

de regresión normal sobre datos continuos latentes. Estos autores generalizan esta idea para modelos de respuesta multinomial, incluyendo el caso donde las categorías multinomiales están ordenadas. En este último caso, los modelos enlazan las probabilidades acumulativas de las respuestas con una estructura de regresión lineal.

Este enfoque tiene un número de ventajas, especialmente en el sistema multinomial, donde puede ser difícil evaluar la función de verosimilitud. Para muestras pequeñas, esta aproximación Bayesiana generalmente se desempeñará mejor que los métodos de máxima verosimilitud tradicionales, los cuales se basan en resultados asintóticos. Además, uno puede elaborar el modelo probit usando mezclas apropiadas de la distribución normal para modelar los datos latentes.

Albert y Chib (1993) modelan las categorías ordenadas de la siguiente manera. Suponga que $Y_1, \ldots, Y_N$ son variables aleatorias observadas, donde Y_i toma valores en alguna de las J categorías ordenadas, $1, \ldots, J$. Sea $p_{ij} = p(Y_i = j)$ y definimos las probabilidades acumulativas como $\eta_{ij} = \sum_{k=1}^{j} p_{ik}$, $j = 1, \ldots, J-1$. Un modelo de regresión para los $\{p_{ij}\}$ esta dado por $\eta_{ij} = \Phi(\gamma_j - \mathbf{x}_i'\beta)$, $i = 1, \ldots, N,\ j = 1, \ldots, J-1$. Uno puede motivar este modelo suponiendo que existe una variable aleatoria continua latente Z_i que se distribuye $Normal(\mathbf{x}_i'\beta, 1)$, y que observamos Y_i, donde $Y_i = j$ si $\gamma_{j-1} < Z_i \leq \gamma_j$ (definimos $\gamma_0 = -\infty$ y $\gamma_J = \infty$). Este problema es un problema de regresión normal donde las variables respuesta están en forma de datos agrupados.

En el modelo, el vector de regresión β y las cotas $\gamma_1, \ldots, \gamma_{J-1}$ son desconocidas. Para asegurar que los parámetros son identificables, es necesario imponer una restricción sobre las cotas; sin pérdida de generalidad, tomamos

	Mean	SD	2.5 %	97.5 %
$\alpha_1(1)$	0.090	0.064	0.004	0.241
$\alpha_1(2)$	0.793	0.084	0.632	0.951
$\alpha_2(1)$	0.348	0.071	0.217	0.495
$\alpha_2(2)$	0.089	0.050	0.007	0.196
$\alpha_3(1)$	0.562	0.078	0.411	0.716
$\alpha_3(2)$	0.118	0.064	0.010	0.248
$\beta_1(1)$	0.118	0.071	0.006	0.262
$\beta_1(2)$	0.716	0.078	0.571	0.880
$\beta_2(1)$	0.448	0.076	0.303	0.598
$\beta_2(2)$	0.241	0.071	0.088	0.378
$\beta_3(1)$	0.435	0.080	0.292	0.605
$\beta_3(2)$	0.043	0.032	0.002	0.123
θ_1	0.470	0.079	0.321	0.629
θ_2	0.530	0.079	0.371	0.679

Tabla 3.10: Frecuencia de visitas.

$\gamma_1 = 0$. La densidad final conjunta de β y $\gamma = (\gamma_2, \ldots, \gamma_{J-1})$ está dada por

$$\pi(\beta, \gamma|\mathbf{y}) = C\pi(\beta, \gamma) \prod_{i=1}^{N} \sum_{j=1}^{J} 1_{(y_i=j)} \left[\Phi(\gamma_j - \mathbf{x}_i'\beta) - \Phi(\gamma_{j-1} - \mathbf{x}_i'\beta)\right],$$

donde $\pi(\beta, \gamma)$ es la distribución inicial. Encontramos la moda de la distribución final de (β, γ) usando el algoritmo de Newton-Raphson (ver sección 2.4.5) y se obtienen las desviaciones estándar aproximadas de las distribución final de (β, γ) usando la segunda derivada del logaritmo de las distribuciones finales evaluada en la moda.

Podemos generalizar el algoritmo de Gibbs para esta situación. Introducimos las variables aleatorias latentes no observadas $Z_1, \ldots, Z_n$ definidas previamente y simuladas de la distribución final conjunta de $(\beta, \gamma, \mathbf{Z})$. Si asignamos una distribución inicial para (β, γ), entonces esta densidad final conjunta está dada por

$$\Pi(\beta, \gamma, \mathbf{Z}|\mathbf{y}) = C \prod_{i=1}^{N} \left[\sqrt{\frac{1}{2\pi}} \exp\left(-\frac{(Z_i - \mathbf{x}_i'\beta)^2}{2} \right) \left\{ \sum_{j=1}^{J} 1_{(Y_i=j)} 1_{(\gamma_{j-1} < Z_i < \gamma_j)} \right\} \right].$$

La distribución final de β condicional a $\mathbf{y}$ y $\mathbf{Z}$ está dada por la distribución normal multivariada

$$Normal_k(\hat{\beta}_Z, (\mathbf{X}'\mathbf{X})^{-1}). \tag{3.11}$$

Las distribuciones finales condicionales completas de $Z_1, \ldots, Z_N$ son independientes con

$$\begin{aligned} Z_i|\beta, \gamma, y_i = j \text{ que se distribuye } Normal(\mathbf{x}_i'\beta, 1) \\ \text{truncada en el lado izquierdo (derecho) por } \gamma_{j-1}(\gamma_j). \end{aligned} \tag{3.12}$$

Finalmente, la densidad condicional completa de γ_j dada $\mathbf{Z}$, $\mathbf{y}$, β, y $\{\gamma_k, k \neq j\}$ está dada (hasta una constante de proporcionalidad) por

$$\prod_{i=1}^{N} \left[1_{(Y_i=j)} 1_{(\gamma_{j-1}<Z_i<\gamma_j)} + 1_{(Y_i=j+1)} 1_{(\gamma_j<Z_i<\gamma_{j+1})}\right]. \tag{3.13}$$

Esta distribución condicional puede verse como una distribución condicional uniforme sobre el intervalo [máx{máx$\{Z_i : Y_i = j\}, \gamma_{j-1}\}$, mín{mín$\{Z_i : Y_i = j+1\}, \gamma_{j+1}\}$]. Para implementar el muestreo de Gibbs, comenzamos con (β, γ) inicializada en el valor del estimador máximo verosímil y simulamos de las distribuciones de las ecuaciones (3.13), (3.12) y (3.11), en ese orden.

Ejemplo. *Clases de estadística.*

Congdon (2005) presenta un ejemplo de 30 estudiantes de estadística clasificados en cinco clases escolares (no aprobado=1, grado D=2, grado C=3, grado B=4, grado A=5) con una variable predictora X=puntaje. Suponemos que $\gamma_0 = \infty$, $\gamma_1 = 0$ y que las distribuciones iniciales para los otros puntos de corte son:

$$\begin{aligned} \gamma_2 &\sim Normal(1, 10) 1_{(0,\gamma_3)} \\ \gamma_3 &\sim Normal(2, 10) 1_{(\gamma_2,\gamma_4)} \\ \gamma_4 &\sim Normal(3, 10) 1_{(3,\infty)}. \end{aligned}$$

Se supone una distribución inicial $Normal(0, 10)$ para el puntaje de SAT, aunque se podría escalar el score dividiendo, por ejemplo, entre 100.

El programa realizado en WinBUGS se presenta en el apéndice C.6.

Los resultados que se obtuvieron para este modelo se presentan en la tabla 3.11.

■

3.4.4. Categorías No Ordenadas con una Distribución Multinormal Latente

Albert y Chib (1993) consideran un modelo para categorías no ordenadas y con variables latentes que tienen distribución multinormal, este puede verse como un caso particular del modelo probit multinormal.

node	mean	sd	MC error	2.50 %	median	97.50 %
D	89.62	8.071	0.3035	75.64	89.01	107.1
b[1]	2.233	0.4475	0.02638	1.404	2.213	3.171
b[2]	0.02516	0.00636	$1.85E-04$	0.01291	0.02501	0.03813
gamma[2]	1.542	0.4549	0.02702	0.7589	1.505	2.526
gamma[3]	2.422	0.4894	0.03123	1.493	2.402	3.44
gamma[4]	3.77	0.58	0.03297	2.677	3.753	4.956
ynew[1]	3.183	0.9539	0.006641	2	3	5
ynew[2]	1.515	0.6711	0.004655	1	1	3
ynew[3]	2.505	0.9178	0.007071	1	2	4
ynew[4]	2.665	0.927	0.006434	1	3	4
ynew[5]	3.695	0.9133	0.00643	2	4	5
ynew[6]	1.627	0.7243	0.004739	1	2	3
ynew[7]	3.195	0.9524	0.006773	2	3	5
ynew[8]	2.356	0.8908	0.006362	1	2	4
ynew[9]	1.863	0.789	0.004918	1	2	4
ynew[10]	3.679	0.9175	0.006827	2	4	5
ynew[11]	3.503	0.9373	0.00623	2	4	5
ynew[12]	3.217	0.9522	0.007238	2	3	5
ynew[13]	2.88	0.9467	0.006801	1	3	5
ynew[14]	3.541	0.9284	0.00681	2	4	5
ynew[15]	3.713	0.9023	0.00667	2	4	5
ynew[16]	3.881	0.8741	0.005729	2	4	5
ynew[17]	2.916	0.9507	0.007503	1	3	5
ynew[18]	3.576	0.9157	0.005828	2	4	5
ynew[19]	3.569	0.9237	0.006019	2	4	5
ynew[20]	2.5	0.9062	0.006612	1	2	4
ynew[21]	3.546	0.9239	0.006137	2	4	5
ynew[22]	3.947	0.8615	0.006136	2	4	5
ynew[23]	3.745	0.9048	0.006815	2	4	5
ynew[24]	3.739	0.9019	0.006779	2	4	5
ynew[25]	3.304	0.9514	0.007005	2	3	5
ynew[26]	3.084	0.9566	0.007407	1	3	5
ynew[27]	4.193	0.7986	0.005686	2	4	5
ynew[28]	4.021	0.8525	0.006233	2	4	5
ynew[29]	4.691	0.5642	0.00529	3	5	5
ynew[30]	3.001	0.9475	0.006864	1	3	5

Tabla 3.11: Clases de estadística.

Sean $Z_1, \ldots, Z_N$ variables aleatorias independientes latentes no observables, donde $Z_i = (Z_{i1}, \ldots, Z_{iJ})$ $(J > 2)$, y definimos $Z_{ij} = \mathbf{x}'_{ij}\beta + \epsilon_{ij}$, $i = 1, \ldots, N$ $j = 1, \ldots, J$, donde $\epsilon_i = (\epsilon_{i1,\ldots,\epsilon_{iJ}})'$ se distribuye $Normal_J(\mathbf{0}, \mathbf{\Sigma})$ y Σ es una matriz de orden $J \times J$ que está parametrizada (por razones de identificabilidad) en términos de un vector de parámetros θ de dimensión tal que no excede $J(J-1)/2$. Podemos considerar a i como un índice de unidades experimentales y a j como un índice de categorías. Para una unidad experimental i observamos una de las J posibles salidas con probabilidades respectivas $p_{i1}, \ldots, p_{iJ}$. La categoría j es observada si $Z_{ij} > Z_{ik}$ para toda $k \neq j$. Las probabilidades multinomiales están dadas por $p_{ij} = P[\mathbf{x}'_{ij}\beta + \epsilon_{ij} > \mathbf{x}'_{ik}\beta + \epsilon_{ik}, \forall k \neq j]$. Note que el cálculo de estas probabilidades implica el cálculo de integrales múltiples de la densidad normal multivariada; así que la estimación por medio de máximo verosimilitud es muy difícil de obtener para una J grande.

El cálculo de las probabilidades multinomiales puede obtenerse por medio del muestreo de Gibbs. Sea $\mathbf{Y} = (y_1, \ldots, y_N)$ el vector de categorías observadas, donde $y_i \in \{1, \ldots, J\}$. Sea $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})'$, el modelo anterior lo podemos escribir como

$$\begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_N \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{3.14}$$

o como $\mathbf{Z} = \mathbf{X}\beta + \epsilon$, donde $\epsilon = (\epsilon'_1, \ldots, \epsilon'_N)'$ se distribuye $Normal_{NJ}(\mathbf{0}, \Omega = \mathbf{I}_N \otimes \Sigma)$. Para implementar el muestreo de Gibbs, necesitamos muestras de las siguientes distribuciones condicionales:

$$\begin{aligned} \beta|\mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N, \theta \\ \mathbf{Z}_1, \ldots, \mathbf{Z}_N|\mathbf{Y}, \beta, \theta \\ \theta|\mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N, \beta. \end{aligned} \tag{3.15}$$

A partir de la representación de (3.14), si se utiliza una distribución inicial difusa para β entonces la teoría normal multivariada estándar establece que $\beta|\mathbf{Z}_1, \ldots, \mathbf{Z}_N, \mathbf{Y}, \theta$ se distribuye $Normal_k(\hat{\beta}_{\mathbf{Z}}, (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1})$, donde $\hat{\beta}_{\mathbf{Z}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{Z}$. Note que el cálculo de los parámetros de la última distribución es fácil, porque Ω^{-1} es una matriz diagonal. Dado $\mathbf{Y}$, β y θ, la colección $\{\mathbf{Z}_i\}$ es una colección de variables aleatorias independientes, donde $\mathbf{Z}_i|\mathbf{Y}, \beta, \theta$ se distribuye $Normal(\mathbf{x}_i\beta, \Sigma)$, $i = 1, \ldots, N$, tal que la Y_i-ésima componente de $\mathbf{Z}_i$ es el máximo. Esto puede simularse a través de la selección de un muestreo de la distribución $Normal(\mathbf{x}_i\beta, \Sigma)$, aceptando la selección si la condición se satisface. Finalmente, considere el muestreo de

$\theta|\mathbf{Z}_1,\ldots,\mathbf{Z}_N,\mathbf{Y},\beta$. Usando una distribución inicial $\pi(\theta)$ para θ, la densidad de esta distribución es proporcional a

$$\pi(\theta)\,|\Omega(\theta)|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{Z}-\mathbf{X}\beta)'\Omega^{-1}(\theta)(\mathbf{Z}-\mathbf{X}\beta)\right\}.$$

Esta distribución no pertenece a ninguna familia paramétrica conocida y es relativamente difícil de simular.

3.5. Selección de Modelos

El desarrollo de métodos de Monte Carlo vía cadenas de Markov ha hecho posible ajustar una gran variedad de modelos. Esta gran herramienta nos hace desear comparar estos modelos e identificar una clase de modelos que nos permitan describir de la manera más adecuada los datos.

Dentro del campo de los modelos clásicos, generalmente la comparación de modelos implica la definición de una medida de ajuste, generalmente una estadística de la devianza, y la complejidad del modelo, usualmente el número de parámetros en el modelo. Conforme aumenta la complejidad del modelo se tiene un mejor ajuste, por esto es que estas dos cantidades son suficientes para la selección de un modelo. Una comparación de modelos usando criterios de información Bayesiana también requieren la especificación del número de parámetros en cada modelo, pero en modelos jerárquicos complejos el número de parámetros puede ser mayor al número de observaciones y este criterio no puede aplicarse. De esta manera es necesario contar con una medida Bayesiana que considere la complejidad y el ajuste del modelo y que pueda utilizarse para comparar modelos de estructuras arbitrarias.

Spiegelhalter et al. (2002) sugieren la medida del *criterio de información de la devianza* (DIC) como una medida de selección de modelos, dada por

$$DIC = D(\bar{\theta}|y) + 2d_e$$

donde $D(\bar{\theta}|y)$ y d_e se describen a continuación.

El parámetro d_e es una estimación del número total efectivo de parámetros o de la dimensión del modelo.

Sea $L^{(t)} = \log[P(Y|\theta^{(t)})]$ el logaritmo de la verosimilitud obtenida en la i-ésima iteración a lo largo de una cadena en una corrida del algoritmo de Gibbs, y sea $D^{(t)} = -2L^{(t)}$ la devianza. Otra definición de la devianza usada principalmente en modelos lineales generalizados es $D^{(t)} = -2(L^{(t)} - L_s^{(t)})$ donde $L^{(t)}$ es el logaritmo de la verosimilitud del modelo saturado obtenida en la i-ésima iteración de la cadena.

Entonces d_e puede aproximarse por la diferencia entre la devianza esperada $E(D|y,\theta)$, dada por la media $\bar{D}$ de las devianzas $D^{(t)}$ obtenidas a partir de la cadena, y la devianza $D(\bar{\theta}|y)$ de la media $\bar{\theta}$ de la distribución final de los parámetros.

Valores pequeños de DIC indicarán un mejor modelo.

Capítulo 4

Otros Modelos

Como hemos mencionado en capítulos anteriores, al analizar variables categóricas en la mayoría de los casos se supone que estos datos siguen una distribución multinomial o Poisson; sin embargo, existen otras distribuciones que pueden aplicarse a los datos y que en algunos casos pueden ser más adecuadas. Una de estas posibles distribuciones es la multinomial negativa. La distribución multinomial negativa permite modelar datos con correlaciones positivas y que presentan varianzas mayores a las medias (sobredispersión relativa a la distribución Poisson).

También hemos estudiado la forma de analizar variables categóricas y modelos para variables categóricas que dependen de otras variables independientes; sin embargo, en ocasiones, estas variables están ordenadas en el tiempo y dependen de valores pasados e inclusive pueden a su vez depender de otras variables independientes. El utilizar modelos de regresión de series de tiempo permite analizar este tipo de datos.

En este capítulo presentaremos algunos modelos utilizados para datos ordenados en el tiempo y que dependen de otras variables, así como modelos de series de tiempo y modelos de regresión de series de tiempo, y que pueden tener una distribución Poisson o multinomial negativa, usando técnicas de estadística Bayesiana.

En la sección 4.1 se estudia una prueba Bayesiana para el análisis de tablas de contingencia utilizando la distribución multinomial negativa. En la sección 4.2 se describen brevemente los modelos de series de tiempo utilizados para el análisis de datos categóricos. En la sección 4.3 se presentan los modelos autoregresivos de valores enteros con rezago 1, utilizando las distribuciones Poisson y multinomial negativa. En la sección 4.4 se generalizan los modelos INAR(1) utilizando una estructura de regresión. En la sección 4.5 se da una breve introducción de los modelos autoregresivos de valores enteros con rezago p. Finalmente en la sección 4.6 se estudian brevemente los modelos

loglineales utilizando una distribución multinomial negativa.

4.1. Análisis de Tablas de Contingencia Basado en la Distribución Multinomial Negativa

Sea $(Y_1, \ldots, Y_T)$ un vector aleatorio con distribución multinomial negativa con índice $\alpha > 0$ y con vector de medias $(\lambda_1, \ldots, \lambda_T)$. El soporte de $(Y_1, \ldots, Y_T)$ es $(y_1, \ldots, y_T)$ tales que y_t, con $t = 1, \ldots, T$, es un entero no negativo. La función de probabilidad de masa de $(Y_1, \ldots, Y_T)$ es

$$\begin{aligned} p(\mathbf{y}|\lambda, \alpha) &= \frac{\Gamma\left(\alpha + \sum_{t=1}^{T} y_t\right)}{\Gamma(\alpha) \prod_{t=1}^{T} y_t!} \left(\frac{\alpha}{\alpha + \sum_{t=1}^{T} \lambda_t}\right)^{\alpha} \prod_{t=1}^{T} \left(\frac{\lambda_t}{\alpha + \sum_{t=1}^{T} \lambda_t}\right)^{y_t} \\ &= \frac{\Gamma\left(\alpha + \sum_{t=1}^{T} y_t\right)}{\Gamma(\alpha) \prod_{t=1}^{T} y_t!} \left(1 - \sum_{t=1}^{T} \pi_t\right)^{\alpha} \prod_{t=1}^{T} \pi_t^{y_t} \\ &= p(\mathbf{y}|\pi, \alpha) \end{aligned}$$

donde

$$\pi = (\pi_1, \ldots, \pi_T), \qquad \pi_t = \frac{\lambda_t}{\alpha + \sum_{t=1}^{T} \lambda_t}, \qquad 1 - \sum_{t=1}^{T} \pi_t = \frac{\alpha}{\alpha + \sum_{t=1}^{T} \lambda_t}.$$

La distribución multinomial negativa es una distribución definida para vectores con entradas en los enteros no negativos; cualesquiera dos componentes del vector con esta distribución tienen correlaciones positivas (a diferencia de un vector con distribución multinomial que presenta cualesquiera dos componentes con correlaciones negativas); y cualquier componente presenta varianza mayor a la media (distinto a una variable aleatoria con distribución Poisson con media y varianza iguales). Para más detalles de las propiedades de la distribución multinomial negativa ver apéndice B.

Uno de los intereses primordiales al analizar tablas de contingencia es probar si las variables que la definen son independientes. En la secciones 3.2.1 y 3.2.2 se presentaron algunas pruebas estadísticas utilizadas para este tipo de hipótesis mediante métodos Bayesianos. En esta sección se presenta un ejemplo de la aplicación de una de estas pruebas.

Waller y Zelterman (1997) presentan la incidencia de cáncer de 1989 en tres grandes ciudades de Ohio, clasificadas de acuerdo al sitio en el que se

Ciudad	1	2	3	4	5	6	7	8	9
Cleveland	71	1052	1258	440	488	159	523	169	268
Cincinnati	52	786	988	270	337	133	378	107	160
Columbus	41	517	715	190	212	91	254	77	137

Tabla 4.1: Muertes de cáncer en tres grandes ciudades de Ohio en 1989 (*National Center for Health Statistics*, 1990). Los sitios principales del tumor y los códigos *ICD-9* (entre paréntesis) son los siguientes: 1 = cavidad oral (140-149); 2 = órganos digestivos y colon (150-159); 3 = pulmón (160-165); 4 = seno (174-175); 5 = genitales (179-187); 6 = órganos urinarios (188-189); 7 = otros y sitios no especificados (170-173, 190-199); 8 = leucemia (204-208); y 9 = tejido linfático (200-203). El área metropolitana de Cincinnati incluye porciones de Kentucky e Indiana.

encuentra el tumor principal de esta enfermedad; la tabla 4.1 presenta los datos.

Los datos tienen una media de 361 y una varianza de $111,760$, además todas las correlaciones entre las ciudades y entre los sitios del tumor principal son positivas; esto implica que una distribución adecuada para los datos puede ser la distribución multinomial negativa.

La prueba utilizada para analizar esta tabla de contingencias fue realizada por Gutiérrez Peña (2005) y se presentó en la sección 3.2.1. Para ésta se obtiene la distribución final de

$$\delta = \delta(\pi) \equiv \sum_l \log\left(\frac{\tilde{\pi}_l}{\tilde{\pi}_l^0}\right)\log(\tilde{\pi}_l);$$

si los valores de δ están concentradas alrededor del cero entonces se concluye que hay independencia.

Un factor de interés que puede surgir al analizar esta tabla es saber si las incidencias para cada una de las ciudades son independientes.

Bajo el supuesto de que los datos tienen una distribución multinomial negativa, en el siguiente programa de WinBUGS se realizaron 100,000 simulaciones de la distribución final de δ:

```
# Modelo
model{
# Familia paramétrica (verosimilitud)
for(t in 1:T){
y[t] ~ dpois(theta[t])
theta[t] <- gamma*lambda[t]
}
```

```
gamma ~ dgamma(alpha,beta)
alpha ~ dgamma(1,1)
# Distribución inicial
for(t in 1:T){
pi[t] <- lambda[t]/(sum(lambda[])+beta)
lambda[t] ~ dgamma(1,1)
}
beta ~ dgamma(1,1)
pi[T+1] <- beta/(sum(lambda[])+beta)
# Prueba de independencia
for(i in 0:I-1){
pimr[i+1] <- pi[i*9+1] + pi[i*9+2] + pi[i*9+3] + pi[i*9+4]
+ pi[i*9+5] + pi[i*9+6] + pi[i*9+7] + pi[i*9+8] + pi[i*9+9]
}
for(j in 1:J){
pimc[j] <- pi[j] + pi[9+j] + pi[18+j]
}
for(i in 1:I){
for(j in 1:J){
pi0[9*(i-1)+j] <- pimr[i]*pimc[j]
}}
pi0[T+1] <- 1-sum(pi0[1:T])
for(t in 1:T+1){
d[t] <- (log(pi[t]/pi0[t]))*(log(pi[t]))
}
delta <- sum(d[])
}

# Datos
list(T=27, I=3, J=9, y = c(71, 1052, 1258, 440, 488, 159, 523,
169, 268, 52, 786, 988, 270, 337, 133, 378, 107, 160, 41, 517,
715, 190, 212, 91, 254, 77, 137))

#Iniciales
list(gamma=1, alpha=1, lambda = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1), beta=1)
```

La gráfica 4.1 muestra la distribución final de δ. Dado que el valor medio de δ es 6.5 y los cuantiles del 2.5 % y 97.5 % son 5.14 y 8.27, respectivamente, entonces se puede concluir que las variables no son independientes, por lo

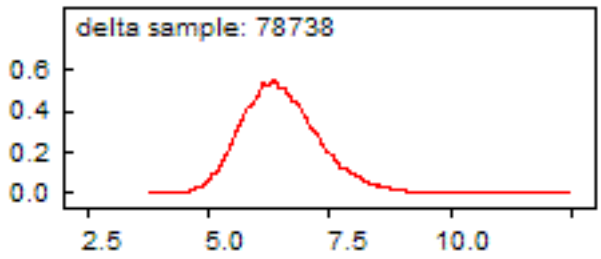

Figura 4.1: Densidad de δ.

que existe una dependencia entre las incidencias de los tipos de cáncer de las ciudades, es decir, si hay mayor incidencia de una enfermedad en una ciudad, entonces se espera una mayor incidencia de las otras enfermedades.

4.2. Modelos para Datos de Series de Tiempo

Los modelos de series de tiempo para datos discretos no han sido muy explotados en comparación con otros modelos. Estos modelos, si bien conceptualmente y en algunos casos matemáticamente son innovadores, también son generalmente restrictivos. Debido a esto, no es claro que exista un modelo dominante para datos discretos de series de tiempo. A continuación se presenta una descripción general de los modelos de series de tiempo para datos continuos y posteriormente para datos discretos, enfocándose en un caso particular de éstos.

4.2.1. Modelos Lineales

Cuando se estudia una variable dependiente continua, los modelos estándar de series de tiempo están bien establecidos. Para *series de tiempo puras*, donde las variables explicativas están definidas por los valores de rezago de la variable dependiente, la clase estándar de modelos lineales es el modelo *autoregresivo de promedios móviles* de orden (p, q), ARMA(p,q); para éstos el valor corriente de Y es igual a la suma ponderada de los p valores pasados, los q valores pasados y el corriente de un error aleatorio independiente e idénticamente distribuido, esto es

$$y_t = \rho_1 y_{t-1} + \cdots + \rho_p y_{t-p} + \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \cdots + \gamma_q \varepsilon_{t-q}$$

para $t = p+1, \ldots, T$, donde ε_t es una variable aleatoria independiente idénticamente distribuida con una cierta distribución $D(0, \sigma^2)$ (usualmente una distribución normal).

Para un modelo de *regresión lineal de series de tiempo* las variables explicativas incluyen regresores exógenos. Los modelos *autoregresivos* o *dinámicos* incluyen regresores exógenos y variables dependientes de rezago en la función de regresión. Un ejemplo es

$$Y_t = \rho Y_{t-1} + \mathbf{x}'_t\beta + \varepsilon_t,$$

donde el error ε_t es una variable aleatoria independiente e idénticamente distribuida con distribución $D(0, \sigma^2)$. Note que este modelo es equivalente a suponer que, dada $Y_{t-1} = y_{t-1}$,

$$Y_t \sim D\left(\rho y_{t-1} + \mathbf{x}'_t\beta, \sigma^2\right),$$

esto es, Y_t dada $Y_{t-1} = y_{t-1}$ se distribuye con media $\rho y_{t-1} + \mathbf{x}'_t\beta$ y varianza σ^2. Si sólo aparecen $\mathbf{x}_t$ y los rezagos de $\mathbf{x}_t$, el modelo es de *rezago distribuido*. Pero, si sólo aparece $\mathbf{x}_t$ como regresor, el modelo es *estático*.

Un modelo de regresión de series de tiempo alternativo es el modelo de *error correlacionado serialmente*. Este comienza con una función de regresión estática

$$Y_t = \mathbf{x}'_t\beta + u_t$$

pero supone que el error u_t está serialmente correlacionado, siguiendo por ejemplo un proceso ARMA. El caso más simple es un error autoregresivo de orden uno, AR(1),

$$u_t = \rho u_{t-1} + \varepsilon_t$$

donde ε_t es una variable aleatoria independiente idénticamente distribuida con distribución $D(0, \sigma^2)$. Entonces el modelo puede reescribirse como

$$Y_t = \rho Y_{t-1} + \beta\mathbf{x}'_t - \beta\rho\mathbf{x}'_{t-1} + \varepsilon_t$$

el cual es un modelo autoregresivo con restricciones no lineales impuestas sobre los parámetros.

Los modelos autoregresivos y de correlación serial pueden combinarse, para producir un modelo autoregresivo con error correlacionado serialmente.

Los modelos de series de tiempo anteriores son los modelos más comunes para variables continuas. Sin embargo, en la literatura puede encontrarse una amplia variedad de éstos y una explicación más detallada. Algunos de éstos pueden estudiarse en Brockwell y Davis (1998, 2002), y Chatfield (2003).

4.2.2. Modelos Discretos

A lo largo del estudio de series de tiempo para datos discretos se han obtenido varios posibles modelos a través de definir de diferente manera la dependencia de Y_t sobre el valor pasado de Y_t, el valor corriente y el pasado de $\mathbf{x}_t$, y los procesos latentes y procesos del error ε_t; a través de diferentes modelos de los procesos latentes; y a través de diferentes extensiones de los modelos básicos.

No existe un sistema simple de clasificación que contenga a todos los modelos de series de tiempo para datos discretos, sin embargo la siguiente clasificación, propuesta por Cameron y Trivedi (1998), puede dar un margen para diferenciar las clases de modelos.

1. Modelos ARMA de valores enteros (INARMA). Definen a y_t como la suma de un entero cuyo valor se determina por el pasado de y_t y un término de innovación independiente. Las distribuciones apropiadas que encabezan las distribuciones marginales de Y_t son tales como la Poisson y la binomial negativa.

2. Modelos autoregresivos o modelos de Markov. Definen la distribución condicional de Y_t como una distribución discreta tal como una Poisson o una binomial negativa, cuyo parámetro de media es una función de los valores de rezago de Y_t.

   En este modelo la distribución condicional de Y_t está especificada, mientras que en el modelo INARMA se especifica la distribución marginal de Y_t.

3. Modelos de errores correlacionados serialmente o modelos de variables latentes. Este modelo define a Y_t de tal manera que dependa de un término estático y un error correlacionado serialmente.

4. Modelos de espacio de estados o modelos de parámetros que varían en el tiempo (*state-space models* o *time-varying parameter models*). Especifican la distribución de Y_t que sea una distribución discreta tal como la Poisson o la binomial negativa, cuya media condicional o parámetros de la media condicional dependan de sus valores en periodos previos.

5. Modelos de Markov ocultos o modelos de cambio de régimen (*hidden Markov models* o *regime shift models*). Especifican la distribución de Y_t que sea una distribución discreta tal como la distribución Poisson o la binomial negativa, con parámetros que varían de acuerdo a cuál, de un

número finito de regímenes, está actuando en un momento determinado. Los regímenes no observados se desarrollan en el tiempo a través de una cadena de Markov.

6. Modelos ARMA discretos (DARMA). Introducen la dependencia del tiempo a través de una mezcla.

Algunos autores clasifican a los modelos como *observation-driven*, con series de tiempo dependientes introducidas por la especificación de los momentos condicionales o las densidades como funciones explícitas de los resultados pasados, o *parameter-driven*, con dependencias inducidas por un proceso de variables latentes.

Otros hacen la diferencia entre modelos *condicionales*, donde los momentos o las densidades están condicionados por $\mathbf{x}_t$ y los resultados pasados de Y_t, y modelos *marginales*, donde se condiciona sólo sobre $\mathbf{x}_t$ y no sobre los resultados pasados de Y_t. Esta es la clasificación más usada para distinguir entre una clase de modelos y otro.

Ejemplo

Como una ilustración de cómo las trayectorias de un proceso autoregresivo de valores enteros de rezago 1 (INAR(1)) difieren de un proceso estándar autoregresivo de rezago 1 (AR(1)), se simuló una muestra de cada uno de los procesos. Para realizar la comparación de las dos trayectorias muestrales, los dos procesos simulados se realizaron de tal manera que tuvieran medias y varianzas iguales. La figura 4.2 (izquierda) presenta las trayectorias de una muestra simulada de 100 observaciones generadas del modelo INAR(1) (ver sección 4.3), en el cual $\alpha = 0.5$ y I_i tiene una distribución Poisson con parámetro $\iota = 1$. La figura 4.2 (derecha) muestra las trayectorias de una muestra simulada de 100 observaciones del modelo estándar AR(1), en el cual I_t toma una distribución normal con media $\iota = 1$ y varianza $\iota(1 + \alpha)$.

Puede verse que al comparar las trayectorias de la figura 4.2 las realizaciones del modelo INAR(1) toman valores enteros no negativos, mientras que las realizaciones del modelo AR(1) son valores reales y algunos de estos son negativos. Como una consecuencia de estas observaciones y de los requerimientos de estacionariedad, puede verse que las realizaciones del proceso INAR(1) tienen muchos valores cercanos a la media (2). Sin embargo, no existe algún patrón en las corridas del proceso AR(1).

4.3. INAR(1)

Los modelos de series de tiempo autoregresivos de valores enteros (INAR) especifican los valores observados de Y_t como los de variables aleatorias de

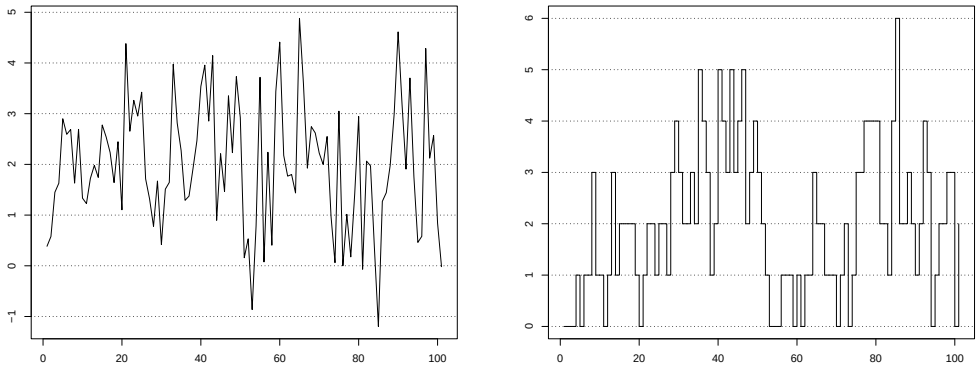

Figura 4.2: Simulación (izquierda) modelo INAR(1) y (derecha) modelo AR(1).

datos de conteo cuyos valores dependen de los resultados pasados y la realización de una variable aleatoria de datos de conteo independiente e idénticamente distribuida I_t cuyos valores no dependen de los resultados pasados. Este modelo es similar al modelo lineal $Y_t = \rho Y_{t-1} + I_t$, sólo que la operación de multiplicar Y_{t-1} por un escalar ρ se reemplaza por una variable aleatoria discreta que depende del valor y_{t-1}. Las diferentes elecciones de la distribución I_t conducen a diferentes distribuciones marginales para Y_t, tal como una distribución Poisson o binomial negativa. Estos modelos tienen la misma estructura de correlación serial que la estructura de los modelos lineales AR para datos continuos. Comenzamos con el estudio de series de tiempo puras, antes de introducir otras variables regresoras.

Los modelos INAR(1) presentados a continuación se describen en Al-Osh y Alzaid (1987), McKenzie (1988), Cameron y Trivedi (1998), y Böckenholt (1999a), y en Böckenholt (1999b) y Zeger (1988) se describen aplicaciones de los modelos de regresión INAR(1).

Los datos que están expresados en términos de conteos tomados secuencialmente en el tiempo, y que están correlacionados, surgen en muchas ocasiones. Ejemplos de este proceso son el número de pacientes en un hospital en un punto específico del tiempo, o el número de personas en una cola de espera para el servicio en un cierto momento. Éstos ejemplifican la acción o influencia de un elemento del proceso en tiempos previos, y una sucesión de arribo (innovación) que tiene una cierta distribución discreta. El propósito de esta sección es modelar tales procesos de una manera análoga a los procesos autoregresivos (AR), en los cuales la sucesión de innovación, por conveniencia

estadística, tiene una distribución normal.

Antes de introducir los modelos INAR(1), presentamos la definición del proceso de adelgazamiento binomial (*binomial thinning procedure*) $\alpha \circ G$ donde el operador '$\circ$' se llama operador de adelgazamiento (*thinning operator*) definido por Steutel y Van Harn (1979).

Sea G una variable aleatoria discreta definida sobre valores enteros no negativos, y sea H_j una sucesión de variables aleatorias binarias independientes e idénticamente distribuidas, independientes de G, tal que $p(H_j = 1) = 1 - p(H_j = 0) = \alpha$ donde $\alpha \in [0, 1]$. El proceso de adelgazamiento binomial está definido por

$$\alpha \circ G = \sum_{j=1}^{G} H_j = B(\alpha, G).$$

$B(\alpha, G)$ es una variable aleatoria binomial que se basa en G ensayos con probabilidad de éxito α.

Definimos un proceso INAR(1) $\{Y_t;\ t = \ldots, -2, -1, 0, 1, 2, \ldots\}$ como:

$$Y_t = \alpha \circ Y_{t-1} + I_t$$

donde $\alpha \in [0, 1]$ y I_t es una sucesión de variables aleatorias con valores enteros no negativos no correlacionados, con media μ y varianza σ^2.

Este modelo INAR(1) simplemente establece que los componentes del proceso al tiempo t, Y_t, son: (i) los supervivientes de los elementos del proceso al tiempo $t-1$, Y_{t-1}, cada uno con probabilidad de supervivencia α y (ii) elementos que entran al sistema en el intervalo $(t-1, t]$ como términos de innovación (I_t).

Note que el operador '$\circ$' tiene las siguientes propiedades: $0 \circ Y = 0$, $1 \circ Y = Y$, $E[\alpha \circ Y] = \alpha E[Y]$ y finalmente, si $\beta \in [0, 1]$, entonces $\beta \circ (\alpha \circ Y) \stackrel{d}{=} (\beta\alpha) \circ Y$, además,

$$\begin{aligned} Y_t &= \alpha_t \circ Y_{t-1} + I_t \\ &= \left(\prod_{j=0}^{k-1} \alpha_{t-j} \right) \circ Y_{t-k} + \sum_{j=1}^{k-1} \left(\prod_{l=0}^{j-1} \alpha_{t-l} \right) \circ I_{t-j} + I_t. \end{aligned}$$

4.3.1. Modelos INAR(1)-Poisson

McKenzie (1988) define los modelos INAR(1) y otros modelos ARMA para datos con distribución Poisson. Sea y_{it} los conteos para una persona i

en el tiempo t (i=1,...,I; t=1,...,T). Cuando estos datos de conteo están generados por un proceso Poisson a tiempo discreto se obtiene que

$$p(y_{it}) = \frac{\exp(-\lambda_{it})\lambda_{it}^{y_{it}}}{y_{it}!}. \tag{4.1}$$

Por simplicidad de notación, a partir de aquí se suprimirá el subíndice i. Un modelo autoregresivo Poisson para T periodos de tiempo puede obtenerse aplicando el proceso de adelgazamiento binomial que puede considerarse análogo a una multiplicación escalar de valores enteros. Aplicando la operación de adelgazamiento binomial los conteos y_t pueden descomponerse en dos enteros no negativos. Una parte, C_{t-1}, es el componente *carry-over* de y_{t-1}, y la otra parte, I_t que es el componente de innovación, refleja la influencia del periodo de tiempo actual, es decir,

$$y_t = C_{t-1} + I_t$$

donde $C_{t-1} = \alpha_t \circ y_{t-1}$, esto es,

$$y_t = \alpha_t \circ y_{t-1} + I_t. \tag{4.2}$$

Un proceso estocástico autoregresivo de primer orden de valores enteros (INAR(1)) con distribuciones marginales Poisson(λ_t) se obtiene cuando Y_0 y I_t son independientes con distribución Poisson con parámetros λ_0 y $\iota_t = \lambda_t - \alpha_t\lambda_{t-1}$, respectivamente.

Cuando Y_0 y I_t son variables aleatorias independientes con distribución Poisson con parámetros λ y $\iota_t = \iota = \lambda(1-\alpha)$, entonces $\{Y_t;\ t = 1, 2, \ldots\}$ es un proceso de Markov estacionario con distribuciones marginales Poisson.

Las probabilidades de transición de este proceso están dadas por la convolución de una variable aleatoria Poisson y una binomial,

$$\begin{aligned} p(y_t|y_{t-1}) &= \begin{cases} \sum_{k=0}^{y_t} \binom{y_{t-1}}{k} \alpha^k (1-\alpha)^{y_{t-1}-k} e^{-\iota} \frac{\iota^{y_t-k}}{(y_t-k)!} & \text{si } y_t \leq y_{t-1} \\ \sum_{k=0}^{y_{t-1}} \binom{y_{t-1}}{k} \alpha^k (1-\alpha)^{y_{t-1}-k} e^{-\iota} \frac{\iota^{y_t-k}}{(y_t-k)!} & \text{si } y_t > y_{t-1} \end{cases} \\ &= y_{t-1}! \exp(-\iota)\iota^{y_t}(1-\alpha)^{y_{t-1}} \sum_{k=0}^{\text{mín}(y_{t-1},y_t)} w_k \left(\frac{\alpha}{(1-\alpha)\iota}\right)^k \end{aligned} \tag{4.3}$$

donde $w_k = ((y_{t-1}-k)!(y_t-k)!k!)^{-1}$.

Esto se debe a que el modelo presenta la estructura de un proceso de Markov de primer orden,

$$p(y_t|y_{t-1}, y_{t-2}, \ldots) = p(y_t|y_{t-1}),$$

y la distribución conjunta para T periodos de tiempo es

$$p(y_1, y_2, \ldots, y_T) = p(y_1) \prod_{t=2}^{T} p(y_t|y_{t-1}),$$

donde $p(y_1)$ está dada por (4.1) y $p(y_t|y_{t-1})$ por (4.3).

La media y la varianza de $\{Y_t\}$ definido en (4.2) son:

$$E[Y_t] = \alpha E[Y_{t-1}] + \iota = \alpha^t E[Y_0] + \sum_{j=0}^{t-1} \alpha^j \iota$$

y

$$\begin{aligned} Var[Y_t] &= Var[E[\alpha \circ Y_{t-1}|Y_{t-1}]] + E[Var[\alpha \circ Y_{t-1}|Y_{t-1}]] + Var[I_t] \\ &= \alpha^{2t} Var[Y_0] + (1-\alpha) \sum_{j=1}^{t} \alpha^{2j-1} E[Y_{t-j}] + \iota \sum_{j=1}^{t} \alpha^{2(j-1)}. \end{aligned}$$

Para cualquier entero no negativo k, podemos calcular la covarianza con rezago k, $\gamma(k) = Cov[Y_{t-k}, Y_t]$,

$$\begin{aligned} \gamma(k) &= Cov\left[Y_{t-k}, \alpha^k \circ Y_{t-k}\right] + \sum_{j=0}^{k-1} Cov\left[Y_{t-k}, \alpha^j \circ I_{t-j}\right] \\ &= \alpha^k Var[Y_{t-k}] \\ &= \alpha^k \gamma(0). \end{aligned}$$

4.3.2. Modelos INAR(1)-Binomial Negativa

Para tomar en cuenta la sobredispersión de Y_t, puede considerarse un proceso INAR(1)-binomial negativo. La distribución marginal de este proceso es binomial negativa con media $\beta\theta$ y varianza $\beta\theta(1+\beta)$, BN(θ,β). Consecuentemente, el proceso INAR(1)-BN para $\{Y_t;\ t = 1, 2, \ldots\}$ se define como $Y_t = \Pi \circ Y_{t-1} + I_t$, donde Π sigue una distribución beta con parámetros $\alpha\theta$ y $(1-\alpha)\theta$, $(0 < \alpha < 1)$, y Y_0 y I_t son independientes con distribución BN(θ,β) y BN($\theta(1-\alpha)$,β), respectivamente .

Un forma más simple de obtener el proceso INAR(1)-BN es a través del proceso INAR(1)-Poisson permitiendo que el parámetro del componente de innovación varíe de acuerdo a una distribución gamma con parámetros $\theta^I = \theta(1-\alpha)$ y β y la probabilidad del proceso de adelgazamiento binomial de acuerdo a una distribución beta con parámetros $\theta^C = \alpha\theta$ y θ^I. Componiendo la parte de innovación de la distribución Poisson con la distribución

gamma se produce una distribución BN(θ^I,β), y componiendo la distribución binomial y la beta se produce una distribución beta-binomial con media (αy_{t-1}) y varianza ($y_{t-1}\alpha(1-\alpha)[(\theta+y_{t-1})/(\theta+1)]$). De esta manera sólo se requiere un parámetro más (β) que en el modelo INAR(1)-Poisson. El proceso INAR(1)-BN resulta entonces en una representación analítica parsimoniosa de las distribuciones marginales y condicionales de Y_t.

Haciendo que θ, θ^I, θ^C y α varíen en el tiempo se denotan por θ_t, θ_t^I, $\theta_{t-1,t}^C$ y $\alpha_{t-1,t}$. Entonces la distribución condicional de Y_t dada $Y_{t-1}=y_{t-1}$, la cual es una convolución de una distribución binomial negativa y una distribución beta-binomial, puede escribirse como

$$\begin{aligned} p(y_t|y_{t-1}) &= \frac{\beta^{y_t} y_{t-1}!\Gamma(\theta_{t-1})}{\Gamma(\theta_{t-1,t}^C)\Gamma(\theta_{t-1}^I)\Gamma(\theta_t^I)\Gamma(\theta_{t-1}+y_{t-1})(1+\beta)^{y_{t-1}+\theta_t^I}} \\ &\times \sum_{k=0}^{\min(y_{t-1},y_t)} w_k \Gamma(\theta_{t-1,t}^C+k)\Gamma(\theta_{t-1}^I+y_{t-1}-k) \\ &\times \Gamma(\theta_t^I+y_t-k), \end{aligned} \tag{4.4}$$

donde $w_k=\{[(1+\beta)/\beta]^k\}/\{(y_{t-1}-k)!(y_t-k)!k!\}$, $\theta_{t-1,t}^C=\alpha_{t-1,t}\sqrt{\theta_{t-1}\theta_t}$ y $\theta_t^I=\theta_t-\theta_{t-1,t}^C$. A partir de esta representación, es claro que la autocorrelación entre α_{t-1} y α_t sólo puede ser positiva. Una propiedad importante de la distribución condicional es que su función de regresión es lineal,

$$E(Y_t|y_{t-1}) = \beta\theta_t^I + \frac{\theta_{t-1,t}^C}{\theta_{t-1}} y_{t-1},$$

y la varianza condicional está dada por

$$V(Y_t|y_{t-1}) = \beta\theta_t^I(1+\beta) + y_{t-1}\frac{\theta_{t-1,t}^C}{\theta_{t-1}}\left(1-\frac{\theta_{t-1,t}^C}{\theta_{t-1}}\right)\frac{\theta_{t-1}+y_{t-1}}{\theta_{t-1}+1}.$$

Debido a que este modelo tiene la estructura de un proceso de Markov de primer orden, se tiene que $p(Y_t|y_{t-1},y_{t-2},\ldots)=p(Y_t|y_{t-1})$, y la distribución conjunta para T periodos de tiempo es $p(y_1,y_2,\ldots,y_T)=p(y_1)\prod_{t=2}^T p(y_t|y_{t-1})$, donde $p(y_1)$ está especificada como una distribución BN(θ_1,β) y la probabilidad $p(y_t|y_{t-1})$ se obtiene por medio de (4.4).

4.3.3. Modelos INAR(1)-Multiomial Negativa

La extensión del proceso INAR(1)-BN a R eventos es sencilla cuando los parámetros de cada uno de los eventos varían de acuerdo a la misma distribución gamma con parámetros θ y β_r ($r=1,\ldots,R$). En este caso la

distribución conjunta de las variables de conteo están dadas por una distribución multinomial negativa con R variables (MN), la cual puede factorizarse en una distribución multinomial y una binomial negativa:

$$\begin{aligned} p(y_{1t},\ldots,y_{Rt}) &= \binom{y_{+t}}{y_{1t},\ldots,y_{Rt}}\left(\frac{\beta_1}{\beta_+}\right)^{y_{1t}}\cdots\left(\frac{\beta_R}{\beta_+}\right)^{y_{Rt}} \\ &\times \frac{\Gamma(\theta+y_{+t})}{\Gamma(\theta)y_{+t}!}(1+\beta_+)^{-\theta}\left(\frac{\beta_+}{1+\beta_+}\right)^{y_{+t}}, \end{aligned} \tag{4.5}$$

donde $\beta_+ = \sum_{r=1}^{R}\beta_r$, y $y_{+t} = \sum_{r=1}^{R} y_{rt}$. Por otra parte, la distribución condicional de $(Y_{1t},\ldots,Y_{Rt})$ dada $(y_{1,t-1},\ldots,y_{R,t-1})$ está dada por

$$\begin{aligned} p(y_{1t},\ldots,y_{Rt} \mid y_{1,t-1},\ldots,y_{R,t-1}) &= \frac{\prod_{r=1}^{R}(\frac{\beta_r}{1+\beta_r})^{y_{r,t-1}+y_{rt}}}{\Gamma(\theta^C)\Gamma(\theta^I)\Gamma(\theta^I)(1+\beta_+)^{\theta+\theta^I}} \\ &\times \sum_{k_1=0}^{\min(y_{1,t-1},y_{1t})}\cdots\sum_{k_R=0}^{\min(y_{R,t-1},y_{Rt})} w_{k_1}\cdots w_{k_R}(1+\beta_+)^{k_+} \\ &\times \Gamma(\theta^C+k_+)\Gamma(\theta^I+y_{+,t-1}-k_+)\Gamma(\theta^I+y_{+t}-k_+), \end{aligned} \tag{4.6}$$

donde $w_{k_r} = (\beta_r^{k_r}(y_{r,t-1}-k_r)!(y_{rt}-k_r)!k_r!)^{-1}$ y $k_+ = \sum_{r=1}^{R} k_r$.

La estructura de la media y de la covarianza del modelo INAR(1)-MN es bastante restrictiva. La esperanza de $\mathbf{y}_t = (y_{1t},\ldots,y_{Rt})$ es $E[\mathbf{y}_t] = \beta\theta$, donde $\beta = (\beta_1,\ldots,\beta_R)$. Así para cada periodo de tiempo, los efectos medios son proporcionales. Similarmente, puede mostrarse que la matriz de covarianzas $\Sigma_{\mathbf{y}}$ de orden $(RT \times RT)$ de $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_T)$ es un producto de Kronecker de dos matrices, $\Sigma_{\mathbf{y}} = \Sigma_\beta \otimes \Sigma_\theta$, donde $\Sigma_\beta = \beta\beta' + diag(\beta)$ y

$$\Sigma_{\theta(T\times T)} = \begin{bmatrix} \theta & \theta^C & \cdots & \theta^C \\ \theta^C & \theta & \cdots & \theta^C \\ \vdots & \vdots & \ddots & \vdots \\ \theta^C & \theta^C & \cdots & \theta \end{bmatrix}.$$

4.3.4. Ejemplo

Zeger (1988) presenta una lista del número de casos de poliomielitis reportados mensualmente por el *U.S. Centers for Disease Control* para los años de 1970 a 1983 y que fueron publicados en *Morbidity and Mortality Weekly Report Annual Summary*; éstos se encuentran en la tabla 4.2.

Los datos mensuales se encuentran graficados en la figura 4.3. La figura 4.4 presenta el histograma de los datos. Los datos provienen de una distribución asimétrica. Además, durante los 168 meses se presentaron 224 casos de

poliomielitis, lo que indica una media de 1.333 casos por mes y una varianza de 3.505. Esto sugiere que los datos pueden tener una distribución binomial negativa, en lugar de una distribución Poisson, debido a que la varianza es mayor a la media. Note también que estos datos podrían tener una distribución Poisson cero-inflada (ver Hall, 2000 y Lambert, 1992), sin embargo, nuestro interés es mostrar el uso de la distribución multinomial negativa, y por lo tanto en este trabajo no exploramos el uso de la distribución Poisson cero-inflada.

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
1970	0	1	0	0	1	3	9	2	3	5	3	5
1971	2	2	0	1	0	1	3	3	2	1	1	5
1972	0	3	1	0	1	4	0	0	1	6	14	1
1973	1	0	0	1	1	1	1	0	1	0	1	0
1974	1	0	1	0	1	0	1	0	1	0	0	2
1975	0	1	0	1	0	0	1	2	0	0	1	2
1976	0	3	1	1	0	2	0	4	0	2	1	1
1977	1	1	0	1	1	0	2	1	3	1	2	4
1978	0	0	0	1	0	1	0	2	2	4	2	3
1979	3	0	0	2	7	8	2	4	1	1	2	4
1980	0	1	1	1	3	0	0	0	0	1	0	1
1981	1	0	0	0	0	0	1	2	0	2	0	0
1982	0	1	0	1	0	1	0	2	0	0	1	2
1983	0	1	0	0	0	1	2	1	0	1	3	6

Tabla 4.2: Número de casos mensuales de poliomielitis de E.U. de 1970 a 1983.

Primeramente, como una muestra de que la distribución Poisson no siempre es la distribución adecuada, se ajustó un modelo INAR(1)-Poisson estacionario. Aplicando el programa en WinBUGS para el modelo especificado en la sección 4.3.1 se obtuvo el modelo ajustado

$$Y_t = \alpha \circ Y_{t-1} + I_t, \qquad t = 1, \ldots, 168,$$

donde

$$Y_0 \sim Poisson(\lambda) \quad y \quad I_t \sim Poisson(\lambda(1-\alpha)) \quad \text{independientes},$$

y usando el programa se obtuvo que $\alpha = 0.3369$ y $\lambda = 1.219$, estos resultados se muestran en la tabla 4.3.

La figura 4.5 muestra las frecuencias relativas obtenidas después de hacer $1,000$ simulaciones realizadas en R. Es claro que este modelo no da un buen ajuste. El error absoluto (diferencia entre la frecuencia relativa estimada y la frecuencia relativa observada) de este ajuste es 0.2764048.

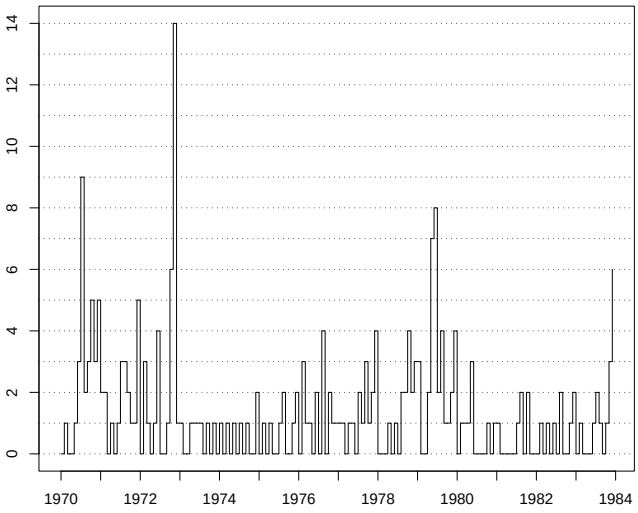

Figura 4.3: Número de los casos mensuales de poliomielitis de E.U. de 1970 a 1983.

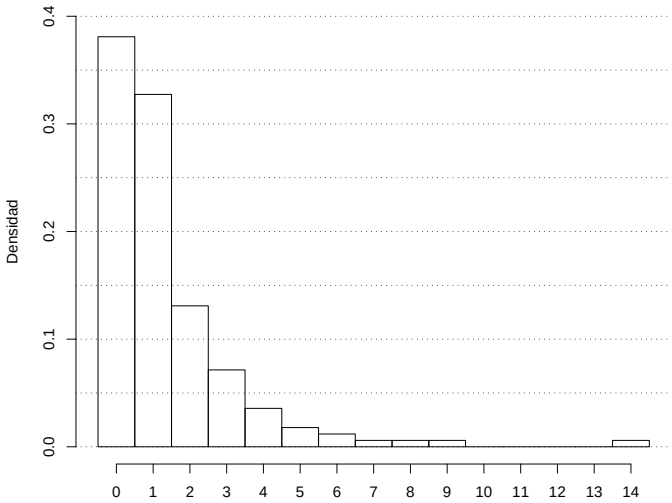

Figura 4.4: Histograma del número de los casos mensuales de poliomielitis de E.U. de 1970 a 1983.

node	mean	sd	MC error	2.5 %	median	97.5 %
α	0.3369	0.06163	7.417E-4	0.2222	0.3352	0.4631
λ	1.219	0.1416	8.914E-4	0.9655	1.21	1.523

Tabla 4.3: Resumen de la distribución final obtenido en WinBUGS para el modelo INAR(1)-Poisson de los datos de poliomielitis.

Lo anterior sugiere que podemos encontrar un mejor ajuste, para lo cual utilizaremos la distribución binomial negativa (note que la observación de noviembre de 1972 es muy alta, si eliminaramos este valor y realizando una prueba estadística no parámetrica ji-cuadrada comprobamos que los datos tienen una distribución binomial negativa con un nivel de significancia de $p = 0.5$, ver Gibbons y Chakraborti, 2003).

Para el ajuste del modelo INAR(1)-BN especificado en la sección 4.3.2 se requieren los parámetros de la distribución binomial negativa. Para esto se realizaron dos programas en WinBUGS para obtener los parámetros que ajustan estos datos a una distribución binomial negativa con índice r y probabilidad π. Los programas se presentan a continuación.

```
# Modelo 1
model{
for(k in 1:K){
y[k] ~ dpois(theta[k])
theta[k] ~ dgamma(r,be)
}
be <- pi/(1-pi)
pi ~ dbeta(1,1)
r ~ dgamma(1,1)
}

# Modelo 2
model{
for(k in 1:K){
y[k] ~ dpois(theta[k])
theta[k] <- gamma[k]*lambda
gamma[k] ~ dgamma(r,be)
}
pi <- be/(lambda+be)
lambda ~ dgamma(1,1)
be ~ dgamma(1,1)
```

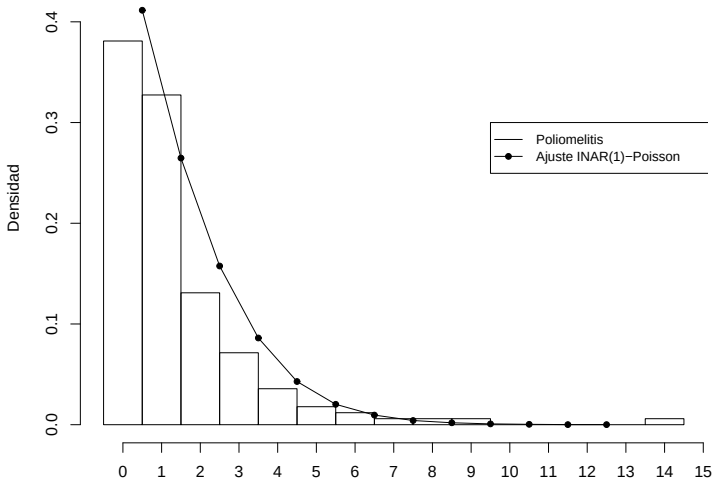

Figura 4.5: Histograma del número de los casos mensuales de poliomielitis de E.U. de 1970 a 1983 y el ajuste INAR(1)-Poisson.

```
r ~ dgamma(1,1)
}
```

Después de realizar 100,000 iteraciones se obtuvo que los parámetros son $r = 1.211$ y $\pi = 0.4704$ para el modelo 1, y $r = 1.211$ y $\pi = 0.4705$ para el modelo 2, respectivamente. Por tal motivo, para el ajuste del modelo INAR(1)-BN se utilizaron los parámetros $r = 1.211$ y $\pi = 0.4705$.

Posteriormente se realizó un programa en WinBUGS para obtener el modelo INAR(1)-BN, utilizando que $\theta = r$ y $\beta = (1 - \pi)/\pi$, el cual se presenta a continuación:

```
model{
y[1] ~ dpois(lambda)
lambda ~ dgamma(theta,betai)
PPI ~ dbeta(thetaC,thetaI)
for(t in 2:T){
for(i in 1:y[t-1]+1){
```

```
H[t,i] ~ dbern(PPI)
}
aG[t] <- sum(H[t,1:y[t-1]+1]) - H[t,1]
mu[t] <- aG[t] + thetaI*beta
y[t] ~ dpois(mu[t])
}
thetaI <- theta*(1-alpha)
thetaC <- theta*alpha
betai <- 1/beta
beta <- (1-pi)/pi
alpha ~ dbeta(1,1)
pi <- 0.4705
theta <- 1.211
}
```

Los resultados obtenidos se presentan en la tabla 4.4.

node	mean	sd	MC error	2.5 %	median	97.5 %	start	sample
alpha	0.3947	0.06674	4.075E-4	0.2589	0.3964	0.5202	1001	100000
thetaC	0.478	0.08082	4.935E-4	0.3135	0.4801	0.63	1001	100000
thetaI	0.733	0.08082	4.935E-4	0.581	0.7309	0.8975	1001	100000

Tabla 4.4: Resultados obtenidos por el programa WinBUGS para el ajuste del modelo INAR(1)-BN para los datos de poliomielitis de E.U. de 1970 a 1983.

Por tanto el modelo ajustado queda definido de la siguiente manera:

$$Y_t = \Pi \circ Y_{t-1} + I_t \qquad t = 1, \ldots, 168$$

donde

$$Y_0 \sim Poisson(\lambda) \qquad I_t \sim Poisson(\lambda^I) \qquad \text{independientes}$$

$$\Pi \sim Beta(\theta^C, \theta^I) \qquad \lambda^I \sim Gamma(\theta^I, \beta)$$

con $\theta^C = \theta\alpha$, $\theta^I = \theta(1-\alpha)$, $\beta = (1-\pi)/\pi$ y $\theta = r$, donde, por los ajustes de la distribución binomial negativa, $r = 1.211$ y $\pi = 0.4705$.

Las frecuencias relativas del número de casos de poliomielitis y del ajuste se presentan en la tabla 4.5 (éstos presentan un error absoluto de 0.1638095).

Para visualizar estos resultados se simuló el modelo INAR(1)-BN en R con los resultados obtenidos; el histograma de los datos se presenta en la figura 4.6. Note que el ajuste obtenido mediante este modelo, suponiendo la

distribución binomial negativa, es mucho mejor que con el modelo anteriormente ajustado, en donde se suponía que los datos seguían una distribución Poisson.

Número de casos	Frecuencias relativas de los casos	Frecuencias relativas del ajuste INAR(1)-BN
0	0.380952381	4.114762e-01
1	0.327380952	2.647321e-01
2	0.130952381	1.575476e-01
3	0.071428571	8.609524e-02
4	0.035714286	4.288095e-02
5	0.017857143	2.024405e-02
6	0.011904762	9.553571e-03
7	0.005952381	4.160714e-03
8	0.005952381	1.928571e-03
9	0.005952381	8.154762e-04
10	0	3.809524e-04
11	0	1.011905e-04
12	0	8.333332e-05
13	0	0
14	0.005952380	0

Tabla 4.5: Frecuencias relativas del número de casos mensuales de poliomielitis de E.U. de 1970 a 1983 y ajuste INAR(1)-BN.

Debido a que la observación de noviembre de 1972 es muy alta comparada con los demás valores ($y_{15} = 14$), podría considerarse como un valor atípico, por lo que se realizó el mismo ajuste sin esta observación. Se obtuvo que el índice de la distribución binomial negativa es $\alpha = 1.46$ y la probabilidad $\pi = 0.529$.

Posteriormente se ajustó el modelo INAR(1)-BN en WinBUGS y finalmente se simularon los datos en R; el histograma de los datos se presenta en la figura 4.7 (éstos presentan un error absoluto de 0.1369461).

Note que estos datos presentan un mejor ajuste que el modelo en donde se consideraron todos los datos, y por tanto podríamos considerar a la observación $y_{15} = 14$ como un valor atípico.

4.4. Modelos de Regresión INAR(1)

En los capítulos 2 y 3 se estudiaron los modelos loglineales desde el punto de vista clásico y Bayesiano (estos modelos a veces se conocen como modelos

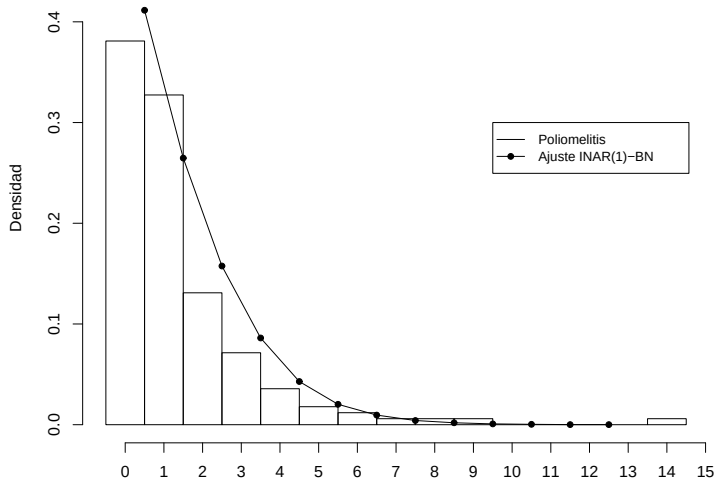

Figura 4.6: Histograma del número de los casos mensuales de poliomielitis de E.U. de 1970 a 1983 y ajuste INAR(1)-BN.

de regresión para datos discretos o datos de conteo), donde se estudian datos discretos sin considerar el tiempo (sin correlación serial). En esta sección estudiamos modelos más explícitos en donde se considera la estructura de correlación que las variables tienen a lo largo de un tiempo determinado (con correlación serial). Para más detalles acerca de esta sección ver Böckenholt (1999a) y Böckenholt (1999b).

4.4.1. Modelos de Regresión INAR(1)-Poisson

Böckenholt (1999a) define una representación de los efectos aleatorios del modelo INAR(1)-Poisson por medio de la descomposición de los parámetros de innovación para la persona i durante el periodo de tiempo t, ι_{it}, en

$$\iota_{it} = \epsilon_i \kappa_{it}$$

donde ϵ_i representa un efecto aleatorio positivo de la persona y $\kappa_{it} = \exp(\mathbf{x}_t\beta_1 + \mathbf{x}_i\beta_2)$ representa los efectos fijos de las covariables específicas de tiempo y persona. Los efectos aleatorios ϵ_i se definen siguiendo una distribución gamma

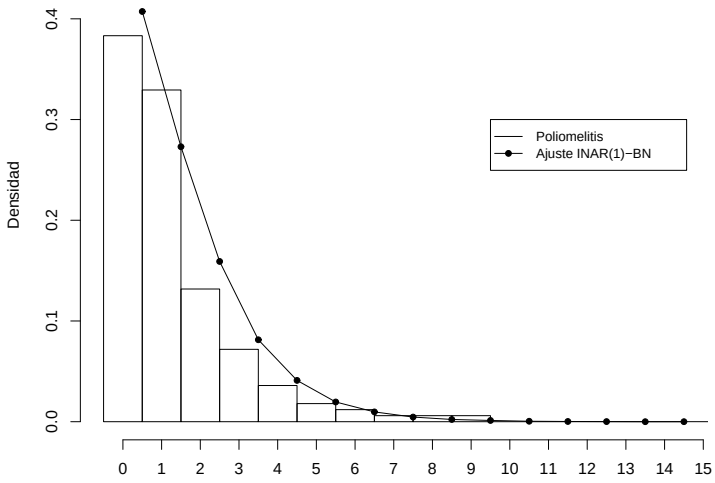

Figura 4.7: Histograma del número de los casos mensuales de poliomielitis de E.U. de 1970 a 1983 y ajuste 2 INAR(1)-BN sin la observación de noviembre de 1972

con valor esperado igual a 1 y parámetro θ_i,

$$f(\epsilon_i) = \frac{\theta_i^{\theta_i}}{\Gamma(\theta_i)} \epsilon_i^{\theta_i - 1} \exp(-\theta_i \epsilon_i).$$

También puede ser de interés expresar las probabilidades de adelgazamiento binomial como una función de las covariables. Se utiliza una función logística para las probabilidades de adelgazamiento

$$\alpha_{it} = \frac{1}{1 + \exp(-(\delta_{0t} + \mathbf{x}_i \delta))},$$

porque ésta mapea el intervalo $(0, 1)$ sobre la línea real.

4.4.2. Modelos de Regresión INAR(1)-Multinomial Negativa

Böckenholt (1999a) define al modelo de regresión INAR(1)-Multinomial Negativo a partir del modelo de regresión INAR(1)-Poisson haciendo que el

parámetro del componente de innovación varíe de acuerdo a una distribución gamma.

También en este caso puede ser de interés expresar las probabilidades de los parámetros de dispersión gamma como una función de las covariables. Como los parámetros individuales son positivos, una función liga apropiada es la exponencial, que relaciona a θ_i con las covariables a través de

$$\theta_i = \exp(\omega_0 + \mathbf{x}_i\omega)^{-1}.$$

En base a estas especificaciones, se la probabilidad marginal está dada por

$$p(y_{i1}, \ldots, y_{iT}) = \int_0^\infty p(y_{i1}) \prod_{t=2}^{T} p(y_{it}|y_{i,t-1}) f(\epsilon_i) d\epsilon_i. \tag{4.7}$$

Una solución analítica de esta integral se presenta en Böckenholt (1999a). De aquí se puede ver que (4.7) se simplifica a la distribución multinomial negativa cuando α_i es igual a 0 para toda i. En este caso, $p(y_{it}|y_{i,t-1}) = p(y_{it})$, y (4.7) se reduce a

$$\begin{aligned} p(y_{i1}, \ldots, y_{iT}) &= \int_0^\infty \prod_{t=1}^{T} p(y_{it}) f(\epsilon_i) d\epsilon_i \\ &= \frac{\theta^{\theta_i}}{\Gamma(\theta_i)} \frac{\Gamma(\theta_i + y_i)}{(\theta_i + \sum_{t=1}^{T} \tau_{it})^{\theta + y_{i+}}} \prod_{t=1}^{T} \frac{\tau_{it}^{y_{it}}}{y_{it}!}, \end{aligned} \tag{4.8}$$

donde τ_{it} es el valor esperado de y_{it}. Como (4.7) generaliza la distribución multinomial negativa, a ésta se le denomina como el modelo INAR(1)-MN. Las distribuciones marginales de (4.7) y (4.8) son idénticas con media y varianza igual a τ_{it} y $(1 + \frac{\tau_{it}}{\theta_i})\tau_{it}$, respectivamente. Las diferencias entre la distribución multinomial negativa con y sin un componente INAR(1) en términos de sus covarianzas y estructuras de autoregresión se discuten a continuación.

Las covarianzas del modelo INAR(1)-MN son sólo no negativas, su tamaño se determina por α_{it}, los respectivos parámetros medios y el parámetro indexado θ_i. A partir de la función generadora de momentos de (4.7), la covarianza de las observaciones adyacentes son (suprimiendo el subíndice i)

$$\gamma_{t-1,t} = \alpha\tau_{t-1} + \tau_{t-1}\tau_t/\theta,$$

y la correlación correspondiente es

$$\rho_{t-1,t} = \frac{\gamma_{t-1,t}}{\sqrt{(1 + \frac{\tau_{t-1}}{\theta})\tau_{t-1}(1 + \frac{\tau_t}{\theta})\tau_t}}.$$

Note que las covarianzas están dadas por una función aditiva de la autocorrelación y las partes de los efectos aleatorios del modelo. Sin embargo, la contribución de la parte de autocorrelación disminuye exponencialmente con el crecimiento de los rezagos entre los periodos de tiempo como puede verse a partir de las covarianzas y las correlaciones para (y_{t-k}, y_t)

$$\gamma_{t-k,t} = \alpha^k \tau_{t-k} + \tau_{t-k}\tau_t/\theta,$$

y

$$\rho_{t-k,t} = \frac{\gamma_{t-k,t}}{\sqrt{(1+\frac{\tau_{t-k}}{\theta})\tau_{t-k}(1+\frac{\tau_t}{\theta})\tau_t}}.$$

Así, cuando $\alpha = 0$, obtenemos las covarianzas de la distribución multinomial negativa con

$$\gamma_{t-k,t} = \tau_{t-k}\tau_t/\theta.$$

La distribución condicional de y_t dada y_{t-1} es una convolución de una distribución binomial negativa y una distribución binomial positiva. Una propiedad importante de la distribución condicional es que su función de regresión es lineal,

$$E(y_t|y_{t-1}) = \frac{\kappa_t}{1+\tau_{t-1}/\theta} + \left(\alpha + \frac{\kappa_t}{\theta+\tau_{t-1}}\right) y_{t-1},$$

y la varianza condicional está dada por

$$V(y_t|y_{t-1}) = \frac{\kappa_t(\kappa_t+\tau_{t-1}+\theta)}{1+\tau_{t-1}/\theta} + \left(\alpha(1-\alpha) + \frac{\kappa_t(\kappa_t+\tau_{t-1}+\theta)}{\theta+\tau_{t-1}}\right) y_{t-1}.$$

Solamente la distribución gamma permite obtener una función de regresión lineal.

Puede ser de interés comparar las funciones de regresión del modelo INAR(1)-MN y las distribuciones multinomiales negativas para rezagos de tamaño k. Bajo la distribución multinomial negativa se obtiene que

$$E(y_t|y_{t-k}) = \frac{\tau_t}{1+\tau_{t-k}/\theta} + \frac{\tau_t}{\theta+\tau_{t-k}} y_{t-k}, \tag{4.9}$$

y bajo el modelo INAR(1)-MN

$$E(y_t|y_{t-k}) = \frac{\tau_t - \alpha^k\tau_{t-k}}{1+\tau_{t-k}/\theta} + \left(\alpha^k + \frac{\tau_t - \alpha^k\tau_{t-k}}{\theta+\tau_{t-k}}\right) y_{t-k}. \tag{4.10}$$

Note que, primero, la pendiente de la función de regresión es más empinada bajo (4.10) que bajo (4.9), pero estas diferencias disminuyen cuando crece el tamaño del rezago; y segundo, el tamaño de θ puede compensar hasta cierto punto las diferencias de las pendientes entre ambos modelos. Cuando existen autocorrelaciones, es probable que el tamaño de θ esté subestimado en (4.9).

4.5. Modelos INAR(p)

Al-Osh y Alzaid (1990) describen los modelos autoregresivos de orden p de valores enteros (INAR(p)) de la siguiente manera. Los procesos INAR(p) son análogos a los procesos AR(p) estándar, utilizados para valores continuos y generalmente con distribución normal, reemplazando la multiplicación escalar por el operador '$\circ$' como se hizo en el proceso INAR(1).

Los modelos INAR(1) definidos en la sección 4.3 son apropiados para modelar datos del tipo de procesos dependientes. Sin embargo, las realizaciones de algunos procesos de conteo $\{Y_t\}$ pueden atribuirse no sólo a su pasado inmediato Y_{t-1} sino también a las realizaciones previas del proceso $\{Y_{t-j}\}_{j=2}^{p}$ para alguna p constante. Consecuentemente para modelar tales procesos es necesario extender los procesos INAR(1) de tal manera que tomen en cuenta estas realizaciones previas. Una forma directa de extender los procesos INAR(1) es considerar la forma estándar de los procesos AR(p) reemplazando las multiplicaciones escalares, $\alpha_i Y_{t-i}$, por la operación $\alpha_i \circ Y_{t-i}$ para $i = 1, \ldots, p$. No obstante, con tales reemplazos, son necesarios algunos supuestos sobre el modelo que rijan la estructura de dependencia del proceso y para que esté bien definido. El proceso $\{Y_t\}$ se dice que es un proceso INAR(p) si admite la siguiente representación

$$Y_t = \sum_{i=1}^{p} \alpha_i \circ Y_{t-i} + I_t \qquad \text{para } t = \ldots, -1, 0, 1, \ldots,$$

donde $\{I_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas de valores enteros no negativos con media μ y varianza σ^2; y α_i $(i = 1, \ldots, p)$ son constantes no negativas tales que $\sum_{i=1}^{p} \alpha_i < 1$. La distribución condicional del vector $(\alpha_1 \circ Y_t,\ \alpha_2 \circ Y_t, \ldots,\ \alpha_p \circ Y_t)$ dada $Y_t = y_t$ es multinomial con parámetros $(\alpha_1, \alpha_2, \ldots, \alpha_p, y_t)$ y es independiente de los resultados pasados del proceso. Es decir, dada $Y_t = y_t$ la variable aleatoria $\alpha_i \circ Y_t$ es independiente de Y_{t-k} y de sus supervivientes $\alpha_j \circ Y_{t-k}$ para $i, j = 1, \ldots, p$ y $k > 0$.

4.6. Modelos Loglineales con la Distribución Multinomial Negativa

Generalmente encontramos diseños de muestreo loglineales modelados por distribuciones Poisson, multinomiales y multinomiales independientes. Estos esquemas de muestreo suponen conteos de celdas independientes, conteos de celdas correlacionados negativamente, y conteos de celdas correlacionados

negativamente dentro de cada uno de los conjuntos mutuamente independientes, respectivamente. Sin embargo, no siempre es adecuado suponer este tipo de distribuciones, ya que en muchos casos los datos presentan una correlación positiva y/o sobredispersión (las varianzas son mayores que las que se esperarían bajo el modelo). Un modelo basado en la distribución multinomial negativa puede dar un mejor ajuste a este tipo de datos, debido a que modela conteos de celdas correlacionados positivamente y las varianzas son mayores a las medias; esta distribución es apropiada para modelar poblaciones en las cuales un conteo observado grande está asociado con conteos grandes en el resto de las celdas. Este tipo de propiedades son muy usadas en eventos que incluyen supervivencia a ciertas enfermedades o conteos de datos longitudinales.

Waller y Zelterman (1997) describen los modelos loglineales utilizando la distribución multinomial negativa.

Para esta sección se intentó desarrollar un ejemplo de los modelos loglineales suponiendo una distribución multinomial negativa; se realizó un programa en WinBUGS para la simulación de los datos de las incidencias de cáncer presentados en Waller y Zelterman (1997) realizando las comparaciones suponiendo una distribución Poisson; sin embargo se presentó una dificultad debido a que WinBUGS no cuenta con la distribución multinomial negativa y se debía dar una representación de ésta a través de distribuciones Poisson-gamma, y a que la gran cantidad de parámetros generó problemas con el ajuste.

Capítulo 5

Conclusiones

Si bien el enfoque Clásico de la Estadística ha predominado por mucho tiempo, la Estadística Bayesiana ha tenido un desarrollo importante en los últimos años. La apertura a estudiar y desarrollar nuevos modelos e implementarlos de tal manera que se pueda hacer uso de la información adicional, conocida o subjetiva, es la gran ventaja de la estadística Bayesiana.

El desarrollo de la tecnología ha abierto la puerta a programas que permiten la aplicación y el análisis eficiente de modelos cada vez más complejos de una manera razonablemente sencilla. WinBUGS ha sido producto de este desarrollo. La implementación de modelos estadísticos a través de métodos de Monte Carlo vía cadenas de Markov es relativamente fácil y bastante eficiente mediante el uso de WinBUGS. Si bien es un *software* que puede ampliarse y perfeccionarse, ha permitido la implementación para hacer pruebas estadísticas, ajustar modelos loglineales y modelos de series de tiempo de manera eficiente, utilizando distribuciones multinomial, Poisson y multinomial negativa.

El análisis de datos categóricos a veces conlleva a construir una tabla de contingencias y analizar la dependencia que puedan tener las variables; si la distribución Poisson y la multinomial son excelentes distribuciones para modelar estos datos, no siempre describen adecuadamente las propiedades que puedan tener los datos. El uso de la distribución multinomial negativa a veces genera un mejor ajuste para éstos; propiedades como varianza mayor a la media o que las correlaciones sean positivas son características de una distribución multinomial negativa. A pesar de que es difícil de implementar, debido a que la mayoría de los *software* no consideran explícitamente esta distribución, pueden obtenerse representaciones mediante el uso de la distribución Poisson-gamma, o específicamente, mediante el uso de la distribución Poisson cuyo parámetro siga una distribución gamma.

El análisis de datos categóricos está bien establecido en la literatura,

especialmente desde el enfoque clásico, pero en años recientes ha habido avances considerables en el enfoque Bayesiano. Sin embargo, mucho de este trabajo se ha concentrado en modelos estáticos y particularmente en el uso de modelos loglineales.

Mucho más reciente es el estudio de modelos dinámicos o de series de tiempo para datos categóricos. La mayor parte de este trabajo se ha realizado desde el enfoque clásico, y queda mucho por hacer desde la perspectiva Bayesiana.

Sin duda existe una gran variedad de modelos que se pueden aplicar al estudiar datos categóricos. Los presentados en este trabajo son algunos de los que han sido más utilizados en la literatura.

Esperamos que esta tesis contribuya a generar un mayor interés en el tema.

Apéndice A

Distribución Dirichlet

Distribución Dirichlet

Sea $(Y_1, \ldots, Y_m)$ un vector aleatorio que tiene distribución Dirichlet con vector de parámetros $(\alpha_1, \ldots, \alpha_m, \alpha_{m+1})$, con $\alpha_i > 0$ para $i = 1, \ldots, m, m+1$, cuya función de densidad de probabilidad conjunta es

$$p(y_1, \ldots, y_m | \alpha_1, \ldots, \alpha_m, \alpha_{m+1}) = \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{m} y_i\right)^{\alpha_{m+1}-1} \prod_{i=1}^{m} y_i^{\alpha_i - 1},$$

donde

$$0 < y_i < 1,\ i = 1, \ldots, m,\ \sum_{i=1}^{m} y_i \leq 1.$$

La distribución Dirichlet se puede ver como una generalización de la distribución beta.

Distribuciones Marginales

Si se integra la función de densidad conjunta de $(Y_1, \ldots, Y_m)$ para una de las variables, digamos Y_m, entonces la función de densidad conjunta de

$(Y_1, \ldots, Y_{m-1})$ es

$$\begin{aligned}
p \quad & (y_1, \ldots, y_{m-1}) \\
= \quad & \int_0^{1-\sum_{i=1}^{m-1} y_i} p(y_1, \ldots, y_m) dy_m \\
= \quad & \int_0^{1-\sum_{i=1}^{m-1} y_i} \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{m} y_i\right)^{\alpha_{m+1}-1} \prod_{i=1}^{m} y_i^{\alpha_i - 1} dy_m \\
= \quad & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \prod_{i=1}^{m-1} y_i^{\alpha_i - 1} \\
& \times \int_0^{1-\sum_{i=1}^{m-1} y_i} \left(1 - \sum_{i=1}^{m-1} y_i - y_m\right)^{\alpha_{m+1}-1} y_m^{\alpha_m - 1} dy_m,
\end{aligned}$$

para resolver esta integral defina $v = y_m/(1 - \sum_{i=1}^{m-1} y_i)$, de tal manera que $y_m = (1 - \sum_{i=1}^{m-1} y_i)v$ y $dy_m = (1 - \sum_{i=1}^{m-1} y_i)dv$, por lo tanto

$$\begin{aligned}
p \quad & (y_1, \ldots, y_{m-1}) \\
= \quad & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \left\{ \prod_{i=1}^{m-1} y_i^{\alpha_i - 1} \right\} \left(1 - \sum_{i=1}^{m-1} y_i\right)^{\alpha_{m+1}-1+\alpha_m-1} \\
& \times \int_0^1 (1-v)^{\alpha_{m+1}-1} v^{\alpha_m - 1} \left(1 - \sum_{i=1}^{m-1} y_i\right) dv \\
= \quad & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \left\{ \prod_{i=1}^{m-1} y_i^{\alpha_i - 1} \right\} \left(1 - \sum_{i=1}^{m-1} y_i\right)^{\alpha_m + \alpha_{m+1} - 1} \\
& \times \frac{\Gamma(\alpha_m)\Gamma(\alpha_{m+1})}{\Gamma(\alpha_m + \alpha_{m+1})} \\
= \quad & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=m}^{m+1} \alpha_i) \prod_{i=1}^{m-1} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{m-1} y_i\right)^{\sum_{i=m}^{m+1} \alpha_i - 1} \prod_{i=1}^{m-1} y_i^{\alpha_i - 1},
\end{aligned}$$

es decir la distribución conjunta de $(Y_1, \ldots, Y_{m-1})$ es Dirichlet con vector de parámetros $(\alpha_1, \ldots, \alpha_{m-1}, \sum_{i=m}^{m+1} \alpha_i)$.

Si se integra la función de densidad conjunta de $(Y_1, \ldots, Y_{m-1})$ para una de las variables, digamos Y_{m-1}, entonces la función de densidad conjunta de

$(Y_1, \ldots, Y_{m-2})$ es

$$
\begin{aligned}
p \quad & (y_1, \ldots, y_{m-2}) \\
= \quad & \int_0^{1-\sum_{i=1}^{m-2} y_i} p(y_1, \ldots, y_{m-1}) dy_{m-1} \\
= \quad & \int_0^{1-\sum_{i=1}^{m-2} y_i} \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=m}^{m+1} \alpha_i) \prod_{i=1}^{m-1} \Gamma(\alpha_i)} \\
& \times \left(1 - \sum_{i=1}^{m-1} y_i\right)^{\sum_{i=m}^{m+1} \alpha_i - 1} \prod_{i=1}^{m-1} y_i^{\alpha_i - 1} dy_{m-1} \\
= \quad & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=m}^{m+1} \alpha_i) \prod_{i=1}^{m-1} \Gamma(\alpha_i)} \prod_{i=1}^{m-2} y_i^{\alpha_i - 1} \\
& \times \int_0^{1-\sum_{i=1}^{m-2} y_i} \left(1 - \sum_{i=1}^{m-2} y_i - y_{m-1}\right)^{\sum_{i=m}^{m+1} \alpha_i - 1} y_{m-1}^{\alpha_{m-1} - 1} dy_{m-1},
\end{aligned}
$$

para resolver esta integral defina $v = y_{m-1}/(1 - \sum_{i=1}^{m-2} y_i)$, de tal manera que $y_{m-1} = (1 - \sum_{i=1}^{m-2} y_i)v$ y $dy_{m-1} = (1 - \sum_{i=1}^{m-2} y_i)dv$, por lo tanto

$$
\begin{aligned}
p \quad & (y_1, \ldots, y_{m-2}) \\
= \quad & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=m}^{m+1} \alpha_i) \prod_{i=1}^{m-1} \Gamma(\alpha_i)} \prod_{i=1}^{m-2} y_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{m-2} y_i\right)^{\sum_{i=m-1}^{m+1} \alpha_i - 2} \\
& \times \int_0^1 (1 - v)^{\sum_{i=m}^{m+1} \alpha_i - 1} v^{\alpha_{m-1} - 1} \left(1 - \sum_{i=1}^{m-2} y_i\right) dv \\
= \quad & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=m}^{m+1} \alpha_i) \prod_{i=1}^{m-1} \Gamma(\alpha_i)} \left\{\prod_{i=1}^{m-2} y_i^{\alpha_i - 1}\right\} \left(1 - \sum_{i=1}^{m-2} y_i\right)^{\sum_{i=m-1}^{m+1} \alpha_i - 1} \\
& \times \frac{\Gamma(\alpha_{m-1}) \Gamma(\sum_{i=m}^{m+1} \alpha_i)}{\Gamma(\sum_{i=m-1}^{m+1} \alpha_i)} \\
= \quad & \frac{\Gamma(\sum_{i=0}^{m} \alpha_i)}{\Gamma(\sum_{i=m-1}^{m+1} \alpha_i) \Gamma \prod_{i=1}^{m-2} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{m-2} y_i\right)^{\sum_{i=m-1}^{m+1} \alpha_i - 1} \prod_{i=1}^{m-2} y_i^{\alpha_i - 1},
\end{aligned}
$$

es decir la distribución conjunta de $(Y_1, \ldots, Y_{m-2})$ es Dirichlet con vector de parámetros $(\alpha_1, \ldots, \alpha_{m-2}, \sum_{i=m-1}^{m+1} \alpha_i)$.

Si continuamos este procedimiento vamos a tener que las distribuciones conjuntas continúan teniendo comportamientos análogos.

Finalmente, la distribución marginal de una de las variables, digamos Y_1, es

$$\begin{aligned} p & (y_1) \\ = & \int_0^{1-y_2} p(y_1, y_2) dy_2 \\ = & \int_0^{1-y_1} \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=3}^{m+1} \alpha_i) \prod_{i=1}^{2} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{2} y_i\right)^{\sum_{i=3}^{m+1} \alpha_i - 1} \prod_{i=1}^{2} y_i^{\alpha_i - 1} dy_2 \\ = & \frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=3}^{m+1} \alpha_i) \prod_{i=1}^{2} \Gamma(\alpha_i)} y_1^{\alpha_1 - 1} \int_0^{1-y_1} (1 - y_1 - y_2)^{\sum_{i=3}^{m+1} \alpha_i - 1} y_2^{\alpha_2 - 1} dy_2, \end{aligned}$$

sea $v = y_2/(1 - y_1)$, entonces $y_2 = (1 - y_1)v$ y $dy_2 = (1 - y_1)dv$

$$\begin{aligned} p(y_1) &= \frac{\Gamma\left(\sum_{i=1}^{m+1} \alpha_i\right)}{\Gamma\left(\sum_{i=3}^{m+1} \alpha_i\right) \prod_{i=1}^{2} \Gamma\left(\alpha_i\right)} y_1^{\alpha_1 - 1} (1 - y_1)^{\sum_{i=2}^{m+1} \alpha_i - 2} \\ &\quad \times \int_0^1 (1 - v)^{\sum_{i=3}^{m+1} \alpha_i - 1} v^{\alpha_2 - 1} (1 - y_1) dv \\ &= \frac{\Gamma\left(\sum_{i=1}^{m+1} \alpha_i\right)}{\Gamma\left(\sum_{i=3}^{m+1} \alpha_i\right) \prod_{i=1}^{2} \Gamma\left(\alpha_i\right)} y_1^{\alpha_1 - 1} (1 - y_1)^{\sum_{i=2}^{m+1} \alpha_i - 1} \\ &\quad \times \frac{\Gamma(\alpha_2) \Gamma\left(\sum_{i=3}^{m+1} \alpha_i\right)}{\Gamma(\sum_{i=2}^{m+1} \alpha_i)} \\ &= \frac{\Gamma\left(\sum_{i=1}^{m+1} \alpha_i\right)}{\Gamma\left(\sum_{i=2}^{m+1} \alpha_i\right) \Gamma(\alpha_1)} y_1^{\alpha_1 - 1} (1 - y_1)^{\sum_{i=2}^{m+1} \alpha_i - 1} \\ &= \frac{\Gamma\left(\sum_{i=1}^{m+1} \alpha_i\right)}{\Gamma(\alpha_1) \Gamma\left(\sum_{i=2}^{m+1} \alpha_i\right)} y_1^{\alpha_1 - 1} (1 - y_1)^{\sum_{i=2}^{m+1} \alpha_i - 1} \end{aligned}$$

es decir la distribución marginal de Y_1 es una distribución beta con parámetros α_1 y $\sum_{i=2}^{m+1} \alpha_i$.

En resumen,

$$\begin{aligned} (Y_1, \ldots, Y_k) &\sim Dirichlet\left(\alpha_1, \ldots, \alpha_k, \sum_{i=k+1}^{m+1} \alpha_i\right), \\ Y_k &\sim Beta\left(\alpha_k, \sum_{i \neq k, i=1}^{m+1} \alpha_i\right). \end{aligned}$$

Las medias, las varianzas y las covarianzas son

$$E[Y_k] = \frac{\alpha_k}{\alpha_+}, \quad Var[Y_k] = \frac{\alpha_k(\alpha_+ - \alpha_k)}{(\alpha_+)^2(\alpha_+ + 1)}, \quad Cov[Y_i, Y_j] = -\frac{\alpha_i \alpha_j}{(\alpha_+)^2(\alpha_+ + 1)}$$

donde $\alpha_+ = \sum_{i=1}^{m+1} \alpha_i$.

Propiedades

Sean $Y_1, \ldots, Y_m$ variables aleatorias, se cumple que

$$p(y_1, \ldots, y_m) = p(y_m|y_1, \ldots, y_{m-1})p(y_{m-1}|y_1, \ldots, y_{m-2}) \cdots p(y_2|y_1)p(y_1).$$

Si $Y_1, \ldots, Y_m$ tienen una distribución Dirichlet con vector de parámetros $(\alpha_1, \ldots, \alpha_T, \alpha_{m+1})$, entonces podríamos encontrar cada una de las probabilidades condicionales y de esta manera definir a la probabilidad conjunta de una distribución Dirichlet en términos del producto de probabilidades marginales condicionadas:

$$\begin{aligned}
p & \quad (y_m|y_1, \ldots, y_{m-1}) \\
= & \quad \frac{p(y_1, \ldots, y_m)}{p(y_1, \ldots, y_{m-1})} \\
= & \quad \frac{\frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{m} y_i\right)^{\alpha_{m+1}-1} \prod_{i=1}^{m} y_i^{\alpha_i - 1}}{\frac{\Gamma(\sum_{i=1}^{m+1} \alpha_i)}{\Gamma(\sum_{i=m}^{m+1} \alpha_i) \prod_{i=1}^{m-1} \Gamma(\alpha_i)} \left(1 - \sum_{i=1}^{m-1} y_i\right)^{\sum_{i=m}^{m+1} \alpha_i - 1} \prod_{i=1}^{m-1} y_i^{\alpha_i - 1}} \\
= & \quad \frac{\Gamma(\alpha_m + \alpha_{m+1})}{\Gamma(\alpha_m)\Gamma(\alpha_{m+1})} \frac{\left(1 - \sum_{i=1}^{m} y_i\right)^{\alpha_{m+1}-1}}{\left(1 - \sum_{i=1}^{m-1} y_i\right)^{\sum_{i=m}^{m+1} \alpha_i - 1}} y_m^{\alpha_m - 1} \\
= & \quad \frac{\Gamma(\alpha_m + \alpha_{m+1})}{\Gamma(\alpha_m)\Gamma(\alpha_{m+1})} \frac{\left(1 - \sum_{i=1}^{m-1} y_i - y_m\right)^{\alpha_{m+1}-1}}{\left(1 - \sum_{i=1}^{m-1} y_i\right)^{\alpha_{m+1}-1}} \\
& \quad \times \frac{y_m^{\alpha_m - 1}}{\left(1 - \sum_{i=1}^{m-1} y_i\right)^{\alpha_m - 1}} \frac{1}{1 - \sum_{i=1}^{m-1} y_i} \\
= & \quad \frac{\Gamma(\alpha_m + \alpha_{m+1})}{\Gamma(\alpha_m)\Gamma(\alpha_{m+1})} \left(1 - \frac{y_m}{1 - \sum_{i=1}^{m-1} y_i}\right)^{\alpha_{m+1}-1} \\
& \quad \times \left(\frac{y_m}{1 - \sum_{i=1}^{m-1} y_i}\right)^{\alpha_m - 1} \left(\frac{1}{1 - \sum_{i=1}^{m-1} y_i}\right).
\end{aligned}$$

Considere el cambio de variable

$$y'_m = \frac{y_m}{1 - \sum_{i=1}^{m-1} y_i},$$

de esta manera

$$\frac{dy'_m}{dy_m} = \frac{1}{1-\sum_{i=1}^{m-1} y_i},$$

por lo tanto

$$p(y'_m|y_1,\ldots,y_{m-1}) = \frac{\Gamma(\alpha_m+\alpha_{m+1})}{\Gamma(\alpha_m)\Gamma(\alpha_{m+1})}\left(1-y'_m\right)^{\alpha_{m+1}-1}\left(y'_m\right)^{\alpha_m-1},$$

es decir, la variable aleatoria

$$Y'_m = \frac{Y_m}{1-\sum_{i=1}^{m-1} Y_i}$$

dada $Y_1,\ldots,Y_{m-1}$ tiene una distribución beta con parámetros α_m y α_{m+1}.

Siguiendo con la secuencia de probabilidades condicionales, tenemos

$$\begin{aligned}
p & \quad (y_{m-1}|y_1,\ldots,y_{m-2}) \\
= & \quad \frac{p(y_1,\ldots,y_{m-1})}{p(y_1,\ldots,y_{m-2})} \\
= & \quad \frac{\frac{\Gamma(\sum_{i=1}^{m+1}\alpha_i)}{\Gamma(\sum_{i=m}^{m+1}\alpha_i)\prod_{i=1}^{m-1}\Gamma(\alpha_i)}\left(1-\sum_{i=1}^{m-1}y_i\right)^{\sum_{i=m}^{m+1}\alpha_i-1}\prod_{i=1}^{m-1}y_i^{\alpha_i-1}}{\frac{\Gamma(\sum_{i=1}^{m+1}\alpha_i)}{\Gamma(\sum_{i=m-1}^{m+1}\alpha_i)\prod_{i=1}^{m-2}\Gamma(\alpha_i)}\left(1-\sum_{i=1}^{m-2}y_i\right)^{\sum_{i=m-1}^{m+1}\alpha_i-1}\prod_{i=1}^{m-2}y_i^{\alpha_i-1}} \\
= & \quad \frac{\Gamma(\sum_{i=m-1}^{m+1}\alpha_i)}{\Gamma(\alpha_{m-1})\Gamma(\sum_{i=m}^{m+1}\alpha_i)}\frac{\left(1-\sum_{i=1}^{m-1}y_i\right)^{\sum_{i=m}^{m+1}\alpha_i-1}}{\left(1-\sum_{i=1}^{m-2}y_i\right)^{\sum_{i=m-1}^{m+1}\alpha_i-1}}y_{m-1}^{\alpha_{m-1}-1} \\
= & \quad \frac{\Gamma(\sum_{i=m-1}^{m+1}\alpha_i)}{\Gamma(\alpha_{m-1})\Gamma(\sum_{i=m}^{m+1}\alpha_i)}\frac{\left(1-\sum_{i=1}^{m-2}y_i-y_{m-1}\right)^{\sum_{i=m}^{m+1}\alpha_i-1}}{\left(1-\sum_{i=1}^{m-2}y_i\right)^{\sum_{i=m}^{m+1}\alpha_i-1}} \\
& \quad \times\frac{y_{m-1}^{\alpha_{m-1}-1}}{\left(1-\sum_{i=1}^{m-2}y_i\right)^{\alpha_{m-1}-1}}\frac{1}{1-\sum_{i=1}^{m-2}y_i} \\
= & \quad \frac{\Gamma(\sum_{i=m-1}^{m+1}\alpha_i)}{\Gamma(\alpha_{m-1})\Gamma(\sum_{i=m}^{m+1}\alpha_i)}\left(1-\frac{y_{m-1}}{1-\sum_{i=1}^{m-2}y_i}\right)^{\sum_{i=m}^{m+1}\alpha_i-1} \\
& \quad \times\left(\frac{y_{m-1}}{1-\sum_{i=1}^{m-2}y_i}\right)^{\alpha_{m-1}-1}\frac{1}{1-\sum_{i=1}^{m-2}y_i}.
\end{aligned}$$

Considere el cambio de variable

$$y'_{m-1} = \frac{y_{m-1}}{1-\sum_{i=1}^{m-2} y_i},$$

de esta manera

$$\frac{dy'_{m-1}}{dy_{m-1}} = \frac{1}{1-\sum_{i=1}^{m-2} y_i},$$

por lo tanto

$$\begin{aligned} p \quad & (y'_{m-1}|y_1,\ldots,y_{m-2}) \\ = \quad & \frac{\Gamma(\sum_{i=m-1}^{m+1}\alpha_i)}{\Gamma(\alpha_{m-1})\Gamma(\sum_{i=m}^{m+1}\alpha_i)}\left(1-y'_{m-1}\right)^{\sum_{i=m}^{m+1}\alpha_i-1}\left(y'_{m-1}\right)^{\alpha_{m-1}-1}, \end{aligned}$$

es decir, la variable aleatoria

$$Y'_{m-1} = \frac{Y_{m-1}}{1-\sum_{i=1}^{m-2} Y_i}$$

dada $Y_1,\ldots,Y_{m-2}$ tiene una distribución beta con parámetros α_{m-1} y $\sum_{i=m}^{m+1}\alpha_i$.

Siguiendo con la secuencia de probabilidades condicionales, tenemos

$$\begin{aligned} p \quad & (y_2|y_1) \\ = \quad & \frac{p(y_1,y_2)}{p(y_1)} \\ = \quad & \frac{\frac{\Gamma(\sum_{i=1}^{m+1}\alpha_i)}{\Gamma(\sum_{i=3}^{m+1}\alpha_i)\prod_{i=1}^{2}\Gamma(\alpha_i)}(1-y_1-y_2)^{\sum_{i=3}^{m+1}\alpha_i-1}y_1^{\alpha_1-1}y_2^{\alpha_2-1}}{\frac{\Gamma(\sum_{i=1}^{m+1}\alpha_i)}{\Gamma(\alpha_1)\Gamma(\sum_{i=2}^{m+1}\alpha_i)}(1-y_1)^{\sum_{i=2}^{m+1}\alpha_i-1}y_1^{\alpha_1-1}} \\ = \quad & \frac{\Gamma(\sum_{i=2}^{m+1}\alpha_i)}{\Gamma(\alpha_2)\Gamma(\sum_{i=3}^{m+1}\alpha_i)}\frac{(1-y_1-y_2)^{\sum_{i=3}^{m+1}\alpha_i-1}}{(1-y_1)^{\sum_{i=2}^{m+1}\alpha_i-1}}y_2^{\alpha_2-1} \\ = \quad & \frac{\Gamma(\sum_{i=2}^{m+1}\alpha_i)}{\Gamma(\alpha_2)\Gamma(\sum_{i=3}^{m+1}\alpha_i)}\frac{(1-y_1-y_2)^{\sum_{i=3}^{m+1}\alpha_i-1}}{(1-y_1)^{\sum_{i=3}^{m+1}\alpha_i-1}}\frac{y_2^{\alpha_2-1}}{(1-y_1)^{\alpha_2-1}}\frac{1}{1-y_1} \\ = \quad & \frac{\Gamma(\sum_{i=2}^{m+1}\alpha_i)}{\Gamma(\alpha_2)\Gamma(\sum_{i=3}^{m+1}\alpha_i)}\left(1-\frac{y_2}{1-y_1}\right)^{\sum_{i=3}^{m+1}\alpha_i-1}\left(\frac{y_2}{1-y_1}\right)^{\alpha_2-1}\frac{1}{1-y_1}, \end{aligned}$$

considere el cambio de variable

$$y'_2 = \frac{y_2}{1-y_1},$$

de esta manera

$$\frac{dy'_2}{dy_2} = \frac{1}{1-y_1},$$

por lo tanto

$$p(y_2'|y_1) = \frac{\Gamma(\sum_{i=2}^{m+1}\alpha_i)}{\Gamma(\alpha_2)\Gamma(\sum_{i=3}^{m+1}\alpha_i)}\,(1-y_2')^{\sum_{i=3}^{m+1}\alpha_i-1}\,(y_2')^{\alpha_2-1}\,,$$

es decir, la variable aleatoria

$$Y_2' = \frac{Y_2}{1-Y_1}$$

dada Y_1 tiene una distribución beta con parámetros α_2 y $\sum_{i=3}^{m+1}\alpha_i$.

Finalmente la distribución de Y_1 es beta con parámetros α_1 y $\sum_{i=2}^{m+1}\alpha_i$.

Por lo tanto, la variable

$$Y_k' = \frac{Y_k}{1-\sum_{i=1}^{k-1}Y_i}$$

dada $Y_1,\ldots,Y_{k-1}$ tiene una distribución beta con parámetros α_k y $\sum_{i=k+1}^{m+1}\alpha_i$.

Asociación con la Distribución Beta

La distribución condicional conjunta de

$$Y_j' = \frac{Y_j}{1-\sum_{i=1}^{k}Y_i} \qquad j=k+1,\ldots,m,$$

dada $Y_1,\ldots,Y_k$, es una distribución Dirichlet con parámetros $\alpha_{k+1},\ldots,\alpha_m$, α_{m+1}.

$$\begin{aligned} p(y_{k+1}',\ldots,y_m'|y_1,\ldots,y_k) &= p(y_{k+1},\ldots,y_m|y_1,\ldots,y_k)J \\ &= \frac{p(y_1,\ldots,y_k,y_{k+1},\ldots,y_m)}{p(y_1,\ldots,y_k)}J \end{aligned}$$

donde, en este caso

$$y_j' = \frac{y_j}{1-\sum_{i=1}^{k}y_i} \qquad \text{y} \qquad y_j = y_j'\left(1-\sum_{i=1}^{k}y_i\right)$$

para $j = k+1, \ldots, m$, entonces

$$
\begin{aligned}
& p(y'_{k+1}, \ldots, y'_m | y_1, \ldots, y_k) \\
&= \frac{\frac{\Gamma(\sum_{i=1}^{m+1}\alpha_i)}{\prod_{i=1}^{m+1}\Gamma(\alpha_i)}\left(1-\sum_{i=1}^{m} y_i\right)^{\alpha_{m+1}-1}\prod_{i=1}^{m} y_i^{\alpha_i-1}}{\frac{\Gamma(\sum_{i=1}^{m+1}\alpha_i)}{\Gamma(\sum_{i=k+1}^{m+1}\alpha_i)\prod_{i=1}^{k}\Gamma(\alpha_i)}\left(1-\sum_{i=1}^{k} y_i\right)^{\sum_{i=k+1}^{m+1}\alpha_i-1}\prod_{i=1}^{k} y_i^{\alpha_i-1}} J \\
&= \frac{\Gamma(\sum_{i=k+1}^{m+1}\alpha_i)\left(1-\sum_{i=1}^{m} y_i\right)^{\alpha_{m+1}-1}}{\prod_{i=k+1}^{m+1}\Gamma(\alpha_i)\left(1-\sum_{i=1}^{k} y_i\right)^{\sum_{i=k+1}^{m+1}\alpha_i-1}} \prod_{i=k+1}^{m} y_i^{\alpha_i-1} J \\
&= \frac{\Gamma(\sum_{i=k+1}^{m+1}\alpha_i)\left(1-\sum_{i=1}^{k} y_i-\sum_{i=k+1}^{m} y_i\right)^{\alpha_{m+1}-1}}{\prod_{i=k+1}^{m+1}\Gamma(\alpha_i)\left(1-\sum_{i=1}^{k} y_i\right)^{\sum_{i=k+1}^{m+1}\alpha_i-1}} \prod_{i=k+1}^{m} y_i^{\alpha_i-1} J,
\end{aligned}
$$

haciendo el cambio de variable $y_j = y'_j\left(1-\sum_{i=1}^{k} y_i\right)$, para $j = k+1, \ldots, m$, tenemos que $dy_j/dy'_j = \left(1-\sum_{i=1}^{k} y_i\right)$, para $j = k+1, \ldots, m$, por tanto el Jacobiano es $J = \left(1-\sum_{i=1}^{k} y_i\right)^{m-k}$,

$$
\begin{aligned}
& p(y'_{k+1}, \ldots, y'_m | y_1, \ldots, y_k) \\
&= \frac{\Gamma(\sum_{i=k+1}^{m+1}\alpha_i)}{\prod_{i=k+1}^{m+1}\Gamma(\alpha_i)} \frac{\left(1-\sum_{i=1}^{k} y_i-\sum_{i=k+1}^{m} y'_j\left(1-\sum_{i=1}^{k} y_i\right)\right)^{\alpha_{m+1}-1}}{\left(1-\sum_{i=1}^{k} y_i\right)^{\sum_{i=k+1}^{m+1}\alpha_i-1}} \\
&\times \left[\prod_{i=k+1}^{m}\left\{y'_j\left(1-\sum_{i=1}^{k} y_i\right)\right\}^{\alpha_i-1}\right]\left(1-\sum_{i=1}^{k} y_i\right)^{m-k} \\
&= \frac{\Gamma(\sum_{i=k+1}^{m+1}\alpha_i)}{\prod_{i=k+1}^{m+1}\Gamma(\alpha_i)}\left(1-\sum_{i=k+1}^{m} y'_j\right)^{\alpha_{m+1}-1}\prod_{i=k+1}^{m} y_j'^{\alpha_i-1},
\end{aligned}
$$

ésta es una distribución Dirichlet con parámetros $\alpha_{k+1}, \ldots, \alpha_m, \alpha_{m+1}$.

Asociación con la Distribución Beta (Variables Independientes)

Las variables aleatorias

$$Z_j = \frac{Y_j}{\sum_{i=j}^{m} Y_i} \qquad j = 1, \ldots, m$$

son variables aleatorias mutuamente independientes con distribución beta, con parámetros α_j y $\sum_{i=j+1}^{m}\alpha_i$.

Asociación con la Distribución Gamma

Sean $X_1, \ldots, X_{m+1}$ variables aleatorias independientes con distribución gamma y parámetros α_i y 1, $X_i \sim Gamma(\alpha_i, 1)$ para $i = 1, \ldots, m, m+1$. Si

$$Y_i = \frac{X_i}{\sum_{i=1}^{m+1} X_i} \qquad i = 1, \ldots, m,$$

entonces

$$(Y_1, \ldots, Y_m) \sim Dirichlet(\alpha_1, \ldots, \alpha_m, \alpha_{m+1}).$$

La función de densidad conjunta de $X_1, \ldots, X_m, X_{m+1}$ es

$$\begin{aligned} p(x_1, \ldots, x_{m+1}) &= \prod_{i=1}^{m+1} \frac{1}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} \exp\{-x_i\} \\ &= \frac{1}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \left\{ \prod_{i=1}^{m+1} x_i^{\alpha_i - 1} \right\} \exp\left\{ -\sum_{i=1}^{m+1} x_i \right\}. \end{aligned}$$

Haciendo las transformaciones $y_{m+1} = \sum_{i=1}^{m+1} x_i$ y $y_i = x_i / (\sum_{i=1}^{m+1} x_i)$ para $i = 1, \ldots, m$, obtenemos que $x_i = y_i y_{m+1}$ para $i = 1, \ldots, m$ y $x_{m+1} = y_{m+1}(1 - \sum_{i=1}^{m} y_i)$ tenemos que

$$\begin{aligned} p(y_1, \ldots, y_{m+1}) &= \frac{1}{\prod_{i=1}^{m+1} \Gamma(\alpha_i)} \left\{ y_{m+1} \left(1 - \sum_{i=1}^{m} y_i \right) \right\}^{\alpha_{m+1} - 1} \\ &\quad \times \left\{ \prod_{i=1}^{m} (y_{m+1} y_i)^{\alpha_i - 1} \right\} \exp\{-y_{m+1}\} J, \end{aligned}$$

con J el Jacobiano

$$\begin{aligned} J &= \frac{\partial(x_1, \ldots, x_m, x_{m+1})}{\partial(y_1, \ldots, y_m, y_{m+1})} \\ &= \begin{vmatrix} y_{m+1} & 0 & \ldots & 0 & y_1 \\ 0 & y_{m+1} & \ldots & 0 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & y_{m+1} & y_m \\ -y_{m+1} & -y_{m+1} & \ldots & -y_{m+1} & 1 - \sum_{i=1}^{m} y_i \end{vmatrix} \\ &= y_{m+1}^m, \end{aligned}$$

por tanto,

$$\begin{aligned} p(y_1,\ldots,y_{m+1}) &= \frac{1}{\prod_{i=1}^{m+1}\Gamma(\alpha_i)}\left\{y_{m+1}\left(1-\sum_{i=1}^{m}y_i\right)\right\}^{\alpha_{m+1}-1} \\ &\quad\times\left\{\prod_{i=1}^{m}(y_{m+1}y_i)^{\alpha_i-1}\right\}\exp\left\{-y_{m+1}\right\}y_{m+1}^{m} \\ &= \frac{1}{\prod_{i=1}^{m+1}\Gamma(\alpha_i)}y_{m+1}^{\sum_{i=1}^{m+1}\alpha_i-1}\left(1-\sum_{i=1}^{m}y_i\right)^{\alpha_{m+1}-1} \\ &\quad\times\left\{\prod_{i=1}^{m}y_i^{\alpha_i-1}\right\}\exp\left\{-y_{m+1}\right\}, \end{aligned}$$

integrando bajo y_{m+1},

$$\begin{aligned} p(y_1,\ldots,y_m) &= \int_0^\infty p(y_1,\ldots,y_{m+1})dy_{m+1} \\ &= \frac{1}{\prod_{i=1}^{m+1}\Gamma(\alpha_i)}\left(1-\sum_{i=1}^{m}y_i\right)^{\alpha_{m+1}-1}\left\{\prod_{i=1}^{m}y_i^{\alpha_i-1}\right\} \\ &\quad\times\int_0^\infty y_{m+1}^{\sum_{i=1}^{m+1}\alpha_i-1}\exp\left\{-y_{m+1}\right\}dy_{m+1} \\ &= \frac{1}{\prod_{i=1}^{m+1}\Gamma(\alpha_i)}\left(1-\sum_{i=1}^{m}y_i\right)^{\alpha_{m+1}-1}\left\{\prod_{i=1}^{m}y_i^{\alpha_i-1}\right\} \\ &\quad\times\Gamma(\sum_{i=1}^{m+1}\alpha_i) \\ &= \frac{\Gamma(\sum_{i=1}^{m+1}\alpha_i)}{\prod_{i=1}^{m+1}\Gamma(\alpha_i)}\left(1-\sum_{i=1}^{m}y_i\right)^{\alpha_{m+1}-1}\left\{\prod_{i=1}^{m}y_i^{\alpha_i-1}\right\}, \end{aligned}$$

ésta es una distribución Dirichlet con parámetros $\alpha_1,\ldots,\alpha_m,\alpha_{m+1}$.

Distribuciones Dirichlet y sus Mezclas

Una densidad de probabilidad Dirichlet con t dimensiones es de la forma

$$\frac{\Gamma(\sum_i k_i)}{\prod_i\Gamma(k_i)}\prod_{i=1}^{t}q_i^{k_i-1}\qquad(k_i>0;\ i=1,\ldots,t;\ \sum q_i=1),$$

donde q_i corresponde a la probabilidad de la celda i. Las k_i son hiperparámetros (parámetros en la distribución inicial).

Denotamos a tal distribución Dirichlet por $D(t; k_1, \ldots, k_t)$ y, si $k_1 = \ldots = k_t = k$, por $D(t, k)$ la cual se conoce como *distribución Dirichlet simétrica* con hiperparámetro constante k.

En un principio en investigaciones acerca de variables multinomiales se usaba una mezcla de distribuciones Dirichlet $D(t, k)$

$$\int_0^\infty D(t, k)\phi(k)dk,$$

donde ϕ es la densidad log-Cauchy

$$\phi(k) = \frac{1}{k[\pi^2 + (\log k)^2]}.$$

La generalización de esta mezcla es

$$\int_0^\infty D(t, t'k)\phi(k)dk,$$

denotamos esta mezcla de densidades iniciales Dirichlet simétricas por $D^*(t, t')$ (ver Good, 1976).

La distribución inicial log-Cauchy fue seleccionada porque es "no informativa"y porque se aproxima a la densidad impropia de Jeffreys-Haldane $1/k$ que propuso para "representar ignorancia"del valor de una variable positiva.

En este trabajo suponemos que la distribución inicial es $D^*(t, 1)$, con t que toma valores r, s y rs, y q_i los valores $p_{i.}$, $p_{.j}$ y p_{ij}.

Necesitamos la siguiente fórmula

$$p\left((m_i)|D(t, k)\right) = \frac{\Gamma(tk)N!\prod\Gamma(m_i + k)}{\Gamma(k)^t\Gamma(N + tk)\prod m_i!} \quad \left(\sum m_i = N\right).$$

(Esta fórmula se reduce a 1 cuando $N = 0$ y a $1/t$ cuando $N = 1$.) Consecuentemente

$$p((m_i)|D^*(t, t')) = \Phi((m_i), t, t'),$$

donde

$$\begin{aligned}\Phi((m_i), t, t') &= \frac{N!}{\prod m_i!}\int_0^\infty \frac{\Gamma(tt'k)\prod\Gamma(m_i + t'k)}{\Gamma(t'k)^t\Gamma(N + tt'k)}\phi(k)dk \\ &= \frac{N!}{\prod m_i!}\int_0^\infty \frac{\Gamma(tk)\prod\Gamma(m_i + k)}{\Gamma(k)^t\Gamma(N + tk)}\phi\left(\frac{k}{t'}\right)\frac{dk}{t'}.\end{aligned}$$

Como $\phi(k)$ es proporcional a $1/k$, cuando k no es muy grande, podemos esperar que $\Phi((m_i), t, t')$ no será muy diferente de $\Phi((m_i), t, 1)$.

Apéndice B

Distribución Multinomial Negativa

B.1. Distribución Binomial Negativa

Una variable aleatoria Y tiene una distribución binomial negativa con índice r y parámetro π ($r = 1, 2, \ldots, 0 < \pi < 1$), denotada por $Nb(y|\pi, r)$, si su función de masa de probabilidad es

$$p(y|\pi, r) = \binom{r+y-1}{r-1}\pi^r(1-\pi)^y, \qquad y = 0, 1, 2, \ldots.$$

La interpretación usual de esta distribución está dada por el número de fracasos antes del r-ésimo éxito.

Propiedades

La media y varianza son

$$E[y] = r\frac{1-\pi}{\pi} \qquad \text{y} \qquad Var[y] = r\frac{1-\pi}{\pi^2}.$$

Si $r(1-\pi) > 1$, la moda es el menor entero mayor o igual que $r(1-\pi)/\pi$; si $r(1-\pi) = 1$, hay dos modas, en 0 y 1; si $r(1-\pi) < 1$, la moda es 0.

Si $r = 1$, Y tiene una distribución geométrica o Pascal. Además, la suma de k variables aleatorias independientes cuya distribución es binomial negativa con índice r_i, $i = 1, \ldots, k$, y parámetro π, respectivamente, es una variable aleatoria que se distribuye binomial negativa con índice $r = \sum_{i=1}^{k} r_i$ y parámetro π.

La función característica es

$$\begin{aligned} E[e^{ity}] &= \sum_{y=0}^{\infty} e^{ity}\binom{r+y-1}{r-1}\pi^r(1-\pi)^y \\ &= \sum_{y=0}^{\infty}\binom{r+y-1}{r-1}\pi^r\left[(1-\pi)e^{it}\right]^y \\ &= \frac{\pi^r}{(1-(1-\pi)e^{it})^r}\sum_{y=0}^{\infty}\binom{r+y-1}{r-1}\left(1-(1-\pi)e^{it}\right)^r\left[(1-\pi)e^{it}\right]^y \\ &= \left(\frac{\pi}{1-(1-\pi)e^{it}}\right)^r. \end{aligned}$$

Análisis Conjugado

La distribución inicial conjugada para una distribución binomial negativa es la distribución beta con parámetros α y β ($\alpha > 0$ y $\beta > 0$), $\pi \sim Beta(\pi|\alpha,\beta)$, tal que la función de densidad de probabilidad de π es

$$p(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, \qquad 0<\pi<1,$$

donde $\Gamma(\cdot)$ es la función gamma dada por

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt.$$

Note que una distribución inicial simétrica sobre π se obtiene haciendo $\alpha = \beta$ y cuando $\alpha = \beta = 1$ se reduce a una distribución inicial uniforme.

La distribución final de π es una distribución beta con parámetros $\alpha + r$ y $\beta + y$, es decir,

$$\begin{aligned} p(\pi|y,r) &= \frac{p(y|\pi,r)p(\pi)}{\int_0^1 p(y|\pi,r)p(\pi)d\pi} \\ &= \frac{\binom{r+y-1}{r-1}\pi^r(1-\pi)^y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{\int_0^1\binom{r+y-1}{r-1}\pi^r(1-\pi)^y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi} \\ &= \frac{\pi^{\alpha+r-1}(1-\pi)^{\beta+y-1}}{\int_0^1\pi^{\alpha+r-1}(1-\pi)^{\beta+y-1}d\pi} \\ &= \frac{\Gamma(\alpha+r+\beta+y)}{\Gamma(\alpha+r)\Gamma(\beta+y)}\pi^{\alpha+r-1}(1-\pi)^{\beta+y-1}, \qquad 0<\pi<1, \end{aligned}$$

y se denota por

$$\pi \sim Beta(\pi|\alpha + r, \beta + y).$$

Suponga que $Y_1, \ldots, Y_T$ son variables aleatorias independientes con distribución $Nb(y_i|\pi, r_i)$ para $i = 1, \ldots, T$. Entonces, si la distribución inicial de π es $Beta(\pi|\alpha, \beta)$, la distribución final de π es

$$\begin{aligned}
p(\pi|y_1, \ldots, y_T) &= \frac{p(y_1, \ldots, y_T|\pi)p(\pi)}{\int_0^1 p(y_1, \ldots, y_T|\pi)p(\pi)d\pi} \\
&= \frac{\left\{\prod_i \binom{r_i+y_i-1}{r_i-1}\pi^{r_i}(1-\pi)^{y_i}\right\}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}}{\int_0^1 \left\{\prod_i \binom{r_i+y_i-1}{r_i-1}\pi^{r_i}(1-\pi)^{y_i}\right\}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi} \\
&= \frac{\pi^{\alpha+\sum_i r_i-1}(1-\pi)^{\beta+\sum_i y_i-1}}{\int_0^1 \pi^{\alpha+\sum_i r_i-1}(1-\pi)^{\beta+\sum_i y_i-1}d\pi} \\
&= \frac{\Gamma\left(\alpha+\sum_i r_i+\beta+\sum_i y_i\right)}{\Gamma\left(\alpha+\sum_i r_i\right)\Gamma\left(\beta+\sum_i y_i\right)}\pi^{\alpha+\sum_i r_i-1}(1-\pi)^{\beta+\sum_i y_i-1},
\end{aligned}$$

cuya distribución es beta con parámetros $\alpha + \sum_{i=1}^T r_i$ y $\beta + \sum_{i=1}^T y_i$, es decir,

$$\pi \sim Beta\left(\pi|\alpha + \sum_{i=1}^T r_i, \beta + \sum_{i=1}^T y_i\right).$$

Predicción

La distribución predictiva inicial de $Y = y$ es

$$\begin{aligned}
p(y) &= \int_0^1 p(y|\pi)p(\pi)d\pi \\
&= \int_0^1 \binom{r+y-1}{r-1}\pi^r(1-\pi)^y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi \\
&= \binom{r+y-1}{r-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \pi^{\alpha+r-1}(1-\pi)^{\beta+y-1}d\pi \\
&= \binom{r+y-1}{r-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+r)\Gamma(\beta+y)}{\Gamma(\alpha+r+\beta+y)},
\end{aligned}$$

cuya distribución es beta-binomial-negativa de parámetros α, β y r, y se denota por $Nbb(y|\alpha, \beta, r)$.

La distribución predictiva inicial de $Y_1 = y_1, \ldots, Y_T = y_T$ es

$$\begin{aligned}
p & (y_1, \ldots, y_T) \\
= & \int_0^1 p(y_1, \ldots, y_T|\pi)p(\pi)d\pi \\
= & \int_0^1 \left\{ \prod_i \binom{r_i + y_i - 1}{r_i - 1} \pi^{r_i}(1-\pi)^{y_i} \right\} \\
& \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi \\
= & \left\{ \prod_i \binom{r_i + y_i - 1}{r_i - 1} \right\} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
& \times \int_0^1 \pi^{\alpha+\sum_i r_i - 1}(1-\pi)^{\beta+\sum_i y_i - 1}d\pi \\
= & \left\{ \prod_i \binom{r_i + y_i - 1}{r_i - 1} \right\} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma\left(\alpha + \sum_i r_i\right)\Gamma\left(\beta + \sum_i y_i\right)}{\Gamma\left(\alpha + \sum_i r_i + \beta + \sum_i y_i\right)}.
\end{aligned}$$

Sea $Y_1, \ldots, Y_T$ variables aleataorias independientes cada una con distribución $Nb(y_i|\pi, r_i)$ para $i = 1, \ldots, T$, entonces $\sum_{i=1}^T Y_i$ tiene una distribución $Nb(\sum_{i=1}^T y_i|\pi, \sum_{i=1}^T r_i)$. La distribución predictiva inicial de $Y_+ = \sum_{i=1}^T Y_i = y_+$ es

$$\begin{aligned}
p(y_+) & = \int_0^1 p(y_+|\pi)p(\pi)d\pi \\
& = \int_0^1 \binom{r_+ + y_+ - 1}{r_+ - 1} \pi^{r_+}(1-\pi)^{y_+} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi \\
& = \binom{r_+ + y_+ - 1}{r_+ - 1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{\alpha+r_+-1}(1-\pi)^{\beta+y_+-1}d\pi \\
& = \binom{r_+ + y_+ - 1}{r_+ - 1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+r_+)\Gamma(\beta+y_+)}{\Gamma(\alpha+r_++\beta+y_+)},
\end{aligned}$$

cuya distribución es beta-binomial-negativa de parámetros α, β y r_+, denotada por $Nbb(y|\alpha, \beta, r_+)$, con $y_+ = \sum_{i=1}^T y_i$ y $r_+ = \sum_{i=1}^T r_i$.

La distribución predictiva final de $Y = y$ es

$$\begin{aligned} p(y|y_1,\ldots,y_T) &= \int_0^1 p(y|\pi)p(\pi|y_1,\ldots,y_T)d\pi \\ &= \int_0^1 \binom{r+y-1}{r-1}\pi^r(1-\pi)^y \frac{\Gamma(\alpha+\sum_i r_i+\beta+\sum_i y_i)}{\Gamma(\alpha+\sum_i r_i)\Gamma(\beta+\sum_i y_i)} \\ &\quad \times\pi^{\alpha+\sum_i r_i-1}(1-\pi)^{\beta+\sum_i y_i-1}d\pi \\ &= \binom{r+y-1}{r-1}\frac{\Gamma(\alpha+\sum_i r_i+\beta+\sum_i y_i)}{\Gamma(\alpha+\sum_i r_i)\Gamma(\beta+\sum_i y_i)} \\ &\quad \times\int_0^1 \pi^{\alpha+r+\sum_i r_i-1}(1-\pi)^{\beta+y+\sum_i y_i-1}d\pi \\ &= \binom{r+y-1}{r-1}\frac{\Gamma(\alpha+\sum_i r_i+\beta+\sum_i y_i)}{\Gamma(\alpha+\sum_i r_i)\Gamma(\beta+\sum_i y_i)} \\ &\quad \times\frac{\Gamma(\alpha+r+\sum_i r_i)\Gamma(\beta+y+\sum_i y_i)}{\Gamma(\alpha+r+\sum_i r_i+\beta+y+\sum_i y_i)}, \end{aligned}$$

cuya distribución es beta-binomial-negativa de parámetros $\alpha+\sum_{i=1}^T r_i$, $\beta+\sum_{i=1}^T y_i$ y r, es decir,

$$y \sim Nbb\left(y|\alpha+\sum_{i=1}^T r_i, \beta+\sum_{i=1}^T y_i, r\right).$$

B.2. Distribución Poisson-Gamma

Una variable aleatoria Y tiene distribución Poisson-gamma con parámetros α, β y λ ($\alpha>0$, $\beta>0$, $\lambda=1,2,\ldots$), $Pg(y|\alpha,\beta,\lambda)$, si su función de masa de probabilidad es

$$p(y|\alpha,\beta,\lambda) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\lambda}{\beta+\lambda}\right)^y\left(\frac{\beta}{\beta+\lambda}\right)^\alpha, \qquad y=0,1,2,\ldots.$$

La distribución se genera con una mezcla de una distribución Poisson con una distribución gamma, tal que

$$\begin{aligned} Y &\sim Poisson(y|\lambda\gamma) \\ \gamma &\sim Gamma(\gamma|\alpha,\beta). \end{aligned}$$

$$\begin{aligned}
p(y|\alpha,\beta,\lambda) &= \int_0^\infty p(y|\lambda\gamma)p(\gamma|\alpha,\beta)d\gamma \\
&= \int_0^\infty \frac{(\lambda\gamma)^y}{y!}\exp\{-\lambda\gamma\}\frac{\beta^\alpha}{\Gamma(\alpha)}\gamma^{\alpha-1}\exp\{-\beta\gamma\}d\gamma \\
&= \frac{(\lambda)^y}{y!}\frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \gamma^{\alpha+y-1}\exp\{-\gamma(\beta+\lambda)\}d\gamma \\
&= \frac{(\lambda)^y}{y!}\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+y)}{(\beta+\lambda)^{\alpha+y}} \\
&= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\lambda}{\beta+\lambda}\right)^y\left(\frac{\beta}{\beta+\lambda}\right)^\alpha.
\end{aligned}$$

La media y la varianza de esta variable aleatoria son

$$E[y] = \lambda\frac{\alpha}{\beta} \qquad \text{y} \qquad Var[y] = \lambda\frac{\alpha(\beta+\lambda)}{\beta^2}.$$

La distribución Poisson-gamma es una generalización de la distribución binomial negativa, $Nb(y|\alpha,\beta/(\beta+\lambda))$, para valores enteros de α.

Esta representación puede usarse para simular y hacer inferencias Bayesianas sobre una distribución binomial negativa.

Si $\alpha\lambda > \beta+\lambda$, la moda es el menor entero mayor o igual a $\lambda(\alpha-1)/\beta-1$; si $\alpha\lambda = \beta+\lambda$, hay dos modas, en 0 y en 1; si $\alpha\lambda < \beta+\lambda$, la moda es 0.

La función característica es

$$\begin{aligned}
&E\ [e^{ity}] \\
&= \sum_{i=0}^\infty e^{ity}\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\lambda}{\beta+\lambda}\right)^y\left(\frac{\beta}{\beta+\lambda}\right)^\alpha \\
&= \sum_{i=0}^\infty \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\lambda e^{it}}{\beta+\lambda}\right)^y\left(\frac{\beta}{\beta+\lambda}\right)^\alpha \\
&= \left(\frac{\beta}{\beta+\lambda}\right)^\alpha\left(1-\frac{\lambda e^{it}}{\beta+\lambda}\right)^{-\alpha}\sum_{i=0}^\infty \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\lambda e^{it}}{\beta+\lambda}\right)^y\left(1-\frac{\lambda e^{it}}{\beta+\lambda}\right)^\alpha \\
&= \left(\frac{\beta}{\beta+\lambda-\lambda e^{it}}\right)^\alpha.
\end{aligned}$$

B.3. Distribución Multinomial Negativa

Sea $\mathbf{Y} = (Y_1,\ldots,Y_T)$ un vector aleatorio con distribución multinomial negativa con índice $\alpha > 0$ y con vector de medias $\lambda = (\lambda_1,\ldots,\lambda_T)$. El

soporte de $(Y_1, \ldots, Y_T)$ es $(y_1, \ldots, y_T)$ tales que y_t, con $t = 1, \ldots, T$, es un entero no negativo. La función de probabilidad de masa de $(Y_1, \ldots, Y_T)$ es

$$\begin{aligned} p(\mathbf{y}|\lambda, \alpha) &= \frac{\Gamma\left(\alpha + \sum_{t=1}^{T} y_t\right)}{\Gamma(\alpha) \prod_{t=1}^{T} y_t!} \left(\frac{\alpha}{\alpha + \sum_{t=1}^{T} \lambda_t}\right)^{\alpha} \prod_{t=1}^{T} \left(\frac{\lambda_t}{\alpha + \sum_{t=1}^{T} \lambda_t}\right)^{y_t} \\ &= \frac{\Gamma\left(\alpha + \sum_{t=1}^{T} y_t\right)}{\Gamma(\alpha) \prod_{t=1}^{T} y_t!} \left(1 - \sum_{t=1}^{T} \pi_t\right)^{\alpha} \prod_{t=1}^{T} \pi_t^{y_t} \\ &= p(\mathbf{y}|\pi, \alpha) \end{aligned}$$

donde

$$\pi = (\pi_1, \ldots, \pi_T), \qquad \pi_t = \frac{\lambda_t}{\alpha + \sum_{t=1}^{T} \lambda_t}, \qquad 1 - \sum_{t=1}^{T} \pi_t = \frac{\alpha}{\alpha + \sum_{t=1}^{T} \lambda_t}.$$

Dos características importantes de esta distribución son que las correlaciones entre los conteos y_i son positivas y las varianzas son mayores que las medias. Las medias, las varianzas y las correlaciones son

$$E[y_t] = \lambda_t, \qquad Var[y_t] = \frac{\lambda(\alpha + \lambda_t)}{\alpha}, \qquad Cov(y_i, y_j) = \frac{\lambda_i \lambda_j}{\alpha} \; \forall i \neq j.$$

La interpretación usual del índice α está dada por el número de eventos en la (T+1)-ésima celda.

Análisis Conjugado

La distribución inicial conjugada para una distribución multinomial negativa es la distribución Dirichlet con vector de parámetros $\beta = (\beta_1, \ldots, \beta_{T+1})$, $Dirichlet(\pi|\beta)$, con función de densidad de probabilidad dada por

$$p(\pi|\beta) = \frac{\Gamma\left(\sum_{t=1}^{T+1} \beta_t\right)}{\prod_{t=1}^{T+1} \Gamma(\beta_t)} \left(1 - \sum_{t=1}^{T} \pi_t\right)^{\beta_{T+1}-1} \prod_{t=1}^{T} \pi_t^{\beta_t - 1},$$

donde $\beta_t > 0$ para $t = 1, \ldots, T+1$, y $0 < \pi_t < 1$ para $t = 1, \ldots, T$.

La función de distribución final de $\pi = (\pi_1, \ldots, \pi_T)$ dada $\mathbf{Y} = \mathbf{y}$ es

$$\begin{aligned}
p & (\pi|\mathbf{y}) \\
= & \frac{p(\mathbf{y}|\pi)p(\pi)}{\int_0^1 p(\mathbf{y}|\pi)p(\pi)d\pi} \\
= & \frac{\frac{\Gamma(\alpha+\sum y_t)}{\Gamma(\alpha)\prod y_t!}(1-\sum \pi_t)^{\alpha}\{\prod \pi_t^{y_t}\}\frac{\Gamma(\sum \beta_t)}{\prod\Gamma(\beta_t)}(1-\sum \pi_t)^{\beta_{T+1}-1}\{\prod \pi_t^{\beta_t-1}\}}{\int_0^1 \frac{\Gamma(\alpha+\sum y_t)}{\Gamma(\alpha)\prod y_t!}(1-\sum \pi_t)^{\alpha}\{\prod \pi_t^{y_t}\}\frac{\Gamma(\sum \beta_t)}{\prod\Gamma(\beta_t)}(1-\sum \pi_t)^{\beta_{T+1}-1}\{\prod \pi_t^{\beta_t-1}\}d\pi} \\
= & \frac{\left\{\prod_{t=1}^T \pi_t^{\beta_t+y_t-1}\right\}\left(1-\sum_{t=1}^T \pi_t\right)^{\beta_{T+1}+\alpha-1}}{\int_0^1 \left\{\prod_{t=1}^T \pi_t^{\beta_t+y_t-1}\right\}\left(1-\sum_{t=1}^T \pi_t\right)^{\beta_{T+1}+\alpha-1} d\pi} \\
= & \frac{\Gamma\left(\sum_{t=1}^{T+1}\beta_t+\alpha+\sum_{t=1}^T y_t\right)}{\Gamma\left(\beta_{T+1}+\alpha\right)\prod_{t=1}^T\Gamma\left(\beta_t+y_t\right)}\left(1-\sum_{t=1}^T \pi_t\right)^{\beta_{T+1}+\alpha-1}\prod_{t=1}^T \pi_t^{\beta_t+y_t-1}.
\end{aligned}$$

Esta distribución es una distribución Dirichlet con vector de parámetros $(\beta_1 + y_1, \ldots, \beta_T + y_T, \beta_{T+1} + \alpha)$, es decir,

$$\pi \sim Dirichlet\left((\pi_1, \ldots, \pi_T)|(\beta_1 + y_1, \ldots, \beta_T + y_T, \beta_{T+1} + \alpha)\right)$$

Predicción

La distribución predictiva inicial de $\mathbf{Y} = \mathbf{y}$ es

$$\begin{aligned}
p(\mathbf{y}) &= \int_0^1 p(y_1, \ldots, y_T|\pi)p(\pi)d\pi \\
&= \int_0^1 \frac{\Gamma\left(\alpha+\sum_{t=1}^T y_t\right)}{\Gamma(\alpha)\prod_{t=1}^T y_t!}\left(1-\sum_{t=1}^T \pi_t\right)^{\alpha}\left\{\prod_{t=1}^T \pi_t^{y_t}\right\} \\
&\quad \times\frac{\Gamma(\sum_{t=1}^{T+1}\beta_t)}{\prod_{t=1}^{T+1}\Gamma(\beta_t)}\left(1-\sum_{t=1}^T \pi_t\right)^{\beta_{T+1}-1}\left\{\prod_{t=1}^T \pi_t^{\beta_t-1}\right\}d\pi \\
&= \frac{\Gamma\left(\alpha+\sum_{t=1}^T y_t\right)}{\Gamma(\alpha)\prod_{t=1}^T y_t!}\frac{\Gamma(\sum_{t=1}^{T+1}\beta_t)}{\prod_{t=1}^{T+1}\Gamma(\beta_t)} \\
&\quad \times\int_0^1 \left(1-\sum_{t=1}^T \pi_t\right)^{\beta_{T+1}+\alpha-1}\left\{\prod_{t=1}^T \pi_t^{\beta_t+y_t-1}\right\}d\pi \\
&= \frac{\Gamma\left(\alpha+\sum_{t=1}^T y_t\right)}{\Gamma(\alpha)\prod_{t=1}^T y_t!}\frac{\Gamma(\sum_{t=1}^{T+1}\beta_t)}{\prod_{t=1}^{T+1}\Gamma(\beta_t)}\frac{(\beta_{T+1}+\alpha)\prod_{t=1}^T\Gamma\left(\beta_t+y_t\right)}{\Gamma\left(\sum_{t=1}^{T+1}\beta_t+\sum_{t=1}^T y_t+\alpha\right)}
\end{aligned}$$

La distribución predictiva final de $\mathbf{Y} = \mathbf{y}$ dada la información inicial $\mathbf{Y_0} = (Y_{01}, \ldots, Y_{0T}) = (y_{01}, \ldots, y_{0T}) = \mathbf{y_0}$ es

$$\begin{aligned}
p(\mathbf{y}|\mathbf{y_0}) &= \int_0^1 p(\mathbf{y}|\pi)p(\pi|\mathbf{y_0})d\pi \\
&= \int_0^1 \frac{\Gamma\left(\alpha + \sum_{t=1}^T y_t\right)}{\Gamma(\alpha)\prod_{t=1}^T y_t!}\left(1 - \sum_{t=1}^T \pi_t\right)^{\alpha}\prod_{t=1}^T \pi_t^{y_t} \\
&\quad \times \frac{\Gamma\left(\sum_{t=1}^{T+1}\beta_t + \alpha_0 + \sum_{t=1}^T y_{0t}\right)}{\Gamma\left(\beta_{T+1} + \alpha_0\right)\prod_{t=1}^T \Gamma\left(\beta_t + y_{0t}\right)}\left(1 - \sum_{t=1}^T \pi_t\right)^{\beta_{T+1}+\alpha_0-1} \\
&\quad \times \left\{\prod_{t=1}^T \pi_t^{\beta_t + y_{0t} - 1}\right\} d\pi \\
&= \frac{\Gamma\left(\alpha + \sum_{t=1}^T y_t\right)}{\Gamma(\alpha)\prod_{t=1}^T y_t!}\frac{\Gamma\left(\sum_{t=1}^{T+1}\beta_t + \alpha_0 + \sum_{t=1}^T y_{0t}\right)}{\Gamma\left(\beta_{T+1} + \alpha_0\right)\prod_{t=1}^T \Gamma\left(\beta_t + y_{0t}\right)} \\
&\quad \times \int_0^1 \left(1 - \sum_{t=1}^T \pi_t\right)^{\beta_{T+1}+\alpha+\alpha_0-1}\prod_{t=1}^T \pi_t^{\beta_t + y_t + y_{0t} - 1} d\pi \\
&= \frac{\Gamma\left(\alpha + \sum_{t=1}^T y_t\right)}{\Gamma(\alpha)\prod_{t=1}^T y_t!}\frac{\Gamma\left(\sum_{t=1}^{T+1}\beta_t + \alpha_0 + \sum_{t=1}^T y_{0t}\right)}{\Gamma\left(\beta_{T+1} + \alpha_0\right)\prod_{t=1}^T \Gamma\left(\beta_t + y_{0t}\right)} \\
&\quad \times \frac{\left(\beta_{T+1} + \alpha + \alpha_0\right)\prod_{t=1}^T \Gamma\left(\beta_t + y_t + y_{0t}\right)}{\Gamma\left(\sum_{t=1}^{T+1}\beta_t + \sum_{t=1}^T (y_t + y_{0t}) + \alpha + \alpha_0\right)}.
\end{aligned}$$

Relación con la Distribución Poisson

Sea Y una variable aletoria con distribuión binomial negativa con índice r y media μ (parámetro $\pi = r/(r+\mu)$, con $r = 1, 2, \ldots$, $\mu > 0$ y $0 < \pi < 1$), con función de masa de probabilidad

$$\begin{aligned}
p(y) &= \binom{r+y-1}{r-1}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^y \\
&= \binom{r+y-1}{r-1}\pi^r(1-\pi)^y.
\end{aligned}$$

Si $r \to \infty$ entonces Y tiene una distribución Poisson con media μ, con función de densidad de probabilidad

$$p(y) = e^{-\mu}\mu^y\frac{1}{y!}.$$

Es decir,

$$\begin{aligned} p(y) &= \binom{r+y-1}{r-1}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^y \\ &= \left(\prod_{j=0}^{y-1}(r+j)\right)\frac{1}{y!}\left(\frac{r}{r+\mu}\right)^r\left(\frac{1}{r+\mu}\right)^y\mu^y \\ &= \left(\prod_{j=0}^{y-1}\frac{r+j}{r+\mu}\right)\left(\frac{r}{r+\mu}\right)^r\left(\frac{1}{r+\mu}\right)^y\mu^y\frac{1}{y!} \\ &= \left(\prod_{j=0}^{y-1}\frac{1+\frac{j}{r}}{1+\frac{\mu}{r}}\right)\left(\frac{1}{1+\frac{\mu}{r}}\right)^r\left(\frac{1}{r+\mu}\right)^y\mu^y\frac{1}{y!} \\ &\to e^{-\mu}\mu^y\frac{1}{y!} \qquad \text{cuando} \quad r\to\infty, \end{aligned}$$

esto es porque

$$\lim_{r\to\infty}\left(\prod_{j=0}^{y-1}\frac{1+\frac{j}{r}}{1+\frac{\mu}{r}}\right)=1, \qquad \lim_{r\to\infty}\left(1+\frac{x}{r}\right)^r=e^x \qquad \text{y} \qquad \lim_{r\to\infty}\left(\frac{1}{r+\mu}\right)^y=1.$$

Por lo tanto la distribución Poisson es un caso especial de la distribución binomial negativa.

B.4. Distribución Poisson-Gamma Multivariada

Sea $\mathbf{Y}=(Y_1,\ldots,Y_T)$ un vector aleatorio con distribución Poisson-gamma multivariada con índice $\alpha>0$ y con vector de medias $\lambda=(\lambda_1,\ldots,\lambda_T)$. El soporte de $(Y_1,\ldots,Y_T)$ es $(y_1,\ldots,y_T)$ tales que y_t es un entero no negativo, con $t=1,\ldots,T$. La distribución Poisson-gamma está generada por una mezcla de distribuciones Poisson independientes con una distribución gamma, es decir,

$$\begin{aligned} Y_t &\sim Poisson(y_t|\lambda_t\gamma) \qquad \text{con } t=1,\ldots,T \text{ independientes,} \\ \gamma &\sim Gamma(\gamma|\alpha,\beta). \end{aligned}$$

Su función de masa de probabilidad es

$$\begin{aligned}
p & (y_1, \ldots, y_T) \\
= & \int_0^\infty p(y_1, \ldots, y_T|\gamma) p(\gamma) d\gamma \\
= & \int_0^\infty \left(\prod_{t=1}^T \frac{(\gamma\lambda_t)^{y_t}}{y_t!} \exp\{-\gamma\lambda_t\} \right) \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \gamma^{\alpha-1} \exp\{-\beta\gamma\} \right) d\gamma \\
= & \frac{\prod_{t=1}^T \lambda_t^{y_t}}{\prod_{t=1}^T y_t!} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \gamma^{\sum_{t=1}^T y_t + \alpha - 1} \exp\left\{ -\gamma \left(\sum_{t=1}^T \lambda_t + \beta \right) \right\} d\gamma \\
= & \frac{\prod_{t=1}^T \lambda_t^{y_t}}{\prod_{t=1}^T y_t!} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma\left(\sum_{t=1}^T y_t + \alpha \right)}{\left(\sum_{t=1}^T \lambda_t + \beta \right)^{\sum_{t=1}^T y_t + \alpha}} \\
= & \frac{\Gamma\left(\sum_{t=1}^T y_t + \alpha \right)}{\Gamma(\alpha) \prod_{t=1}^T y_t!} \left(\frac{\beta}{\sum_{t=1}^T \lambda_t + \beta} \right)^\alpha \prod_{t=1}^T \left(\frac{\lambda_t}{\sum_{t=1}^T \lambda_t + \beta} \right)^{y_t}.
\end{aligned}$$

Dos características importantes de esta distribución son que las correlaciones entre los conteos y_i son positivas y las varianzas son mayores que las medias. Las medias, las varianzas y las correlaciones son

$$E[y_t] = \lambda_t \frac{\alpha}{\beta}, \qquad Var[y_t] = \lambda_t \alpha \frac{(\beta + \lambda_t)}{\beta^2}, \qquad Cov(y_i, y_j) = \lambda_i \lambda_j \frac{\alpha}{\beta^2} \; \forall i \neq j.$$

Si $\alpha = \beta$, la distribución Poisson-gamma multivariada se transforma en una distribución multinomial negativa. Esta representación puede usarse para simular y hacer inferencias Bayesianas sobre la distribución multinomial negativa.

Apéndice C

Programas

Los siguientes programas fueron realizados por el Dr. Peter Congdon y se encuentran disponibles en la página `http://alpha.qmul.ac.uk/~ugfa117/` o en la página `http://www.geog.qmul.ac.uk/staff/congdon.html`. Estos programas se realizaron en el *software* WinBUGS, el cual es un software libre y puede obtenerse de la página `http://www.mrc-bsu.cam.ac.uk/bugs/`. WinBUGS es un *software* especializado en la implementación de métodos de Monte Carlo basados en cadenas de Markov; realizado para el análisis de una gran variedad de modelos abordados desde el punto de vista Bayesiano.

C.1. Programa de Movilidad Social

Modelo

```
model {# INICIALES
# Factores para los orígenes
u1[1] <- 0; for (i in 2:I) { u1[i] ~ dnorm(0,0.01) }
# Factores para los destinos
u2[1] <- 0; for (i in 2:I) { u2[i] ~ dnorm(0,0.01) }
# Factores para las diagonales
v1[1] <- 0; for (i in 2:I) { v1[i] ~ dnorm(0,0.01) }
u ~ dnorm(0,0.01);
v ~ dnorm(0,0.01);
# MOVILIDAD PERFECTA
# for (i in 1:I) { for (j in 1:I) {
# m[i,j] ~ dpois(mu[i,j]); log(mu[i,j]) <- u+u1[i]+u2[j]; }}
# MOVILIDAD CUASI-PERFECTA
for(i in 2:I) { for(j in 1:i-1) {
m[i,j] ~ dpois(mu[i,j]); log(mu[i,j]) <- u+u1[i]+u2[j]; }}
for(i in 1:I-1) { for(j in i+1:I) {
```

```
m[i,j] ~ dpois(mu[i,j]); log(mu[i,j]) <- u+u1[i]+u2[j]; }}
for(i in 1:I) { m[i,i] ~ dpois(mu[i,i]); log(mu[i,i]) <- v+v1[i]; }
# AJUSTE
for(i in 1:I) { for(j in 1:I) {
devG[i,j]<-m[i,j]*log((m[i,j]+0.5)/(mu[i,j]+0.5))-(m[i,j]-mu[i,j]);
devX[i,j] <- (m[i,j]-mu[i,j])*(m[i,j]-mu[i,j])/mu[i,j]; }}
G2 <- 2 * sum( devG[,] );
X2 <- sum( devX[,] )
}
```

Datos

```
list( m = structure(.Data = c(50, 45, 8, 18, 8, 28, 174, 84, 154, 55,
11, 78, 110, 23, 96, 14, 150, 185, 714, 447, 0, 42, 72, 320, 411),
.Dim = c(5,5)), I = 5)
```

Iniciales

```
list(u1=c(NA,1,1,1,1), u2=c(NA,1,1,1,1), v1=c(NA,1,1,1,1), u=1, v=1)
```

C.2. Programa *Inter Sib Marriage*

Modelo

```
model { # INICIALES
# Factores para la esposa
u1[1] <- 0; for (i in 2:I) { u1[i] ~ dnorm(0,0.01) } # Iniciales
# Factores para el esposo
u2[1] <- 0; for (j in 2:J) { u2[j] ~ dnorm(0,0.01) } # Iniciales
# Media
u ~ dnorm(0,0.01);
# VEROSIMILITUD
# MODELO DE INDEPENDENCIA
for (i in 1:I) { for (j in 1:J) { m[i,j] ~ dpois(mu[i,j]);
# mu con estructura de ceros donde delta=0
mu.r[i,j] <- delta[i,j]*mu[i,j];
# Modelo loglineal
log(mu[i,j]) <- delta[i,j]*(u + u1[i]+u2[j]);}}
# BONDAD DE AJUSTE
for (i in 1:I) { for (j in 1:I) {
devG[i,j]<-m[i,j]*log((m[i,j]+0.5)/(mu[i,j]+0.5))-(m[i,j]-mu[i,j]);
devX[i,j] <- (m[i,j]-mu[i,j])*(m[i,j]-mu[i,j])/mu[i,j]; } }
mu.r.sum <- sum(mu.r[,]);
G2 <- 2 * sum( devG[,] );
X2 <- sum( devX[,] )
```

```
}
```

Datos

```
list(m=structure(.Data=c(NA, 5, 17, NA, 6, 5, NA, 0, 16, 2, NA, 2,
NA, 10, 11, 10, NA, NA, NA, 9, 6, 20, 8, 0, 1), .Dim=c(5,5)),
delta=structure(.Data=c(0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1,
1, 0, 0, 0, 1, 1, 1, 1, 1, 1), .Dim=c(5,5)), I = 5, J = 5)
```

Iniciales

```
list(u1=c(NA,1,1,1,1), u2=c(NA,1,1,1,1), u=1)
```

C.3. Programa de Movilidad Social (continuación)

Modelo

```
CUASI-SIMETRIA
model { # INICIALES
# Factores para orígenes/destinos:
u1[1] <- 0; for (i in 2:I) { u1[i] ~ dnorm(0,0.001) }
u2[1] <- 0; for (i in 2:I) { u2[i] ~ dnorm(0,0.001) }
# primer renglón
for (j in 1:I) { u12[1,j] <- 0 } # contraste primer estrato, esquina
# primer columna
for (i in 2:I) { u12[i,1] <- 0 }
# interacciones de la diagonal superior
for (i in 2:I-1) { for (j in i+1:I) { u12[i,j] ~ dnorm(0,0.001); }}
# interacciones de la diagonal inferior
for (i in 3:I) { for (j in 2:i-1) { u12[i,j] <- u12[j,i]; }}
# diagonal principal
for (i in 2:I) { u12[i,i] ~ dnorm(0,0.001); }
u ~ dnorm(0,0.001);
# VEROSIMILITUD
for (i in 2:I) { for (j in 1:i-1) {
m[i,j] ~ dpois(mu[i,j]);
log(mu[i,j]) <- u+u1[i]+u2[j]+u12[i,j] }}
for (i in 1:I-1) { for (j in i+1:I) {
m[i,j] ~ dpois(mu[i,j]);
log(mu[i,j]) <- u+ u1[i]+u2[j]+u12[i,j] }}
for (i in 1:I) { m[i,i] ~ dpois(mu[i,i]);
log(mu[i,i]) <- u + u1[i]+ u2[i]+u12[i,i];}
for (i in 1:I) { for (j in 1:I) {
devG[i,j]<-m[i,j]*log((m[i,j]+0.5)/(mu[i,j]+0.5))-(m[i,j]-mu[i,j]);
```

```
devX[i,j] <- (m[i,j]-mu[i,j])*(m[i,j]-mu[i,j])/mu[i,j]; }}
# Devianzas
for (i in 1:I) { devG.r[i,i] <- 0; devX.r[i,i] <- 0;
for (j in 1:i-1) {
devG.r[i,j]<-m[i,j]*log((m[i,j]+0.5)/(mu[i,j]+0.5))-(m[i,j]-mu[i,j]);
devX.r[i,j] <- (m[i,j]-mu[i,j])*(m[i,j]-mu[i,j])/mu[i,j]; }
for(j in i+1:I){
devG.r[i,j]<-m[i,j]*log((m[i,j]+0.5)/(mu[i,j]+0.5))-(m[i,j]-mu[i,j]);
devX.r[i,j] <- (m[i,j]-mu[i,j])*(m[i,j]-mu[i,j])/mu[i,j]; }}
G2 <- 2*sum( devG[,] ); X2 <- sum( devX[,] );
G2.r <- 2*sum( devG.r[,] ); X2.r <- sum( devX.r[,] );
}
```

Modelo

```
# MOVILIDAD SOCIAL
model {# INICIALES
delta[1] <- 1; gamma[1] <- 1;beta[1] <- 1;alpha[1] <- 1;
for (i in 2:IM) { delta[i] ~ dgamma(0.01,0.01) }
for (i in 2:I) { beta[i] ~ dgamma(0.01,0.01);
alpha[i] ~ dgamma(0.01,0.01);
gamma[i] ~ dgamma(0.01,0.01); }
mu.d ~ dgamma(0.01,0.01); mu.od ~ dgamma(0.01,0.01);
# VEROSIMILITUD
for (i in 2:I) { for (j in 1:i-1) {
m[i,j] ~ dpois(mu[i,j]);
mu[i,j] <- mu.d*alpha[i]*beta[j]*delta[i-j];}}
for (i in 1:I-1) { for (j in i+1:I) {
m[i,j] ~ dpois(mu[i,j]);
mu[i,j] <- mu.d*alpha[i]*beta[j]*delta[j-i]; }}
for (i in 1:I) { m[i,i] ~ dpois(mu[i,i]);
mu[i,i] <- mu.od*gamma[i]; }
# AJUSTE
for (i in 1:I) { for (j in 1:I) {
devG[i,j]<-m[i,j]*log((m[i,j]+0.5)/(mu[i,j]+0.5))-(m[i,j]-mu[i,j]);
devX[i,j] <- (m[i,j]-mu[i,j])*(m[i,j]-mu[i,j])/mu[i,j]; }}
G2 <- 2 * sum( devG[,] );
X2 <- sum( devX[,] )
}
```

Datos

```
# Datos para el modelo de cuasi-simetría
list(m = structure(.Data=c(50, 45, 8, 18, 8, 28, 174, 84, 154, 55,
11, 78, 110, 223, 96, 14, 150, 185, 714, 447, 0, 42, 72, 320, 411),
```

```
.Dim = c(5,5)), I=5)
# Datos para el modelo de movilidad social
list(m = structure(.Data=c(50, 45, 8, 18, 8, 28, 174, 84, 154, 55,
11, 78, 110, 223, 96, 14, 150, 185, 714, 447, 0, 42, 72, 320, 411),
.Dim = c(5,5)), I=5, IM=4)
```

Iniciales

```
# Iniciales para el modelo de cuasi-simetría
list(u1=c(NA,1,1,1,1), u2=c(NA,1,1,1,1), u=0, u12=structure(.Data=
c(NA, NA, NA, NA, NA, NA, 0, 0, 0, 0, NA, NA, 0, 0, 0, NA, NA, NA,
0, 0, NA, NA, NA, NA, 0), .Dim = c(5,5)))
# Iniciales para el modelo de movilidad social
list(alpha = c(NA,0.9,0.7,2,1.2), beta = c(NA,0.7,0.5,1.6,1.3),
gamma = c(NA,1,1,1,1), delta = c(NA,0.6,0.27,0.084), mu.d=200,
mu.od=200)
```

C.4. *Matched Pairs* por Grupo de Sangre

Modelo

```
model { mu.g ~ dnorm(0,0.001);
alpha[1] <- 0; gamma[1] <- 0;
for (i in 2:I) { alpha[i] ~ dnorm(0,0.001)
gamma[i] ~ dnorm(0,0.001) }
for (i in 1:I-1) { for (j in i+1:I) { delta[i,j] ~ dnorm(0,0.001)
delta[j,i] <- delta[i,j]
log(mu[i,j]) <- mu.g + delta[i,j] + alpha[i]
log(mu[j,i]) <- mu.g + delta[j,i] + alpha[j] }}
for (i in 1:I) { log(mu[i,i]) <- mu.g+gamma[i]
psi[i] <- exp(alpha[i])
for ( j in 1:I) { m[i,j] ~ dpois(mu[i,j]);
devG[i,j]<-m[i,j]*log((m[i,j]+0.5)/(mu[i,j]+0.5))-(m[i,j]-mu[i,j]) }}
tdev <- sum(devG[,])
}
```

Datos

```
list(I=4, m = structure(.Data = c(64, 18, 8, 3, 66, 74, 14, 6, 4, 2,
4, 2, 12, 10, 12, 2), .Dim = c(4,4)))
```

Iniciales

```
list(alpha=c(NA, 0, 0, 0), gamma=c(NA, 0, 0, 0))
```

C.5. Frecuencia de Visita

Modelo

```
model {
for (k in 1:K) { prior.theta[k] <- 1 }
# Componentes Dirichlet
post.theta[1] <- sum(f[,])+prior.theta[1]
post.theta[2] <- sum(F[,])-sum(f[,])+prior.theta[2]
# Dirichlet
for (k in 1:K) { theta[k] <- theta.s[k]/sum(theta.s[]) }
theta.s[1] ~ dgamma(post.theta[1],1)
theta.s[2] ~ dgamma(post.theta[2],1)
# Dirichlet para alpha
for (i in 1:I) { alpha[i,1] <- xalpha.1[i]/sum(xalpha.1[1:I])
alpha[i,2] <- xalpha.2[i]/sum(xalpha.2[1:I])
xalpha.1[i] ~ dgamma(post.alpha.1[i],1)
xalpha.2[i] ~ dgamma(post.alpha.2[i],1) }
# Dirichlet for beta
for (j in 1:J) { beta[j,1] <- xbeta.1[j]/sum(xbeta.1[1:J])
beta[j,2] <- xbeta.2[j]/sum(xbeta.2[1:J])
xbeta.1[j] ~ dgamma(post.beta.1[j],1)
xbeta.2[j] ~ dgamma(post.beta.2[j],1) }
# actualiza los componentes Dirichlet para alpha y beta
for (i in 1:I){ prior.alpha.1[i] <- 1
prior.alpha.2[i] <- 1
post.alpha.1[i] <- sum(f[i,])+prior.alpha.1[i]
post.alpha.2[i] <- sum(F[i,])-sum(f[i,])+prior.alpha.2[i] }
for (j in 1:J){ prior.beta.1[j] <- 1
prior.beta.2[j] <- 1
post.beta.1[j] <- sum(f[,j])+prior.beta.1[j]
post.beta.2[j] <- sum(F[,j])-sum(f[,j])+prior.beta.2[j] }
# modelo de muestreo
for (i in 1:I) { for (j in 1:J) { f[i,j] ~ dbin(s[i,j],F[i,j])
# probabilidades de las observaciones de la celda i,j de la clase 1
s[i,j] <- s.c[i,j,1]/sum(s.c[i,j,])
for (k in 1:K) { s.c[i,j,k] <- theta[k]*alpha[i,k]*beta[j,k] }}}
}
```

Datos

```
list(I=3, J=3, K=2, F = structure(.Data = c(43, 16, 3, 6, 11, 10, 9,
18, 16), .Dim = c(3,3)))
```

Iniciales

```
list(f = structure(.Data = c(23,8,2,3,5,5,5,9,8), .Dim = c(3,3))
```

C.6. Clases de Estadística

Modelo

```
model { for (i in 1:2) { z[i] ~ dnorm(mu[i],1)I(,gamma[1]) }
for (i in 3:26) { z[i] ~ dnorm(mu[i],1) I(gamma[y[i]-1],gamma[y[i]])
}
for (i in 27:30) { z[i] ~ dnorm(mu[i],1) I(gamma[4],) }
for (i in 1:30) { mu[i] <- b[1]+b[2]*(SATM[i]-mean(SATM[]))
LL[i] <- -0.5*pow(z[i]-mu[i],2)-0.5*log(6.28)
# residuales
r[i] <- z[i]-mu[i]
# datso latentes y categorías predictivas
znew[i] ~ dnorm(mu[i],1)
ynew[i] <- step(gamma[1]-znew[i])
+ 2*step(znew[i]-gamma[1])*step(gamma[2]-znew[i])
+ 3*step(znew[i]-gamma[2])*step(gamma[3]-znew[i])
+ 4*step(znew[i]-gamma[3])*step(gamma[4]-znew[i])
+ 5*step(znew[i]-gamma[4])
Match[i] <- equals(ynew[i],y[i]) }
# total de concordancia
TMatch <- sum(Match[])/30
D <- -2*sum(LL[])
b[1] ~ dnorm(0,0.0001)
b[2] ~ dnorm(0,10)
gamma[1] <- 0
gamma[2] ~ dnorm(1,0.1) I(min[2],max[2])
gamma[3] ~ dnorm(2,0.1) I(min[3],max[3])
gamma[4] ~ dnorm(3,0.1) I(min[4],max[4])
min[2] <- ranked(z[3:9],7); max[2] <- ranked(z[10:16],1)
min[3] <- ranked(z[10:16],7); max[3] <- ranked(z[17:26],1)
min[4] <- ranked(z[17:26],10); max[4] <- ranked(z[27:30],1) }
```

Iniciales

```
list(b = c(0,0), gamma = c(NA,1,2,3))
list(b = c(1.5,0.024), gamma = c(NA,1,2.2,3.6), z = c(-0.4, -1.1,
0.5, 0.6, 0.7, 0.4, 0.6, 0.5, 0.4, 1.7, 1.7, 1.6, 1.6, 1.7, 1.7,
1.8, 2.7, 2.8, 2.8, 2.6, 2.8, 2.9, 2.9, 2.8, 2.7, 2.7, 4.5, 4.5,
```

```
5.1, 4.1))
```

Datos

```
y[ ] SATM[ ]; 1 557; 1 463; 2 525; 2 533; 2 582; 2 471; 2 557;
2 517; 2 488; 3 581; 3 572; 3 559; 3 543; 3 574; 3 582; 3 591;
4 545; 4 576; 4 576; 4 525; 4 574; 4 595; 4 584; 4 584; 4 563;
4 553; 5 609; 5 599; 5 649; 5 549; END
```

C.7. Multinomial Negativa

En este apartado se presentan algunos programas de simulación e inferencia para datos con distribución multinomial negativa.

C.7.1. Simulación

La simulación de los datos se hizo utilizando el programa R. La simulación se hizo a través de la distribución Poisson-gamma multivariada, es decir, utilizando una mezcla de variables independientes con distribución Poisson con una distribución gamma.

Los datos que se generan son 1000 valores que provienen de una distribución Poisson-gamma multivariada de dimensión 3, con índice $r = 10$, y vector de medias $\lambda = (5, 10, 15)$.

```
## Simulación Multinomial Negativa
## Simulación Poisson-Gamma
## Z ~ Poisson(lambda*gama)
## gama ~ Gamma(alpha,beta)

## Supuestos
n <- 1000
dimen <- 3
lambda <- c(5,10,15)
r <- 10
seed <- 1000000

## Números aleatorios
alpha <- r
beta <- r
theta <- array(1,dim=c(n,dimen))
z1 <- array(1,dim=c(n,dimen))
set.seed(seed)
```

```
gama <- rgamma(n,shape=alpha,rate=beta)
for(i in 1:n){ for(j in 1:dimen){
theta[i,j] <- gama[i]*lambda[j]
}}
set.seed(seed)
for(i in 1:n){ for(j in 1:dimen){
z1[i,j] <- rpois(1,theta[i,j])
}}

## Gráficas
op <- par(mfrow=c(1,3))
for(k in 1:dimen){
maxz <- max(z1[,k])
z <- seq(0,maxz,by=1)
probz <- dnbinom(z,size=r,mu=lambda[k])
hist(z1[,k],prob=TRUE,right=FALSE,breaks=c(0:maxz))
lines(z+0.5,probz,col=\blue",type=\o")
}
```

C.7.2. Inferencia

Para hacer inferencia de datos con distribución binomial negativa y multinomial negativa se realizaron programas de simulación usando el programa WinBUGS. Los datos utilizados en los ejemplos se generaron por medio del programa realizado en R presentado anteriormente. Se presentan distintos programas debido a que la inferencia se puede realizar a través de distintas aproximaciones.

Distribución Binomial Negativa

Se simularon 50 datos de una distribución binomial negativa con índice $r = 5$ y parámetro $p = 0.75$. Se realizaron 4 modelos en el programa WinBUGS para obtener la inferencia de los datos debido a que este cálculo se puede hacer por medio de diferentes aproximaciones.

```
## Datos: números aleatorios con distribución BinNeg(5,0.75)
list(y=c(18, 17, 22, 11, 3, 29, 13, 7, 15, 23, 19, 17, 7, 13, 11, 18,
21, 12, 9, 7, 5, 7, 18, 26, 12, 4, 10, 6, 14, 23, 16, 21, 13, 8, 27,
19, 13, 24, 26, 10, 28, 2, 14, 4, 12, 14, 5, 20, 19, 20), K=50, r=5)
```

Modelo 1:

$$y \sim BinNeg(r, \pi).$$

Iniciales:

$$\pi \sim Beta(1,1).$$

```
## Modelo 1
model{
## Familia paramétrica (verosimilitud)
for(k in 1:K){ y[k] ~ dnegbin(pic,r) }
## Distribución inicial
pic <- 1-pi
pi ~ dbeta(1,1)
## Predicción
y.new ~ dnegbin(pic,r)
}

## Iniciales
list(pi=0.75,y.new=1)
```

Modelo 2:

$$\begin{aligned} y &\sim Poisson(\gamma\lambda) \\ \gamma &\sim Gamma(r,1). \end{aligned}$$

Iniciales:

$$\lambda = \frac{\pi}{1-\pi}, \qquad \pi \sim Beta(1,1).$$

```
## Modelo 2
model{
## Familia paramétrica (verosimilitud)
for(k in 1:K){
y[k] ~ dpois(theta[k])
theta[k] <- gamma[k]*lambda
gamma[k] ~ dgamma(r,1)
}
## Distribución inicial
lambda <- (pi)/(1-pi)
pi ~ dbeta(1,1)
## Predicción
y.new ~ dpois(theta.new)
theta.new <- gamma.new*lambda
gamma.new ~ dgamma(r,1)
}
```

```
## Iniciales
list(pi=0.75, y.new=1, gamma.new=1, gamma=c(1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1))
```

Modelo 3:

$$\begin{aligned} y &\sim Poisson(\theta) \\ \theta &\sim Gamma(r, \beta). \end{aligned}$$

Iniciales:

$$\beta = \frac{1-\pi}{\pi}, \qquad \pi \sim Beta(1,1).$$

```
## Modelo 3
model{
## Familia paramétrica (verosimilitud)
for(k in 1:K){
y[k] ~ dpois(theta[k])
theta[k] ~ dgamma(r,be)
} ## Distribución inicial
be <- (1-pi)/pi
pi ~ dbeta(1,1)
## Predicción
y.new ~ dpois(theta.new)
theta.new ~ dgamma(r,be)
}

## Iniciales
list(pi=0.75, y.new=1, theta.new=1, theta=c(1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1))
```

Modelo 4:

$$\begin{aligned} y &\sim Poisson(\gamma\lambda) \\ \gamma &\sim Gamma(r, \beta). \end{aligned}$$

Iniciales:

$$\pi = \frac{\lambda}{\lambda+\beta}, \qquad \lambda \sim Gamma(1,1), \qquad \beta \sim Gamma(1,1).$$

```
## Modelo 4
```

```
model{
## Familia paramétrica (verosimilitud)
for(k in 1:K){
y[k] ~ dpois(theta[k])
theta[k] <- gamma[k]*lambda
gamma[k] ~ dgamma(r,be)
}
## Distribución inicial
pi <- lambda/(lambda+be)
lambda ~ dgamma(1,1)
be ~ dgamma(1,1)
## Predicción
y.new ~ dpois(theta.new)
theta.new <- gamma.new*lambda
gamma.new ~ dgamma(r,be)
}

## Iniciales
list(lambda=1, be=1, y.new=1, gamma.new=1, gamma=c(1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1))
```

Distribución Multinomial Negativa

Se simularon 99 muestras de una distribución multinomial negativa con vector de medias (5,10,15) e índice $\alpha = 10$. Se realizaron 2 modelos en WinBUGS para hacer la inferencia de los datos mediante dos aproximaciones diferentes.

```
## Datos
## Multinomial negativa, media=(5,10,15), indice=10
list(T=3,K=99, alpha=10, beta=1, a=c(1,1,1,1))
y[,1]=c(3, 9, 6, 5, 5, 8, 4, 5, 4, 3, 4, 7, 4, 3, 6, 4, 7, 3, 3, 7,
2, 10, 6, 8, 6, 4, 5, 5, 4, 7, 7, 3, 6, 8, 8, 2, 16, 2, 3, 8, 1, 9,
6, 2, 5, 4, 3, 3, 1, 8, 2, 3, 2, 5, 5, 1, 3, 8, 5, 6, 7, 6, 2, 6, 4,
9, 0, 3, 5, 1, 9, 8, 3, 3, 5, 11, 5, 3, 6, 2, 3, 7, 4, 8, 7, 8, 2,
8, 7, 2, 6, 2, 9, 2, 6, 1, 4, 4, 7, 1)
y[,2]=c(6, 7, 7, 7, 6, 6, 11, 13, 14, 10, 15, 16, 16, 10, 8, 6, 8,
18, 14, 17, 4, 7, 14, 10, 8, 10, 16, 7, 14, 7, 14, 7, 21, 8, 14, 5,
16, 7, 14, 5, 2, 19, 13, 9, 17, 8, 7, 3, 11, 9, 14, 7, 6, 4, 9, 11,
7, 4, 5, 10, 18, 12, 9, 17, 12, 13, 8, 4, 10, 5, 16, 12, 7, 7, 8,
11, 9, 7, 12, 4, 10, 23, 6, 11, 13, 13, 11, 8, 8, 2, 16, 6, 7, 4,
12, 7, 5, 8, 8, 7)
y[,3]=(9, 15, 26, 17, 12, 11, 12, 13, 18, 16, 10, 8, 23, 12, 11, 7,
```

```
23, 21, 8, 19, 11, 18, 23, 27, 13, 12, 15, 13, 22, 19, 21, 9, 12,
16, 24, 9, 12, 11, 12, 9, 9, 22, 14, 12, 27, 25, 8, 18, 9, 13, 14,
7, 12, 19, 13, 10, 24, 12, 10, 8, 19, 17, 10, 15, 21, 16, 13, 14,
16, 5, 21, 25, 16, 11, 13, 21, 19, 11, 12, 10, 14, 22, 10, 15, 23,
26, 18, 17, 9, 9, 24, 6, 18, 9, 14, 9, 21, 11, 20, 7)
```

Modelo 1:

$$\begin{aligned} y_t &\sim Poisson(\gamma\lambda_t) \quad t = 1, \ldots, T \\ \gamma &\sim Gamma(\alpha, \beta). \end{aligned}$$

Iniciales:

$$\lambda_t = \frac{\beta\pi_t}{1 - \sum_{i=1}^{T} \pi_i},$$

$$\pi_t^c \sim Beta(aa[t], bb[t]), \qquad aa[t] = a[t], \qquad bb[t] = \sum_{i=t+1}^{T+1} a_i,$$

$$\pi_t = \pi_t^c * \left(1 - \sum_{i=1}^{t} \pi_i\right) \; t = 2, \ldots, T, \qquad \pi_1 = \pi_1^c, \qquad \pi_{T+1} = 1 - \sum_{t=1}^{T} \pi_t.$$

```
## Modelo 1
model{
## Familia paramétrica (verosimilitud)
for(k in 1:K){
for(t in 1:T){
y[k,t] ~ dpois(theta[k,t])
theta[k,t] <- gamma[k]*lambda[t]
}
gamma[k] ~ dgamma(alpha,beta)
}
## Distribución inicial
for(t in 1:T){
lambda[t] <- beta*pi[t]/(1-sum(pi[1:T]))
}
pi[T+1] <- 1-sum(pi[1:T])
for(i in 1:T){
pip[i] ~ dbeta(aa[i],bb[i])
aa[i] <- a[i]
bb[i] <- sum(a[i:T])+a[T+1]-a[i]
}
for(j in 2:T){
pi[j] <- pip[j]*(1-sum(pi[1:j-1]))
```

```
}
pi[1] <- pip[1]
## Predicción
for(t in 1:T){
y.new[t] ~ dpois(theta.new[t])
theta.new[t] <- gamma.new*lambda[t]
}
gamma.new ~ dgamma(alpha,beta)
}

## Datos
## Multinomial negativa, media=(5,10,15), indice=10
list(T=3, K=99, alpha=10, beta=1, a=c(1,1,1,1)) # Datos

## Iniciales
list(pip=c(0.33,0.33,0.34), y.new=c(1,1,1), gamma.new=1, gamma=c(1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1))
```

Modelo 2:

$$\begin{aligned} y_t &\sim Poisson(\gamma\lambda_t) \quad t = 1, \ldots, T \\ \gamma &\sim Gamma(\alpha, \beta). \end{aligned}$$

Iniciales:

$$\pi_t = \frac{\lambda_t}{\sum_{i=1}^{T} \lambda_i + \beta} \qquad \pi_{T+1} = \frac{\beta}{\sum_{i=1}^{T} \lambda_i + \beta}$$
$$\lambda_t \sim Gamma(1,1) \qquad \beta \sim Gamma(1,1)$$

```
## Modelo
model{
## Familia paramétrica (verosimilitud)
for(k in 1:K){
for(t in 1:T){
y[k,t] ~ dpois(theta[k,t])
theta[k,t] <- gamma[k]*lambda[t]
}
gamma[k] ~ dgamma(alpha,beta)
```

```
}
## Distribución inicial
for(t in 1:T){
pi[t] <- lambda[t]/(sum(lambda[])+beta)
lambda[t] ~ dgamma(1,1)
}
beta ~ dgamma(1,1)
pi[T+1] <- beta/(sum(lambda[])+beta)
## Predicción
for(t in 1:T){
y.new[t] ~ dpois(theta.new[t])
theta.new[t] <- gamma.new*lambda[t]
}
gamma.new ~ dgamma(alpha,beta)
}

## Datos
## Multinomial negativa, media=(5,10,15), indice=10
list(T=3,K=99,alpha=10) # Datos

## Iniciales
list(lambda=c(1,1,1), beta=1, y.new=c(1,1,1), gamma.new=1, gamma=c(1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1))
```

Bibliografía

[1] Agresti, A. (1984): *Analysis of Ordinal Categorical Data*. Wiley, New York.

[2] Agresti, A. (1996): *An Introduction to Categorical Data Analysis*. Wiley, New York.

[3] Agresti, A. (2002): *Categorical Data Analysis*. 2ª edición, Wiley, New York.

[4] Aitkin, M. (1979): "A Simultaneous Test Procedure for Contingency Table Models". *Applied Statistics*. **28**, 233-242.

[5] Albert, J. H. (1996): "Bayesian Selection of Log-Linear Models". *The Canadian Journal of Statistics*. **24**, 327-347.

[6] Albert, J. H. & Chib, S. (1993): "Bayesian Analysis of Binary and Polychotomous Response Data". *Journal of the American Statistical Association*. **88**, 669-679.

[7] Al-Osh, M. A. & Alzaid, A. A. (1987): "First-Order Integer Valued Autoregressive (INAR(1)) Process". *Journal of Time Series Analysis*. **8**, 261-275.

[8] Al-Osh, M. A. & Alzaid, A. A. (1990): "An Integer-Valued pth-Order Autoregressive Structure (INAR(p)) Process". *Journal of Applied Probability*. **27**, 314-324.

[9] Barndorff-Nielsen, O. E. (1978): *Information and Exponential Families in Statistical Theory*. Wiley, Chichester.

[10] Bernardo, J. M. & Smith, A. F. M. (1994): *Bayesian Theory*. Wiley, Chichester.

[11] Bishop, Y., Fienberg, S. & Holland, P. (1975): *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, London.

[12] Böckenholt, U. (1999a): "An INAR(1) Negative Multinomial Regression Model for Longitudinal Count Data". *Psychometrika*. **64**, 53-67.

[13] Böckenholt, U. (1999b): "Analyzing Multiple Emotions Over Time by Autoregressive Negative Multinomial Regression Models". *Journal of the American Statistical Association*. **94**, 757-765.

[14] Brockwell, P. & Davis, R. (1998): *Time Series: Theory and Methods*. Segunda edición. Springer.

[15] Brockwell, P. & Davis, R. (2002): *Introduction to Time Series and Forecasting*. Springer.

[16] Cameron, A. C. & Trivedi, P. K. (1998): *Regression Analysis of Count Data*. Cambridge University Press, Reino Unido.

[17] Chatfield, C. (2003): *The Analysis of Time Series: An Introduction*. Chapman Hall.

[18] Chen, M.-H., Shao, Q.-M. & Ibrahim, J. G. (2000): *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

[19] Congdon, P. (2001): *Bayesian Statistical Modelling*. Wiley, England.

[20] Congdon, P. (2005): *Bayesian Models for Categorical Data*. Wiley, England.

[21] Dickey, J. M., Jiang, J.-M. & Kadane, J. B. (1987): "Bayesian methods for censored categorical data". *Journal of the American Statistical Association*. **82**, 773-781.

[22] Efstathiou, M., Gutiérrez-Peña, E. & Smith, A. F. M. (1998): "Laplace Approximations for Natural Exponential Families with Cuts". *Scandinavian Journal of Statistics*. **25**, 77-92.

[23] Evans, M. J., Gilula, Z. & Guttman, I. (1989): "Latent Class Analysis of Two-Way Contingency Tables by Bayesian Methods". *Biometrika*. **76**, 557-563.

[24] Forster, J. J. & Smith, P. W. F. (1998): "Model-Based Inference for Categorical Survey Data Subject to Non-Ignorable Non-Response". *Journal of the Royal Statistical Society B*. **60**, 57-70.

[25] Gibbons, J. D. & Chakraborti, S. (2003): *Nonparametric Statistical Inference*. Marcel Dekker, edición 4, USA.

[26] Gilks, W. R., Richardson, S. & Spiegelhalter, D.J. (1996): *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

[27] Glass, D. V. (1954): *Social Mobility in Britain*. Glencoe, III.: Free Press.

[28] Good, I. J. (1967): "A Bayesian Significance Test for Multinomial Distributions". *Journal of the Royal Statistical Society. Series B (Methodological)*. **29**, 339-431.

[29] Good, I. J. (1976): "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables". *The Annals of Statistics*. **4**, 1159-1189.

[30] Goodman, L. A. (1981): "Criteria for Determining Whether Certain Categories in a Cross-Classification Table Should Be Combined, with Special Reference to Occupational Categories in an Occupational Mobility Table". *The American Journal of Sociology.* **87**, 612-650.

[31] Gutiérrez Peña, E. (1997): *Métodos Computacionales en la Inferencia Bayesiana.* IIMAS, UNAM, México.

[32] Gutiérrez Peña, E. (1998): *Análisis Bayesiano de Modelos Jerárquicos Lineales.* IIMAS, UNAM, México.

[33] Gutiérrez-Peña, E. (2005): "Bayesian Methods for Categorical Data". *Encyclopedia of Statistics in Behavioral Science.* **1**, 139-146.

[34] Hall, D. B. (2000): "Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study". *Biometrics.* **56**, 1030-1039.

[35] Ip, E. H. & Wang, Y. J. (2007): "A Note on Cuts for Contingency Tables". Por aparecer en el *Journal of Multivariate Analysis.*

[36] Kass, R. E. & Raftery, A. E. (1995): "Bayes Factor". *Journal of the American Statistical Association.* **90**, 773-795.

[37] Laird, N. M. (1978): "Empirical Bayes Methods for Two-Way Contingency Tables". *Biometrika.* **65**, 581-590.

[38] Lambert, D. (1992): "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing". *American Statistical Association and American Society for Quality Control.* **34**, 1-14.

[39] Leonard, T. (1975): "Bayesian Estimation Methods for Two-Way Contingency Tables". *Journal of the Royal Statistical Society. Series B (Methodological).* **37**, 23-37.

[40] Leonard, T. (2000): *A Course in Categorical Data Analysis.* Chapman & Hall, London.

[41] Leonard, T. & Hsu, J. (1994): "The Bayesian Analysis of Categorical Data". *In Aspects of Uncertainty. A Tribute to D. V. Lindley.* John Wiley & Sons Ltd, New York, 283-310.

[42] Leonard, T. & Hsu, J. (1999): *Bayesian Methods.* Cambridge, Reino Unido.

[43] Lindley, D. V. (1964): "The Bayesian Analysis of Contingency Tables". *The Annals of Mathematical Statistics.* **35**, 1622-1643.

[44] Lovison, G. (1994): "Log-Linear Modelling of Data from Matched Case-Control studies". *Journal of Applied Statistics.* **21**, 125-141.

[45] McKenzie, E. (1988): "Some ARMA Models for Dependent Sequences of Poisson Counts". *Advances in Applied Probability.* **20**, 822-835.

[46] Monroy, E. (2006): *WinBUGS: un Software para Inferencia Bayesiana.* Tesis de Licenciatura (Actuaría). Facultad de Ciencias, UNAM. México.

[47] Park, T. & Brown, M. B. (1994): "Models for Categorical Data with Nonignorable Nonresponse". *Journal of the American Statistical Association.* **89**, 44-52.

[48] Paulino, C. D. & Pereira, C. A. (1995): "Bayesian Methods for Categorical Data under Informative General Censoring". *Biometrika.* **82**, 439-446.

[49] Paulino, C. D., Soares, P. & Neuhaus, J. (2003): "Binomial Regression with Misclassification". *Biometrics.* **59**, 670-675.

[50] Radelet, M. (1981): "Racial Characteristics and the Imposition of the Death Penalty". *Amer. Sociol. Rev.* **46**, 918-927.

[51] Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002): "Bayesian Measures of Model Complexity and Fit". *Journal of the Royal Statistical Society.* Series B, **64**, 583-639.

[52] Steutel, F. W. & Van Harn, K. (1979): "Discrete Analogues of Self-Decomposability and Stability". *Annals of Probability.* **7**, 893-899.

[53] Walker, S. (1996): "A Bayesian Maximum a Posteriori Algorithm for Categorical Data under Informative General Censoring". *The Statistician.* **45**, 293-298.

[54] Waller, L. A. & Zelterman, D. (1997): "Log-Linear Modeling with the Negative Multinomial Distribution". *Biometrics.* **53**, 971-982.

[55] Zeger, S. L. (1988): "A Regression Model for Time Series of Counts". *Biometrika.* Vol. 75, **4**, 621-629.