Revista Colombiana de Estadística
Junio 2009, volumen 32, no. 1, pp. 99 a 121

Reducción de modelos en la presencia de parámetros de perturbación

Reduction of Models in the Presence of Nuisance Parameters

Rafael Farias[1,a], Germán Moreno[1,2,b], Alexandre Patriota[1,c]

[1]Departamento de Estadística, Instituto de Matemática y Estadística, Universidad de São Paulo, São Paulo, Brasil

[2]Escuela de Matemáticas, Universidad Industrial de Santander (UIS), Bucaramanga, Colombia

Resumen

En muchos problemas de inferencia estadística existe interés en estimar solamente algunos elementos del vector de parámetros que definen el modelo adoptado. Generalmente, esos elementos están asociados a las medidas de localización, y los parámetros adicionales -que en la mayoría de las veces están en el modelo solo para controlar la dispersión o la asimetría- son conocidos como parámetros de perturbación o de incomodidad (*nuisance parameters*) de las distribuciones subyacentes. Es común estimar todos los parámetros del modelo y hacer inferencias exclusivamente para los parámetros de interés. Dependiendo del modelo adoptado, este procedimiento puede ser muy costoso, tanto algebraica como computacionalmente, por lo cual conviene reducirlo para que dependa únicamente de los parámetros de interés. En este artículo, hacemos una revisión de los métodos de estimación en la presencia de parámetros de perturbación y consideramos algunas aplicaciones en modelos recientemente discutidos en la literatura.

***Palabras clave*:** estimación, parámetro de perturbación, función de verosimilitud, suficiencia, información auxiliar.

Abstract

In many statistical inference problems, there is interest in estimation of only some elements of the parameter vector that defines the adopted model. In general, such elements are associated to measures of location and the additional terms, known as nuisance parameters, to control the dispersion and asymmetry of the underlying distributions. To estimate all the parameters

[a]Estudiante de doctorado. E-mail: rfarias@ime.usp.br
[b]Profesor asistente. E-mail: gmorenoa@uis.edu.co
[c]Estudiante de doctorado. E-mail: patriota@ime.usp.br

of the model and to draw inferences only on the parameters of interest. Depending on the adopted model, this procedure can be both algebraically is common and computationally very costly and thus it is convenient to reduce it, so that it depends only on the parameters of interest. This article reviews estimation methods in the presence of nuisance parameters and consider some applications in models recently discussed in the literature.

***Key words*:** Estimation, Nuisance parameter, Likelihood function, Sufficiency, Ancillarity.

1. Introducción

Uno de los principales objetivos de la estadística es inferir sobre determinada población apoyada solamente en la información de una parte de ella (muestra). Usualmente, estamos interesados en determinada cantidad como la media, mediana, varianza, asimetría, curtosis, coeficiente de correlación, entre otras. Algunas veces, deseamos encontrar y explicar relaciones entre variables y hacer previsiones sobre los valores futuros de la variable estudiada.

En cualquier situación práctica, inicialmente debemos identificar qué cantidades de la población son de principal interés. Después de definidas estas cantidades, es natural suponer un modelo estadístico que se adecue al problema. Por ejemplo, supóngase que el investigador está interesado en los parámetros de localización y de escala. En este caso específico, el vector de interés es $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)^\top$, y suponiendo el modelo estadístico $\mathcal{F} = \left\{N\left(\mu, \sigma^2\right) : \mu \in I\!R \text{ y } \sigma^2 \in I\!R^+\right\}$, siendo $I\!R$ el conjunto de los números reales y $I\!R^+$ el conjunto de los números reales positivos, tenemos que el vector de interés es el vector que define la familia $\mathcal{F}$; por tanto, no existen parámetros de perturbación. Si $X_1, \ldots, X_n$ es una muestra aleatoria de la población objetivo, entonces, para estimar el vector $\boldsymbol{\theta}$ basta encontrar un estadístico suficiente y completo que sea no sesgado; $\widehat{\boldsymbol{\theta}} = \left(\overline{X}, S^2\right)^\top$, siendo $\overline{X} = \sum_i X_i/n$ y $S^2 = \sum_i \left(X_i - \overline{X}\right)^2/(n-1)$, cumple estas condiciones (véase Lehmann & Casella 1998); entonces, el problema inferencial se resuelve, dado que toda la información de la muestra está concentrada en el estadístico $\widehat{\boldsymbol{\theta}}$.

Si el vector de interés define por completo el modelo estadístico adoptado, estamos en el problema de la inferencia usual. Se deben encontrar estimadores óptimos según algún criterio de optimización. Por ejemplo, estimadores no sesgados de varianza uniformemente mínima (obtenidos minimizando una función de pérdida cuadrática), estimadores invariantes según algún grupo de transformaciones (de escala, de origen, de permutaciones, entre otras), estimadores que minimicen el riesgo máximo generado por un subespacio paramétrico (estimador minimax), estimadores que minimicen el riesgo según alguna distribución *a priori* (estimadores de Bayes). Todos esos estimadores dependen de estadísticos suficientes minimales o completos (si existen) que, a su vez, se relacionen con estadísticos auxiliares. Las propiedades de estos estimadores pueden ser vistas con detalles en Lehmann & Casella (1998) y Lindsey (1996). Si el vector de interés no define por completo el modelo estadístico, entonces existen parámetros de perturbación y es

preciso encontrar estimadores óptimos siguiendo otros criterios, como suficiencia e información parcial.

Para ilustrar la idea de parámetros de perturbación, suponga que $X_1, \ldots, X_n$ es una muestra aleatoria de la población objeto de estudio. Considere que el modelo estadístico propuesto para describir el comportamiento de los datos observados es

$$\mathcal{F} = \left\{ SN(\boldsymbol{\theta}) : \boldsymbol{\theta} = \left(\mu, \sigma^2, \lambda\right)^\top, \quad \text{con} \quad \mu, \lambda \in I\!R \quad \text{y} \quad \sigma^2 \in I\!R^+ \right\} \tag{1}$$

siendo $SN(\mu, \sigma^2, \lambda)$ una distribución normal-asimétrica (*Skew-Normal*), con μ, σ^2 y λ los parámetros de localización, escala y asimetría, respectivamente. La función de densidad de la normal-asimétrica definida por Azzalini (1985) es dada por

$$f\left(x \mid \mu, \sigma^2\right) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right), \quad x \in I\!R \tag{2}$$

siendo $\phi(\cdot)$ y $\Phi(\cdot)$ la función de densidad y la distribución acumulada de la distribución normal estándar, respectivamente. Las propiedades de esta distribución pueden ser encontradas en Azzalini (1985). Considerando que estamos interesados solamente en los parámetros de localización y escala, podemos escribir el vector de parámetros para la distribución definida en (2) como $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, donde $\boldsymbol{\theta}_1 = \left(\mu, \sigma^2\right)^\top$ y $\boldsymbol{\theta}_2 = \lambda$. En este caso, el vector de interés $\boldsymbol{\theta}_1$ no coincide con el vector de parámetros que indexa la familia de distribuciones $\mathcal{F}$ y λ es un parámetro de perturbación para la estimación de $\boldsymbol{\theta}_1$. Obsérvese que, cuando $\lambda = 0$, el modelo (2) se reduce al modelo normal y, por tanto, no existe parámetro de perturbación.

En ciertas ocasiones, la dimensión del vector de parámetros de perturbación crece con el tamaño de la muestra. Neyman & Scott (1948) definen estos parámetros como parámetros incidentales. Para ilustrar esta definición, considere $(Y_1, X_1), \ldots, (Y_n, X_n)$ una muestra aleatoria, cuya relación entre Y_i y X_i está dada por $Y_i = g(\boldsymbol{\theta}_1, x_i) + e_i$ y $X_i = x_i + u_i$, siendo e_i y u_i variables aleatorias independientes para todo $i = 1, \ldots, n$ y $g(\boldsymbol{\theta}_1, x_i)$ una función conocida. Así, el vector de parámetros que define el modelo es $\boldsymbol{\theta}^{(n)} = \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^{(n)\top}\right)^\top$, con $\boldsymbol{\theta}_2^{(n)} = (x_1, \ldots, x_n)^\top$, el vector de parámetros incidentales que generalmente no es de interés del investigador. Este modelo es conocido en la literatura como modelo funcional con errores en las variables y puede ser estudiado con más detalles en Fuller (1987). En este caso, es común hacer inferencias sobre los parámetros de interés usando la función de verosimilitud perfilada, definida en la sección 4.2.

A pesar de que existen diversas formas de tratar modelos que poseen parámetros de perturbación, el enfoque principal de este trabajo se basa en la reducción de modelos. La forma más simple y directa es encontrar una función de verosimilitud ortogonal para el parámetro de interés. Así, en la sección 2.2, introducimos el concepto de verosimilitud ortogonal con algunos ejemplos en modelos asimétricos. En la sección 3, presentamos algunas técnicas de reducción de modelos a través de estadísticos e ilustramos la teoría con algunos ejemplos. En la sección 4, exhibimos dos funciones de verosimilitudes aproximadas que son utilizadas para construir funciones de verosimilitudes ortogonales para los parámetros de interés. Finalizamos el artículo con algunos comentarios de las técnicas presentadas.

El principal objetivo de este artículo es motivar el uso de las técnicas de reducción de modelos ilustrándolas con ejemplos recientemente discutidos en la literatura.

2. Función de verosimilitud

Asumimos en este artículo que $\boldsymbol{\theta}_1$ (la partición de interés) y $\boldsymbol{\theta}_2$ (el vector de parámetros de perturbación) tienen dimensiones p_1 y $p - p_1$, respectivamente. Consideramos también que toda la información de la muestra está contenida en la función de verosimilitud, que está correctamente especificada. El problema consiste en estimar $\boldsymbol{\theta}_1$ minimizando la pérdida de información que puede ocurrir en la estimación de $\boldsymbol{\theta}_2$. La pérdida de información será definida con más detalles en el transcurso del texto.

2.1. Función de verosimilitud genuina

Sea X una variable aleatoria en un espacio de probabilidad $(\Omega, \mathcal{A}, \nu)$, siendo Ω el espacio de posibilidades del experimento, $\mathcal{A} = \sigma(X)$ la σ-álgebra asociada a Ω tal que X es medible y ν una medida de probabilidad aplicada a los elementos de $\mathcal{A}$. Sea $\mathcal{X} \subset \mathbb{R}$ el espacio de valores posibles que X puede asumir. Considere que la distribución de probabilidad de X pertenece a la familia

$$\mathcal{F} = \left\{ F(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top\right)^\top \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p \right\} \tag{3}$$

siendo $F(\cdot \mid \boldsymbol{\theta})$ una función de distribución. Sea $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ una muestra aleatoria de X; denotaremos por $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ la función de verosimilitud genuina asociada a $F(\cdot \mid \boldsymbol{\theta})$. Si X es una variable continua, entonces

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{dF(x_i \mid \boldsymbol{\theta})}{dx_i} = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \tag{4}$$

Si X es una variable discreta, entonces

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{i=1}^{n} \Big[F\big(x_i^+ \mid \boldsymbol{\theta}\big) - F(x_i^- \mid \boldsymbol{\theta}) \Big] = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \tag{5}$$

siendo $\lim_{y \downarrow x} F(y \mid \boldsymbol{\theta}) = F\big(x_i^+ \mid \boldsymbol{\theta}\big)$ y $\lim_{y \uparrow x} F(y \mid \boldsymbol{\theta}) = F(x_i^- \mid \boldsymbol{\theta})$. La función $f(x_i \mid \boldsymbol{\theta})$ denota la función de densidad en el caso continuo y la función de probabilidad en el caso discreto.

En el enfoque clásico es común maximizar la función de verosimilitud $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ en relación con los parámetros del modelo para obtener sus estimadores. Los estimadores de máxima verosimilitud (EMV) son ampliamente usados debido a sus buenas propiedades como invarianza, consistencia, eficiencia y normalidad asintótica, si se satisfacen algunas condiciones de regularidad (ver Lehmann & Casella 1998).

2.2. Función de verosimilitud ortogonal

Suponiendo que $\boldsymbol{X}$ es un vector aleatorio con distribución de probabilidad perteneciente a $\mathcal{F}$, decimos que la función de verosimilitud $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ es ortogonal en relación con la partición de interés si

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = L_1(\boldsymbol{\theta}_1 \mid \boldsymbol{x}) L_2(\boldsymbol{\theta}_2 \mid \boldsymbol{x}) \tag{6}$$

y los vectores $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ tienen variaciones independientes, o sea,

$$\left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top\right)^\top \in \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 = \boldsymbol{\Theta} \subset I\!R^p \tag{7}$$

donde $\boldsymbol{\Theta}_k$ es el espacio paramétrico en que $\boldsymbol{\theta}_k$ puede asumir valores, con $k = 1, 2$. Denotaremos $L_k(\boldsymbol{\theta}_k \mid \boldsymbol{x})$ simplemente por $L_k(\boldsymbol{\theta}_k)$ para $k = 1, 2$.

A partir de la ecuación (6) tenemos que el EMV para $\boldsymbol{\theta}_1$ depende de la función de verosimilitud genuina solamente a través de $L_1(\boldsymbol{\theta}_1)$. En este caso, el EMV de $\boldsymbol{\theta}_1$ no depende de $\boldsymbol{\theta}_2$; luego podemos ignorar la estimación de $\boldsymbol{\theta}_2$, sin que esto interfiera la estimación de los parámetros de interés. Por tanto, podemos definir un nuevo modelo reducido, $\mathcal{F}_1 = \{L_1(\boldsymbol{\theta}_1); \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1\}$, para hacer inferencias sobre $\boldsymbol{\theta}_1$. Es importante notar que, en este caso, la información dada por la estimación de $\boldsymbol{\theta}_2$ es irrelevante en la estimación de $\boldsymbol{\theta}_1$.

Ejemplo 1. Análisis de supervivencia. El principal interés en análisis de supervivencia es estudiar el tiempo hasta la ocurrencia de determinado evento. En esta área de la estadística es común encontrar la presencia de censuras antes de la ocurrencia del evento de interés. En algunas situaciones, es razonable asumir que las censuras no son informativas, o sea, su distribución no comparte parámetros con la función de distribución del tiempo de ocurrencia del evento. Además, se asume también independencia entre las censuras y el evento de interés. Sea T el tiempo hasta la ocurrencia del evento y C el tiempo hasta la censura.

(*) Suponga que $T \sim f(t \mid \boldsymbol{\theta}_1)$ es independiente de $C \sim g(c \mid \boldsymbol{\theta}_2)$, de modo que $\boldsymbol{\theta}_2$ no comparte parámetros con $\boldsymbol{\theta}_1$.

En la práctica se observa el tiempo hasta la ocurrencia del evento o el tiempo hasta la censura, o sea, $Z = \text{mín}\{T, C\}$ y $\delta = I(C \geq T)$. La distribución conjunta de (Z, δ) se obtiene así:

$$\begin{aligned} f(z, \delta = 1 \mid \boldsymbol{\theta}) &= P(\delta = 1 \mid \boldsymbol{\theta}) f(z \mid \delta = 1, \boldsymbol{\theta}) \\ &= P(C \geq T \mid \boldsymbol{\theta}) f(z \mid \boldsymbol{\theta}_1) \\ &= G(z \mid \boldsymbol{\theta}_2) f(z \mid \boldsymbol{\theta}_1) \end{aligned} \tag{8}$$

pues, si $\delta = 1$, entonces $Z = T$.

$$\begin{aligned} f(z, \delta = 0 \mid \boldsymbol{\theta}) &= P(\delta = 0 \mid \boldsymbol{\theta}) f(z \mid \delta = 0, \boldsymbol{\theta}) \\ &= P(C \leq T \mid \boldsymbol{\theta}) g(z \mid \boldsymbol{\theta}_2) \\ &= S(z \mid \boldsymbol{\theta}_1) g(z \mid \boldsymbol{\theta}_2) \end{aligned} \tag{9}$$

y si $\delta = 0$, tendremos $Z = C$. Así, la función de verosimilitud será

$$\begin{aligned} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= f(z, \delta \mid \boldsymbol{\theta}) \\ &= \left[G(z \mid \boldsymbol{\theta}_2) f(z \mid \boldsymbol{\theta}_1)\right]^{\delta} \left[S(z \mid \boldsymbol{\theta}_1) g(z \mid \boldsymbol{\theta}_2)\right]^{1-\delta} \\ &= \left[S(z \mid \boldsymbol{\theta}_1)^{1-\delta} f(z \mid \boldsymbol{\theta}_1)^{\delta}\right] \left[G(z \mid \boldsymbol{\theta}_2)^{\delta} g(z \mid \boldsymbol{\theta}_2)^{1-\delta}\right] \end{aligned} \tag{10}$$

por tanto, la función de verosimilitud puede ser separada en una parte que solo depende del parámetro de interés $\boldsymbol{\theta}_1$ y otra que solo depende del parámetro de perturbación $\boldsymbol{\theta}_2$. Si las censuras no son informativas, podemos usar únicamente $L_1(\boldsymbol{\theta}_1) = S(z \mid \boldsymbol{\theta}_1)^{1-\delta} f(z \mid \boldsymbol{\theta}_1)^{\delta}$ para hacer inferencias sobre $\boldsymbol{\theta}_1$, sin tener pérdida de información.

En la mayoría de las situaciones no es posible tener una función de verosimilitud ortogonal. En algunos modelos, podemos encontrar una reparametrización adecuada, tal que la función de verosimilitud sea ortogonal para el nuevo vector de parámetros. Esto es, podemos definir un nuevo vector de parámetros, $\boldsymbol{\lambda} = \left(\boldsymbol{\lambda}_1^{\top}, \boldsymbol{\lambda}_2^{\top}\right)^{\top}$ con $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_1(\boldsymbol{\theta}_1)$ y $\boldsymbol{\lambda}_2 = \boldsymbol{\lambda}_2(\boldsymbol{\theta})$ de forma que

$$L(\boldsymbol{\lambda}) = L_1^*(\boldsymbol{\lambda}_1) L_2^*(\boldsymbol{\lambda}_2) \tag{11}$$

Asumiendo que $\boldsymbol{\lambda}_1$ es una función biyectiva del vector de interés, podemos usar L_1^* para estimar $\boldsymbol{\lambda}_1$ y, en consecuencia, estimar $\boldsymbol{\theta}_1$. Solo en algunos casos específicos la reparametrización existe y tiene interpretación para el problema analizado.

Lindsey (1996) define varios tipos de reparametrizaciones ortogonales, entre los cuales se pueden citar estimación ortogonal (el EMV de $\boldsymbol{\theta}_1$ no depende del EMV de $\boldsymbol{\theta}_2$), diseño ortogonal (cuando las columnas de la matriz de diseño del modelo de regresión son linealmente independientes), información ortogonal (la matriz de información de Fisher esperada es bloque diagonal en relación a $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$) y la función de verosimilitud ortogonal.

Cuando la función de verosimilitud no es ortogonal y las reparametrizaciones no son viables, se puede escribir la función de verosimilitud de la forma

$$L(\boldsymbol{\theta}) = L_1(\boldsymbol{\theta}_1) L_2(\boldsymbol{\theta}) \tag{12}$$

o sea, siempre será posible factorizar la función de verosimilitud de modo que uno de los factores dependa solamente de $\boldsymbol{\theta}_1$ y otro dependa de una función del vector completo $\boldsymbol{\theta}$. En el caso más extremo, $L_1(\boldsymbol{\theta}_1) = 1$ y $L_2(\boldsymbol{\theta}) = L(\boldsymbol{\theta})$.

Ejemplo 2. Análisis de supervivencia (continuación). Considérese el ejemplo 1 alterando la condición (*) para (**), siendo esta nueva condición definida por:

(**) Suponga que $T \sim f(t \mid \boldsymbol{\theta}_1)$ es independiente de $C \sim g(c \mid \boldsymbol{\theta})$, tal que $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^{\top}, \boldsymbol{\theta}_2^{\top}\right)^{\top}$.

Con la suposición (**), la función de verosimilitud está dada por

$$\begin{aligned} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= f(z, \delta) \\ &= \left[G(z \mid \boldsymbol{\theta}) f(z \mid \boldsymbol{\theta}_1)\right]^{\delta} \left[S(z \mid \boldsymbol{\theta}_1) g(z \mid \boldsymbol{\theta})\right]^{1-\delta} \\ &= \left[S(z \mid \boldsymbol{\theta}_1)^{1-\delta} f(z \mid \boldsymbol{\theta}_1)^{\delta}\right] \left[G(z \mid \boldsymbol{\theta})^{\delta} g(z \mid \boldsymbol{\theta})^{1-\delta}\right] \\ &= L_1(\boldsymbol{\theta}_1) L_2(\boldsymbol{\theta}) \end{aligned} \tag{13}$$

por tanto, si se ignora $L_2(\boldsymbol{\theta})$, se puede perder mucha información en la estimación de $\boldsymbol{\theta}_1$, si usamos únicamente el término $L_1(\boldsymbol{\theta}_1)$.

Existen algunos criterios para escoger la función $L_1(\boldsymbol{\theta}_1)$ tal que conserve toda la información sobre $\boldsymbol{\theta}_1$ contenida en la función de verosimilitud $L(\boldsymbol{\theta})$; por consiguiente, sería razonable despreciar la función $L_2(\boldsymbol{\theta})$ en el proceso de estimación de $\boldsymbol{\theta}_1$. Esto genera la necesidad de definir más precisamente un concepto para pérdida de información, pues sería interesante encontrar $L_1(\boldsymbol{\theta}_1)$ y $L_2(\boldsymbol{\theta})$ tal que la información que $L_2(\boldsymbol{\theta})$ cargue sobre $\boldsymbol{\theta}_1$ sea mínima (o nula). En la próxima sección introducimos algunos conceptos esenciales para determinar tales funciones.

3. Reducción de modelos a través de estadísticos

Sea $\boldsymbol{X}$ un vector aleatorio con distribución de probabilidad perteneciente a $\mathcal{F}$, donde $\mathcal{F} = \left\{F(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^{\top}, \boldsymbol{\theta}_2^{\top}\right)^{\top} \in \boldsymbol{\Theta} \subseteq I\!R^p\right\}$. La reducción de modelos se basa en estadísticos, funciones de $\boldsymbol{X}$, que concentren la mayor parte de la información relevante sobre el vector de interés $\boldsymbol{\theta}_1$ disponible en $\boldsymbol{X}$.

Considere $T = T(\boldsymbol{X})$ y $U = U(\boldsymbol{X})$, estadísticos que dependen únicamente de $\boldsymbol{X}$. La función de densidad conjunta de $(T, U, \boldsymbol{X})$ es dada por

$$f(t, u, \boldsymbol{x} \mid \boldsymbol{\theta}) = f(t \mid \boldsymbol{\theta}) f(u \mid t, \boldsymbol{\theta}) f(\boldsymbol{x} \mid t, u, \boldsymbol{\theta}) \tag{14}$$

Factorizando el lado izquierdo de esta ecuación, obtenemos

$$f(t, u \mid \boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{x} \mid \boldsymbol{\theta}) = f(t \mid \boldsymbol{\theta}) f(u \mid t, \boldsymbol{\theta}) f(\boldsymbol{x} \mid t, u, \boldsymbol{\theta}) \tag{15}$$

Como los estadísticos T y U son determinados por $\boldsymbol{X}$, sus distribuciones condicionales en $\boldsymbol{X}$ son degeneradas. Se sigue que

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = f(t \mid \boldsymbol{\theta}) f(u \mid t, \boldsymbol{\theta}) f(\boldsymbol{x} \mid t, u, \boldsymbol{\theta}) \quad \text{c.s. } \nu \tag{16}$$

siendo que "c.s. ν" significa "casi segura ν", o sea, la relación (16) vale para todo $\boldsymbol{x} \in (\mathcal{X}^n - A)$ tal que $\nu(A) = 0$, donde ν es la medida de probabilidad aplicada a los elementos de A.

3.1. Función de verosimilitud marginal y condicional

En la teoría de la verosimilitud introducida por Fisher, la función de verosimilitud ordinaria es la función de densidad conjunta (o probabilidad) de la muestra $\boldsymbol{X}$ en función del vector de parámetros que define por completo la familia. Siguiendo la idea de la factorización dada antes, podemos definir dos nuevas funciones de verosimilitud.

Definición 1. Sea $\boldsymbol{T}$ un estadístico cuya distribución solo depende de $\boldsymbol{\theta}_1$. La función de verosimilitud marginal está dada por

$$L_M(\boldsymbol{\theta}_1; \boldsymbol{t}) = f(\boldsymbol{t} \mid \boldsymbol{\theta}_1) \quad \text{c.s. } \nu \tag{17}$$

Suponga que $(\boldsymbol{U}, \boldsymbol{T})$ sea un estadístico tal que sea posible obtener la factorización

$$f(t, u \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = f(t \mid \boldsymbol{\theta_1}) f(u \mid t, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \tag{18}$$

Despreciando el término $f(u \mid t, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$, tenemos la función de verosimilitud marginal $L_M(\boldsymbol{\theta}_1; \boldsymbol{t})$ basada en $\boldsymbol{T} = t$.

Definición 2. Sean $\boldsymbol{U}$ y $\boldsymbol{T}$ dos estadísticos tales que la distribución de $\boldsymbol{T}|\boldsymbol{U}$ no dependa de $\boldsymbol{\theta}_2$. La función de verosimilitud condicional está dada por

$$L_C(\boldsymbol{\theta}_1; \boldsymbol{t} \mid \boldsymbol{u}) = f(\boldsymbol{t} \mid \boldsymbol{u}, \boldsymbol{\theta}_1) \quad \text{c.s. } \nu \tag{19}$$

Suponga que $(\boldsymbol{U}, \boldsymbol{T})$ sea un estadístico tal que es posible obtener la factorización

$$f(t, u \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = f(u \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) f(t \mid u, \boldsymbol{\theta_1}) \tag{20}$$

Despreciando el término $f(u \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$, tenemos la función de verosimilitud condicional $L_C(\boldsymbol{\theta}_1; \boldsymbol{t} \mid \boldsymbol{u})$ basada en $\boldsymbol{T} \mid \boldsymbol{U} = u$.

Las funciones de verosimilitudes marginales y condicionales también pueden usarse para hacer inferencias sobre $\boldsymbol{\theta}_1$, pero el precio es la pérdida de información, dado que en los dos casos dejamos de considerar una parte de la función de verosimilitud original. Se pierde el mínimo de información si son utilizados estadísticos con propiedades óptimas como *I-suficiencia*, *I-auxiliar* y *ausencia de información parcial en el sentido extendido*, conceptos definidos en las siguientes secciones.

3.2. Estadístico suficiente y auxiliar

Fisher definió el concepto de estadístico suficiente y auxiliar (*ancillary statistic*) para una familia de distribuciones, esto es, cuando el parámetro de interés coincide con el parámetro que determina por completo la familia. Lindsey (1996) llama a estas clases de estadísticos *F-suficientes* y *F-auxiliares* (*F* por *Full*, total, pues definen totalmente la familia). En el transcurso del texto hablaremos simplemente de estadísticos suficientes y auxiliares, y se definen así:

Definición 3. Un estadístico $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ es suficiente para el vector de parámetros $\boldsymbol{\theta}$ si $f(\boldsymbol{x} \mid t, \boldsymbol{\theta}) = f(\boldsymbol{x} \mid t)$ no depende de $\boldsymbol{\theta}$ c.s. ν.

Para encontrar estadísticos suficientes para una familia se puede utilizar el criterio de la factorización[1] (Halmos & Savage 1949) definido por:

Definición 4. Un estadístico $\boldsymbol{T}$ es suficiente para el vector de parámetros $\boldsymbol{\theta}$ si la función de verosimilitud puede ser factorizada de la forma $L(\boldsymbol{\theta}) = g(\boldsymbol{t} \mid \boldsymbol{\theta})h(\boldsymbol{x})$.

Un ejemplo básico de aplicación de este criterio es el siguiente.

Ejemplo 3. Distribución Poisson. Sea $X_1, \ldots, X_n$ una muestra aleatoria de $X \sim P(\lambda)$, distribución de Poisson de parámetro λ. La función de verosimilitud está dada por

$$\begin{aligned} L(\lambda \mid X_1, \ldots, X_n) &= P(X_1 = x_1 \mid \lambda) \ldots P(X_n = x_n \mid \lambda) \\ &= \frac{\lambda^{x_1} \exp^{-\lambda}}{x_1!} \cdots \frac{\lambda^{x_n} \exp^{-\lambda}}{x_n!} \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \exp^{-n\lambda}}{\prod_{i=1}^n x_i!} \\ &= \left(\lambda^{\sum_{i=1}^n x_i} \exp^{-n\lambda}\right) \frac{1}{\prod_{i=1}^n x_i!} \end{aligned} \tag{21}$$

Por el criterio de la factorización, tenemos que $T = \sum_{i=1}^n x_i$ es un estadístico suficiente para λ.

Definición 5. Un estadístico $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{X})$ es auxiliar para $\boldsymbol{\theta}$ si la distribución de $\boldsymbol{U}$ no depende de $\boldsymbol{\theta}$, o sea, $f(\boldsymbol{u} \mid \boldsymbol{\theta}) = f(\boldsymbol{u})$ c.s. ν.

Asumiendo que $\boldsymbol{T}$ y $\boldsymbol{U}$ son estadísticos suficiente y auxiliar para $\boldsymbol{\theta}$, respectivamente, una consecuencia de las definiciones 3 y 5 es que la función de verosimilitud para $\boldsymbol{\theta}$ puede factorizarse como

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f(\boldsymbol{t} \mid \boldsymbol{\theta}) f(\boldsymbol{x} \mid \boldsymbol{t}) \;\; \text{y} \;\; L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \boldsymbol{u}, \boldsymbol{\theta}) f(\boldsymbol{u}) \quad \text{c.s. } \nu \tag{22}$$

Por tanto, dependiendo del estadístico usado, podemos reducir el modelo $\mathcal{F}$, para $\mathcal{F}_1 = \left\{F(\boldsymbol{t} \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\right\}$ o $\mathcal{F}_1^* = \left\{F(\boldsymbol{x} \mid \boldsymbol{u}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\right\}$.

Ejemplo 4. Distribución alfa-normal. Sea $X_1, \ldots, X_n$ una muestra aleatoria de $X \sim \alpha N(\alpha)$, alfa-normal estándar definida inicialmente por Durrans (1992) y estudiada recientemente por Jones (2004), cuya densidad es dada por

$$f(x \mid \alpha) = \alpha \phi(x) \Phi(x)^{\alpha - 1}, \quad x \in I\!R \tag{23}$$

estando $\phi(\cdot)$ y $\Phi(\cdot)$ definidas en (2). La función de verosimilitud está dada por

$$L(\alpha \mid \boldsymbol{x}) = \alpha^n \left[\prod_{i=1}^n \phi(x_i)\right] \left[\prod_{i=1}^n \Phi(x_i)\right]^{\alpha - 1} \tag{24}$$

Por el criterio de la factorización, tenemos que $T = \prod_i \Phi(X_i)$ es un estadístico suficiente para α.

[1] También conocido en la literatura como criterio de factorización de Neyman-Fisher.

Ejemplo 5. Distribución normal asimétrica. Sea $Y_1, \ldots, Y_n$ una muestra aleatoria de la variable $Y \sim SN\big(0, \sigma^2, \lambda\big)$ definida en (2), con $\sigma^2 = 1$. Usando las propiedades de la distribución *Normal-Asimétrica* derivadas por Azzalini (1985), tenemos que $U = \sum_{i=1}^{n} Y_i^2 \sim \chi^2\,(n)$, distribución chi-cuadrado con n grados de libertad. Entonces, por la definición 5, el estadístico U es auxiliar para λ.

Si optamos por un estadístico suficiente $\boldsymbol{T}$, es deseable que este sea minimal (función de todos los estadísticos suficientes), pues así tendremos la mayor reducción posible en los datos (Pace & Salvan 1997, Lehmann & Casella 1998). Si optamos por un estadístico auxiliar $\boldsymbol{U}$, es conveniente que la misma sea maximal, o sea, no existe otro estadístico auxiliar que sea función de este.

Como el objetivo de este trabajo es estimar solo una parte del vector $\boldsymbol{\theta}$, es conveniente definir estadísticos que contengan información solo sobre una partición del vector que define la familia o modelo en cuestión, es decir, estadísticos que generalicen los conceptos de suficiencia e información auxiliar introducidos por Fisher. A continuación definimos los conceptos de información parcial y ausencia parcial de información.

3.3. Suficiencia y ausencia parcial de información

Definición 6. Si $(\boldsymbol{T}, \boldsymbol{U})$ es suficiente para $\boldsymbol{\theta}$ y, en (16), $f(u \mid t, \boldsymbol{\theta}) = f(u \mid t, \boldsymbol{\theta}_2)$, o sea, la densidad de $\boldsymbol{U} \mid \boldsymbol{T}$ solo depende de $\boldsymbol{\theta}_2$, entonces decimos que $\boldsymbol{T}$ es parcialmente suficiente para $\boldsymbol{\theta}_1$. Además, si los campos de variación de $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ son independientes entre sí, entonces $\boldsymbol{T}$ es llamada *S-suficiente* para $\boldsymbol{\theta}_1$.

Ejemplo 6. Distribución exponencial truncada. Sea $X_1, \ldots, X_n$ una muestra aleatoria de $\boldsymbol{X}$ con distribución exponencial truncada perteneciente a $\mathcal{F} = \big\{E(\boldsymbol{\theta}) : \boldsymbol{\theta} = (\alpha, \beta)^\top \in \boldsymbol{\Theta} = I\!R \times (0, \infty)\big\}$, cuya densidad es dada por

$$f(x \mid \alpha) = \frac{1}{\beta} \exp\left\{ -\frac{(x - \alpha)}{\beta} \right\}, \quad x \in (\alpha, \infty) \tag{25}$$

y su función de verosimilitud por

$$L(\alpha, \beta \mid \boldsymbol{x}) = \beta^{-n} \exp\left\{ \frac{n\alpha}{\beta} \right\} \exp\left\{ -\frac{\sum_i x_i}{\beta} \right\} I(\alpha)_{\left(-\infty, x_{(1)}\right)} \tag{26}$$

donde $x_{(1)} = \text{mín}\{x_1, \ldots, x_n\}$. Utilizando el criterio de la factorización, tenemos que $\boldsymbol{V} = \Big(X_{(1)}, \sum_i X_i\Big)$ es suficiente para $\boldsymbol{\theta} = (\alpha, \beta)^\top$. Al mismo tiempo, el vector $\boldsymbol{V^*} = (U, T)$, con $U = X_{(1)}$ y $T = 2n \sum_i \Big(X_i - X_{(1)}\Big)$, también es suficiente, pues es función $1:1$ de $\boldsymbol{V}$. El estadístico $\boldsymbol{V^*}$ también es completo[2], pues satisface la condición

$$I\!E\big[g(\boldsymbol{V^*})\big] = 0 \Longleftrightarrow g(\boldsymbol{V^*}) = 0, \ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \ \text{ c.s. } \ \nu \tag{27}$$

[2]Si X es una variable aleatoria con distribución perteneciente a una familia $\mathcal{F}_\theta$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, se dice que un estadístico T es completo si para cualquier función medible g se verifica $I\!E_\theta\big[g(T)\big] = 0$, si y solo si $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, $g(T) = 0$, c.s. ν.

Dado que U es estadístico suficiente y completo, T es estadístico auxiliar para β, y esto vale para todo $\beta \in (0, \infty)$, por el Teorema de Basu[3], U y T son independientes y la distribución de $T \mid U$ es igual a la distribución de T, y esta última no depende de α, pues $U \sim E(\alpha, n/\beta)$ y $T \sim \beta\, \chi^2\,(2n)$. Entonces U es un estadístico parcialmente suficiente para α y también es *S-suficiente*, pues $(\alpha, \beta) \in I\!R \times (0, \infty)$.

Definición 7. Si $\boldsymbol{T}$ es degenerada y, en (16), $f(u \mid t, \boldsymbol{\theta}) = f(u \mid \boldsymbol{\theta}_2)$, o sea, la densidad de $\boldsymbol{U}$ solo depende de $\boldsymbol{\theta}_2$, decimos que $\boldsymbol{U}$ es parcialmente auxiliar para $\boldsymbol{\theta}_1$. Además, si los campos de variación de $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ son independientes entre sí, entonces se dice que $\boldsymbol{U}$ es *S-auxiliar* para $\boldsymbol{\theta}_1$.

Ejemplo 7. Distribución normal asimétrica (continuación). Considere el ejemplo 5, $SN(0, \sigma^2, \lambda)$, con σ^2 desconocido. El estadístico $U \sim \sigma^2 \chi^2(n)$ es parcialmente auxiliar para λ, y como los parámetros varían independientemente, entonces U también es *S-auxiliar*.

En las definiciones 6 y 7 establecemos los conceptos de suficiencia e información auxiliar parcial para particiones de un vector. Con tales definiciones es posible retirar de la función de verosimilitud parte de la información que no es relevante en el proceso de estimación del parámetro de interés. Por ejemplo, si el vector $(\boldsymbol{U}, \boldsymbol{T})$ es suficiente para el vector completo $\boldsymbol{\theta}$ y $\boldsymbol{T}$ es un estadístico parcialmente suficiente para $\boldsymbol{\theta}_1$, entonces la función de verosimilitud puede ser factorizada de la forma

$$L(\boldsymbol{\theta}) = f(t|\boldsymbol{\theta}) f(u \mid t, \boldsymbol{\theta}_2) f(\boldsymbol{x} \mid t, u) \quad \text{c.s. } \nu \tag{28}$$

Así, se puede proponer un modelo reducido usando únicamente $f(t \mid \boldsymbol{\theta})$. Si $\boldsymbol{U}$ es parcialmente auxiliar para $\boldsymbol{\theta}_1$, entonces

$$L(\boldsymbol{\theta}) = f(t \mid u, \boldsymbol{\theta}) f(u \mid \boldsymbol{\theta}_2) f(\boldsymbol{x} \mid t, u) \quad \text{c.s. } \nu \tag{29}$$

Por tanto, el modelo reducido puede usar solo $f(t \mid u, \boldsymbol{\theta})$.

A pesar de reducir la función de verosimilitud, esta no se torna ortogonal y, por tanto, el parámetro de perturbación continúa presente. La función de verosimilitud será ortogonal, usando las definiciones 6 y 7, solo cuando exista un estadístico $\boldsymbol{T}^*$ parcialmente suficiente para $\boldsymbol{\theta}_1$ y parcialmente auxiliar para $\boldsymbol{\theta}_2$, o exista un estadístico $\boldsymbol{U}^*$ parcialmente suficiente para $\boldsymbol{\theta}_2$ y parcialmente auxiliar para $\boldsymbol{\theta}_1$. Además, los vectores de parámetros $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ deben variar independientemente, o sea, el campo de variación de $\boldsymbol{\theta}_1$ debe ser igual para cada $\boldsymbol{\theta}_2$ fijo, y viceversa. Esta propiedad puede encontrarse en la familia exponencial de rango completo (ver Lindsey 1996).

Por tanto, si las anteriores condiciones se satisfacen, el estadístico $\boldsymbol{T}^*$ separa la función de verosimilitud de la forma

$$L(\boldsymbol{\theta}) = f(\boldsymbol{t}^* \mid \boldsymbol{\theta}_1) f(\boldsymbol{x} \mid \boldsymbol{t}^*, \boldsymbol{\theta}_2) = L_1(\boldsymbol{\theta}_1) L_2(\boldsymbol{\theta}_2) \tag{30}$$

y usando el estadístico $\boldsymbol{U}^*$, obtenemos

$$L(\boldsymbol{\theta}) = f(\boldsymbol{x} \mid \boldsymbol{u}^*, \boldsymbol{\theta}_1) f(\boldsymbol{u}^* \mid \boldsymbol{\theta}_2) = L_1(\boldsymbol{\theta}_1) L_2(\boldsymbol{\theta}_2) \tag{31}$$

[3]El Teorema de Basu dice que dos estadísticos U y T son independientes si U es suficiente y completo para θ y T es auxiliar para θ.

Ejemplo 8. Análisis de supervivencia (continuación). Considere el ejemplo 2. Supóngase también que $T \sim \exp(\lambda)$ y $C \sim \exp(\kappa\lambda)$. En este caso, $\boldsymbol{\theta} = (\lambda, \kappa)$, siendo λ el parámetro de interés y κ el parámetro de perturbación. Haciendo $A = \sum_i \delta_i z_i$, $B = \sum_i (1-\delta_i) z_i$ y $d = \sum_i \delta_i$, se puede mostrar que $\lambda A \mid d \sim gamma(d, 1)$, $\lambda B \mid d \sim gamma(d, \kappa)$ y $d \sim Bin\,(n, 1/(1+\kappa))$. Por consiguiente, la distribución conjunta de $W = A/B$ y d no depende de λ. La función de verosimilitud está dada por

$$\begin{aligned} L(\lambda, \kappa) &= \lambda^n \kappa^{n-d} \exp\{\lambda(1+\kappa)\Sigma_i z_i\} \\ &= \lambda^n \kappa^{n-d} \exp\{\lambda(1+\kappa)(A+B)\} \\ &= \lambda^n \kappa^{n-d} \exp\{\lambda(1+\kappa)B(1+W)\} \end{aligned} \tag{32}$$

Por el criterio de la factorización, se nota que (B, W, d) es suficiente para (λ, κ). Haciendo $U^* = (W, d)$ tenemos que $B \mid U^* \sim gamma(d, \lambda W)$. Así, se pueden hacer inferencias sobre λ usando solo la distribución de $B \mid U^*$. El estimador de máxima verosimilitud de λ usando esta distribución está dado por $\widehat{\lambda} = d/(BW) = d/A$.

Definición 8. Un estadístico $\boldsymbol{T^*}$ que sea parcialmente suficiente para $\boldsymbol{\theta}_1$, y parcialmente auxiliar para $\boldsymbol{\theta}_2$ y cuyos parámetros sean ortogonales, es llamado "corte propio" (*proper cut*) por Lindsey (1996); también se denomina estadístico que define un corte de Bardorff-Nielsen en el modelo $\mathcal{F}$.

Si $\boldsymbol{T^*}$ define un corte de Bardorff-Nielsen para $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, entonces $\boldsymbol{T^*}$ es un estadístico *S-suficiente* para $\boldsymbol{\theta}_1$ y *S-auxiliar* para $\boldsymbol{\theta}_2$. Además, la función de verosimilitud es ortogonal y siempre puede ser escrita de la forma

$$L(\boldsymbol{\theta}) = f(\boldsymbol{t^*} \mid \boldsymbol{\theta}_1) f(\boldsymbol{x} \mid \boldsymbol{t^*}, \boldsymbol{\theta}_2) \tag{33}$$

En este caso no tendremos pérdida de información al usar el modelo $L_1(\boldsymbol{\theta}_1)$ dado en (30) o (31).

Es raro encontrar estadísticos $\boldsymbol{T^*}$ y $\boldsymbol{U^*}$ con estas propiedades. Jorgensen (1993) usó la definición de modelo saturado para introducir nuevos conceptos de suficiencia e información auxiliar, con el objetivo de reducir al máximo el modelo. El concepto de modelo saturado corresponde a la idea de un parámetro para cada observación, y se define a continuación.

Definición 9. Se dice que un modelo estadístico $\mathcal{F} = \{F(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ es saturado si, para todo $X \in \mathcal{X}$, el estimador de máxima verosimilitud $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(X)$ es único y función 1:1 de X.

En las definiciones 10 y 11 considere que el vector $(\boldsymbol{T}, \boldsymbol{U})$ es suficiente para $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Definición 10. Sea $\boldsymbol{T}$ un estadístico *S-auxiliar* para $\boldsymbol{\theta}_2$; entonces

$$L(\boldsymbol{\theta}) = f(\boldsymbol{t} \mid \boldsymbol{\theta}_1) f(\boldsymbol{u} \mid \boldsymbol{t}, \boldsymbol{\theta}) = L_1(\boldsymbol{\theta}_1) L_2(\boldsymbol{\theta}) \tag{34}$$

Para $\boldsymbol{\theta}_1$ fijo, si $f(\boldsymbol{u} \mid \boldsymbol{t}, \boldsymbol{\theta})$ es un modelo saturado, entonces se dice que el estadístico $\boldsymbol{T}$ es *I-suficiente* para $\boldsymbol{\theta}_1$.

Definición 11. Sea $\boldsymbol{U}$ un estadístico *S-suficiente* para $\boldsymbol{\theta}_2$; entonces

$$L(\boldsymbol{\theta}) = f(\boldsymbol{t} \mid \boldsymbol{u}, \boldsymbol{\theta}_1) f(\boldsymbol{u} \mid \boldsymbol{\theta}) = L_1(\boldsymbol{\theta}_1) L_2(\boldsymbol{\theta}) \tag{35}$$

para $\boldsymbol{\theta}_1$ fijo, si $f(\boldsymbol{u} \mid \boldsymbol{\theta})$ es un modelo saturado, entonces se dice que el estadístico $\boldsymbol{U}$ es *I-auxiliar* para $\boldsymbol{\theta}_1$.

En la definición 10, toda la información relevante sobre $\boldsymbol{\theta}_1$ está contenida en el primer término $f(\boldsymbol{t} \mid \boldsymbol{\theta}_1)$. En la definición 11, la idea es contraria: no existe información relevante sobre $\boldsymbol{\theta}_1$ en el segundo término $f(\boldsymbol{u} \mid \boldsymbol{\theta})$. Además, en la definición 10, cuando $\boldsymbol{\theta}_1$ está fijo, la saturación del modelo $L_2(\boldsymbol{\theta}) = f(\boldsymbol{u} \mid \boldsymbol{t}, \boldsymbol{\theta})$ no garantiza que el estadístico $\boldsymbol{U}$ sea totalmente no informativo para diferentes valores de $\boldsymbol{\theta}_1$.

Sea $f\left(\boldsymbol{u} \mid \boldsymbol{t}, \boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2\right)$ la función de verosimilitud $f(\boldsymbol{u} \mid \boldsymbol{t}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ cuando substituimos $\boldsymbol{\theta}_2$ por su EMV $\widehat{\boldsymbol{\theta}}_2$. Pace & Salvan (1997) argumentan que si $f\left(\boldsymbol{u} \mid \boldsymbol{t}, \boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2\right)$ fuera no identificable o no existiera el EMV para $\boldsymbol{\theta}_1$, entonces $L_2(\boldsymbol{\theta})$ podría ser ignorado en la estimación de $\boldsymbol{\theta}_1$. Este concepto de falta de información se denomina *ausencia de información parcial en el sentido extendido.*

Ejemplo 9. Distribución exponencial truncada (continuación). Considere el ejemplo 6, donde $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución $\mathrm{E}(\alpha, \beta)$. El parámetro de escala β es el parámetro de interés y α es el parámetro de perturbación.

Por el ejemplo 6, tenemos que el vector de estadísticos $\boldsymbol{V}^* = (U, T)$, con $U = X_{(1)}$ y $T = 2n \sum_i \left(X_i - X_{(1)}\right)$, es suficiente para (α, β); además, $U \sim E(\alpha, n/\beta)$ y $T \sim \beta\chi^2\,(2n)$ son independientes. El estadístico T es *S-auxiliar* para α, pues la distribución de T no depende de α y la distribución $U \mid T = t$ es igual a la de la distribución marginal de U por la independencia. Fijando el valor de β en la distribución de $U \mid T = t$, el EMV de α es $\widehat{\alpha} = U$; luego el modelo es saturado, y consecuentemente T es *I-suficiente* para β. Así, toda la información relevante que la muestra tiene sobre β está contenida en la distribución marginal de T. Entonces, el factor ignorado en la función de verosimilitud será $L_2(\theta) = f(u \mid t, \theta)$.

Si sustituimos α por su estimador de máxima verosimilitud en L_2, tenemos la nueva función de verosimilitud dada por

$$f(u \mid t, \beta, \widehat{\alpha}) = \frac{n}{b} \exp\left\{ -\frac{(u-u)}{\beta} \right\} = \frac{n}{\beta} \tag{36}$$

Como L_2 es una función decreciente en β, tenemos que su EMV no está definido, y dado $T = t$, la distribución de U no es informativa en la estimación de β en el sentido extendido.

Ejemplo 10. El test exacto de Fisher es una de las pruebas más famosos para verificar si existe asociación entre variables categóricas, este test se deriva de la distribución binomial como veremos a continuación.

En la tabla 1 presentamos una tabla de contingencia 2×2, básicamente, una tabla de contingencia es la representación de las frecuencias conjuntas entre dos o más características que deseamos estudiar. Sea A y B la representación de dos eventos independientes de interés, A' y B' sus respectivos eventos complementarios; tal que a es el número de ocurrencias del evento $A \cap B$, b es el número de ocurrencias del evento $A' \cap B$, c es el número de ocurrencias del evento $A \cap B'$ y d es el número de ocurrencias del evento $A' \cap B'$.

Tabla 1: Tabla de Contingencia 2×2.

	A	A'	Total
B	a	b	m
B'	c	d	$m-n$
Total	t	$n-t$	n

Si n y m son fijos, entonces $a \sim Bin(m, p_1)$ y $c \sim Bin(m-n, p_2)$, donde $Bin(n, p)$ denota la distribución binomial con parámetros n y p. La función de verosimilitud puede ser escrita como

$$f(a, c \mid p_1, p_2) = \binom{m}{a}\binom{m-n}{c} p_1^a (1-p_1)^b p_2^c (1-p_2)^d \tag{37}$$

Suponga que estamos interesados en estimar la razón de ventajas (*odds ratio*) $\theta_1 = \dfrac{p_1(1-p_2)}{(1-p_1)p_2}$. Así, haciendo $\theta_2 = p_2$, la función de verosimilitud puede ser reescrita de la forma

$$f(a, t \mid \theta_1, \theta_2) = \binom{m}{a}\binom{m-n}{t-a} \frac{\theta_1^a \theta_2^\top (1-\theta_2)^{n-t}}{(1-\theta_2+\theta_1\theta_2)^m} \tag{38}$$

La distribución condicional de a dado t está dada por

$$f(a \mid t, \theta_1) = \frac{\binom{m}{a}\binom{m-n}{t-a}\theta_1^a}{\sum_k \binom{m}{k}\binom{m-n}{t-k}\theta_1^k} \tag{39}$$

La distribución de $a \mid t$ se utiliza para hacer pruebas de asociación entre A y B. La distribución de t está dada por

$$f(t \mid \theta_1, \theta_2) = \frac{\theta_1^a \theta_2^\top (1-\theta_2)^{n-t}}{(1-\theta_2+\theta_1\theta_2)^m} \sum_k \binom{m}{k}\binom{m-n}{t-k}\theta_1^k \tag{40}$$

Si $\theta_1 = 1$, la derivada del logaritmo de $f(t \mid \theta_1, \theta_2)$ es

$$\frac{\partial \log f(t \mid \theta_1, \theta_2)}{\partial \theta_2} = \frac{t}{\theta_2} - \frac{n-t}{1-\theta_2} \tag{41}$$

y el estimador de máxima verosimilitud de θ_2 es t/n. Así, $f(t \mid \theta_1, \theta_2)$ es un modelo saturado y, por tanto, el estadístico T es *I-auxiliar* para θ_1.

4. Funciones de seudoverosimilitudes

Cuando existen modelos donde no es posible aplicar las técnicas vistas en las secciones anteriores o la información contenida en L_2 no puede ser ignorada debido a su importancia en la inferencia, es indispensable definir otras alternativas. Las funciones de seudoverosimilitudes pueden utilizarse como una aproximación a la clase de funciones de verosimilitudes genuinas. Las funciones de verosimilitudes canónica, perfilada, perfilada corregida, perfilada modificada, predictivas bayesianas (no bayesianas) y cuasiverosimilitud son algunos ejemplos de funciones de seudoverosimilitudes. Para ilustrar este tipo de técnicas de reducción de modelos, en esta sección presentamos ejemplos de las funciones de verosimilitudes canónica y perfilada.

4.1. Función de verosimilitud canónica

Sea $L(\boldsymbol{\theta})$ una función de verosimilitud de dos parámetros, con $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \subset I\!R^2$, es decir, los parámetros de interés θ_1 y de perturbación θ_2 son escalares. Ya vimos que si $L(\boldsymbol{\theta}) = L_1(\theta_1)L_2(\theta_2)$, entonces $L(\boldsymbol{\theta})$ será ortogonal en relación con la partición de interés y la inferencia sobre θ_1 estará basada integralmente en $L_1(\theta_1)$. Dado que no siempre es posible obtener con exactitud esta separación, Hinde & Aitkin (1987) propusieron realizar una aproximación a esta factorización. La idea central es considerar una factorización aproximada para la función de verosimilitud original, esto es,

$$L(\theta_1, \theta_2) \approx L_1(\theta_1)L_2(\theta_2) \tag{42}$$

donde la distancia entre las funciones de verosimilitudes original y aproximada es la menor posible. Las funciones $L_1(\theta_1)$ y $L_2(\theta_2)$ se obtienen por una descomposición de autofunciones de $L(\boldsymbol{\theta})$. Estas funciones se llaman verosimilitudes canónicas para los parámetros θ_1 y θ_2, respectivamente.

Para determinar las funciones $L_1(\theta_1)$ y $L_2(\theta_2)$, Hinde y Aitkin consideraron tres casos, dependiendo de la naturaleza del espacio paramétrico: i) ambos discretos; ii) uno discreto y el otro continuo y iii) ambos continuos. La idea principal de los autores es integrar (o sumar) $L(\theta_1, \theta_2)L_2(\theta_2)$ con respecto al parámetro de perturbación θ_2; el resultado es la función de verosimilitud canónica para el parámetro de interés θ_1. A continuación se presenta un ejemplo clásico para ilustrar esta técnica.

Ejemplo 11. Distribución normal. Sea X una variable aleatoria con distribución $N(\mu, 1)$. Defina $\theta_1 = |\mu|$ y $\theta_2 = \text{signo}(\mu)$, esto es, $\theta_1 \in I\!R^+$ y $\theta_2 \in \{-1, 1\}$. Suponga que estamos interesados en hacer inferencias sobre $\theta_1 = |\mu|$. La función de verosimilitud genuina es

$$L(\theta_1, \theta_2 \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(x - \theta_1\theta_2\right)^2\right\} \tag{43}$$

Sean $T = |X|$ y $S = \text{signo}(X)$, entonces

$$\begin{aligned} L(\theta_1, \theta_2 \mid t, s) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(ts - \theta_1\theta_2\right)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\Big(t^2 - 2ts\theta_1\theta_2 + \theta_2^2\Big)\right\} \end{aligned} \tag{44}$$

luego, T y S son conjuntamente suficientes para θ_1 y θ_2. Note que la función de verosimilitud no es ortogonal.

Siguiendo la idea de Hinde y Aitkin, se debe minimizar

$$\sum_{j=1}^{2} \int_{\boldsymbol{\Theta}_1} \Big[L\big(\theta_1, \theta_{2j}\big) - L_1(\theta_1) L_2\big(\theta_{2j}\big)\Big]^2 d\theta_1 \tag{45}$$

cuyas soluciones son

$$L(\theta_1, -1 \mid x) L_2(-1 \mid x) + L(\theta_1, 1 \mid x) L_2(1 \mid x) = \lambda L_1(\theta_1 | x) \tag{46}$$

$$\int_{\boldsymbol{\Theta}_1} L(\theta_1, 1 \mid x) L_1(\theta_1 \mid x)\, d\theta_1 = \lambda L_2(1 \mid x) \quad \text{y} \tag{47}$$

$$\int_{\boldsymbol{\Theta}_1} L(\theta_1, -1 \mid x) L_1(\theta_1 \mid x)\, d\theta_1 = \lambda L_2(-1 \mid x) \tag{48}$$

En la expresión (46), $L_1(\theta_1 \mid x)$ depende de las cantidades desconocidas $L_2(1 \mid x)$ y $L_2(-1 \mid x)$. En la expresión (48) las cantidades $L_2(1 \mid x)$ y $L_2(-1 \mid x)$ dependen de $L_1(\theta_1 \mid x)$. Con el fin de simplificar la notación en este problema, considere $M_1 = L(\theta_1, 1 \mid x)$, $M_2 = L(\theta_1, -1 \mid x)$, $N_1 = L_2(1 \mid x)$ y $N_2 = L_2(-1 \mid x)$. Como N_1 y N_2 no dependen de los parámetros, (46) y (48) pueden reescribirse matricialmente de la forma:

$$\lambda^2 \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} = \boldsymbol{M} \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} \tag{49}$$

donde $\boldsymbol{M}$ es la matriz de dimensión 2×2 cuyo elemento en la posición (j, j') es dado por

$$m_{jj'} = \int_{\boldsymbol{\Theta}_1} M_j M_{j'}\, d\theta_1 \tag{50}$$

Resolviendo las integrales para cada elemento de la matriz $\boldsymbol{M}$, tenemos que

$$\boldsymbol{M} = \begin{bmatrix} \frac{1}{2\sqrt{\pi}} \Phi\big(-\sqrt{2}x\big) & \frac{1}{4\sqrt{\pi}} \exp\big\{-x^2\big\} \\ \frac{1}{4\sqrt{\pi}} \exp\big\{-x^2\big\} & \frac{1}{2\sqrt{\pi}} \Phi\big(\sqrt{2}x\big) \end{bmatrix} \tag{51}$$

donde $\Phi(\cdot)$ es la función de distribución acumulada de la distribución normal estándar. Los autovalores de la matriz $\boldsymbol{M}$ están dados por

$$\eta_1 = \frac{1 + \sqrt{\big(2\Phi\big(\sqrt{2}x\big) - 1\big)^2 + \exp\big\{-2x^2\big\}}}{2} \tag{52}$$

y

$$\eta_2 = \frac{1 - \sqrt{\left(2\Phi\left(\sqrt{2}x\right) - 1\right)^2 + \exp\left\{-2x^2\right\}}}{2} \tag{53}$$

Se comprueba fácilmente que la suma de los autovalores η_1 y η_2 es 1. Ahora, dado que la solución de la ecuación (46) es $\lambda L_1(\theta_1)$, y en la ecuación (49) tenemos la relación $\lambda = \sqrt{\eta}$, entonces la función de verosimilitud canónica será completamente informativa cuando $\eta_{\text{máx}} = 1$ (siendo $\eta_{\text{máx}}$ el mayor autovalor de $\boldsymbol{M}$).

El autovector asociado a $\eta_{\text{máx}}$ es $\boldsymbol{b} = (r(x), 1)$, donde $r(x) = (v^2(x) + 1)^{1/2} - v(x)$, con $v(x) = \exp\{x^2\}\left(2\Phi\left(\sqrt{2}x\right) - 1\right)$. Remplazando en la ecuación (46) con $\lambda_{\text{máx}} = \sqrt{\eta_{\text{máx}}}$, tenemos, $\sqrt{\eta_{\text{máx}}} \cdot L_1(\theta_1) = M_1 \cdot r(x) + M_2 \cdot 1$, y por consiguiente

$$L_1(\theta_1) = \frac{1}{\sqrt{\eta_{\text{máx}}}}\left[\exp\left\{-\frac{1}{2}(x+\theta_1)^2\right\} \cdot r(x) + \exp\left\{-\frac{1}{2}(x-\theta_1)^2\right\} \cdot 1\right]$$

Para cualquier valor que tome el parámetro de perturbación θ_2, la función de verosimilitud canónica $L_1(\theta_1)$ será siempre igual.

Las principales ventajas de la función de verosimilitud canónica son fundamentalmente que la inferencia sobre θ_1 se basa integralmente en $L_1(\theta_1)$; y la función de verosimilitud canónica siempre existe para modelos con dos parámetros, en contraste con las funciones verosimilitudes marginal y condicional, que generalmente no existen. Dos de las principales desventajas de este método son: tiene álgebra pesada, aun para espacios paramétricos de baja dimensión y para cada configuración de la función de verosimilitud existe una solución particular.

4.2. Función de verosimilitud perfilada

Inferir sobre el parámetro de interés a partir de la función de verosimilitud marginal o condicional es muy adecuado, porque estas son verosimilitudes genuinas; el problema es que no siempre es posible su construcción. Una solución es sustituir en la verosimilitud original el vector de parámetros de perturbación por una estimativa consistente; la función resultante se conoce como función de verosimilitud perfilada.

Formalmente, sea $(X_1, \ldots, X_n)$ una muestra aleatoria de la variable X con distribución de probabilidad en la familia $\mathcal{F} = \left\{F(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top\right)^\top \in \boldsymbol{\Theta}\right\}$, siendo $\boldsymbol{\theta}_1$ el vector de parámetros de interés y $\boldsymbol{\theta}_2$ el vector de parámetros de perturbación. Sea $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\theta}}_1^\top, \widehat{\boldsymbol{\theta}}_2^\top\right)^\top$ el estimador de máxima verosimilitud del vector $\boldsymbol{\theta}$ completo, y $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\theta}_j)$ el estimador de máxima verosimilitud de $\boldsymbol{\theta}_i$ cuando $\boldsymbol{\theta}_j$ está fijo, para $i, j = 1, 2$. La función de verosimilitud perfilada es definida por

$$L_p\left(\boldsymbol{\theta}_1\right) = L\left(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\right) \tag{54}$$

donde $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ denota la función de verosimilitud genuina y $\widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$ denota el estimador de máxima verosimilitud de $\boldsymbol{\theta}_2$ para $\boldsymbol{\theta}_1$ fijo.

La expresión (54) sugiere un procedimiento de maximización en dos etapas. La primera etapa consiste en calcular el valor $\widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$ que maximice $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ con respecto a $\boldsymbol{\theta}_2$, suponiendo $\boldsymbol{\theta}_1$ constante. La segunda etapa busca el valor $\boldsymbol{\theta}_1$ que maximice $L_p(\boldsymbol{\theta}_1)$.

La inferencia aproximada sobre $\boldsymbol{\theta}_1$ se hace tratando $L_p(\boldsymbol{\theta}_1)$ como una función de verosimilitud genuina basada en un modelo solamente con el parámetro $\boldsymbol{\theta}_1$. Usar la función de verosimilitud perfilada es semejante a tratar el parámetro de perturbación como si fuese conocido. Tal procedimiento puede conducir a algunos problemas; por ejemplo, inconsistencia e ineficiencia de los estimadores de los parámetros de interés.

Veamos dos ejemplos.

Ejemplo 12. Distribución normal. Suponga que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución normal, $N(\mu, \sigma^2)$. Luego, su función de verosimilitud genuina es

$$L\left(\mu, \sigma^2; x\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \tag{55}$$

Dado μ, el EMV de σ^2 es $\dfrac{\Sigma_i(x_i-\mu)^2}{n}$. Y dada σ^2, el EMV de μ es $\overline{x}$. Por tanto, la función de verosimilitud perfilada de μ es

$$L_p(\mu; x) = \left\{\frac{\Sigma_i(x_i-\mu)^2}{n} 2e\pi\right\}^{-n/2} \tag{56}$$

y la función de verosimilitud perfilada de σ^2 es

$$L_p\left(\sigma^2; x\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{\Sigma_i(x_i-\overline{x})^2}{2\sigma^2}\right\} \tag{57}$$

En este caso, considerando las funciones de verosimilitudes perfiladas; los EMV coinciden con los estimadores usuales cálculados a partir de la función de verosimilitud genuina.

La función de verosimilitud perfilada también se utiliza bastante en modelos con errores en las variables, donde el número de parámetros de perturbación crece con el tamaño de la muestra (parámetros incidentales). Presentamos un ejemplo de este modelo.

Ejemplo 13. Modelo con errores en las variables. Considere $(Y_1, X_1), \ldots, (Y_n, X_n)$ una muestra aleatoria cuya relación entre Y_i y X_i es dada por $Y_i = \alpha + \beta x_i + e_i$ y $X_i = x_i + u_i$, siendo $e_i \sim N(0, \lambda)$ y $u_i \sim N(0, \kappa)$ variables aleatorias independientes para todo $i = 1, \ldots, n$. El logaritmo de la función de verosimilitud (ℓ) para este

modelo es ilimitado, y es necesario hacer algunas suposiciones extras para limitarla. Suponiendo que la razón $\sigma = \lambda/\kappa$ es conocida, el logaritmo de la función de verosimilitud se torna limitado y dado por

$$\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \log L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \tag{58}$$

siendo

$$\ell_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \propto -\frac{1}{2}\log(\sigma\kappa) - \frac{1}{2}\log(\kappa) - \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma\kappa} - \frac{(X_i - x_i)^2}{2\kappa} \tag{59}$$

Aquí, $\boldsymbol{\theta}_1 = (\alpha, \beta, \kappa)^\top$ es el vector de parámetros de interés y $\boldsymbol{\theta}_2 = (x_1, \ldots, x_n)^\top$ es el vector de parámetros incidentales (de perturbación). El estimador de máxima verosimilitud para x_i está dado por

$$\widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1) = \widehat{x}_i = \frac{\kappa\beta(Y_i - \alpha) + \sigma\kappa X_i}{\beta^2\kappa + \sigma\kappa} \tag{60}$$

Sustituyendo (60) en la log-verosimilitud genuina (58), tenemos

$$\ell_p(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \ell_{pi}\Big(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\Big) \tag{61}$$

siendo

$$\ell_{pi}\Big(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)\Big) \propto -\frac{1}{2}\log(\sigma\kappa) - \frac{1}{2}\log(\kappa) - \frac{\big(y_i - \alpha - \beta\widehat{x}_i\big)^2}{2\sigma\kappa} - \frac{\big(X_i - \widehat{x}_i\big)^2}{2\kappa} \tag{62}$$

Los EMV para α, β y κ, cuando σ es conocida, se obtienen igualando a cero las derivadas de ℓ_p en relación con los parámetros de interés. Los estimadores son dados por

$$\begin{aligned}
\widehat{\alpha} &= \overline{Y} - \widehat{\beta}\overline{X} \\
\widehat{\beta} &= \frac{S_Y - \sigma S_X^2 + \sqrt{\big(S_Y^2 - \sigma S_X^2\big)^2 - 4\sigma S_{YX}^2}}{2S_{YX}} \\
\widehat{\kappa} &= \sum_{i=1}^{n} \frac{\Big(Y_i - \widehat{\alpha} - \widehat{\beta}X_i\Big)^2}{2n\Big(\widehat{\beta}^2 + \sigma\Big)}
\end{aligned} \tag{63}$$

siendo,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(X_i - \overline{X}\Big)^2$$
$$S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \overline{Y}\Big)^2$$
$$S_{YX} = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \overline{Y}\Big)\Big(X_i - \overline{X}\Big)$$

Patefield (1978) mostró que el EMV $\widehat{\kappa}$ converge en probabilidad para $\kappa/2$. En este caso, el estimador consistente es dado por $2\widehat{\kappa}$. Mak (1982) estudió las propiedades de los estimadores en presencia de parámetros incidentales. El autor demostró que el estimador del vector de parámetros de interés existe y converge para una distribución normal multivariada con media igual al vector de parámetros de interés, si se satisfacen las condiciones

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} x_i = \lim_{n\to\infty}\overline{x}_n < \infty, \qquad \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 < \infty$$

y

$$\lim_{n\to\infty}\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}|x_i|^{2+\delta} = 0$$

para todo $\delta > 0$.

Mak (1982) también demostró que la matriz de covarianzas asintótica de los parámetros de interés no es la inversa de la matriz de información de Fisher esperada y debe ser sustituida por la siguiente matriz

$$\mathrm{Cov}\Big(\widehat{\boldsymbol{\theta}}_1\Big) = \frac{1}{n}\boldsymbol{A}(\boldsymbol{\theta}_1)^{-1}\boldsymbol{V}(\boldsymbol{\theta}_1)\boldsymbol{A}(\boldsymbol{\theta}_1)^{-1}$$

siendo,

$$\boldsymbol{V}(\boldsymbol{\theta}_1) = \frac{1}{n}\mathrm{Var}\left(\frac{\partial \ell_p}{\partial \boldsymbol{\theta}_1}\right) \quad \text{y} \quad \boldsymbol{A}(\boldsymbol{\theta}_1) = \frac{1}{n}\mathrm{E}\left(\frac{\partial^2 \ell_p}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^t}\right)$$

Para terminar, resaltamos que las principales ventajas de usar la función de verosimilitud perfilada cuando el número de parámetros de perturbación no crece con el tamaño de la muestra son:

- La función de verosimilitud perfilada siempre existe.

- La función de verosimilitud perfilada no depende del parámetro de perturbación.
- La matriz de información observada perfilada de $\boldsymbol{\theta}_1$ se define de forma análoga a la información observada de $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.
- El estadístico de la razón de verosimilitudes tiene distribución asintótica igual a la basada en la razón de verosimilitudes genuinas, esto es,

$$W_p(\boldsymbol{\theta}_1) = 2\left\{ L_p\left(\widehat{\boldsymbol{\theta}_1}\right) - L_p(\boldsymbol{\theta}_1) \right\} \xrightarrow{\mathcal{D}} \chi^2(p_1)$$

  siendo que $\xrightarrow{\mathcal{D}}$ significa convergencia en distribución y p_1 la dimensión de $\boldsymbol{\theta}_1$.

Las demostraciones de estas propiedades están en Cordeiro (1992).

La principal desventaja es que la función de verosimilitud perfilada, generalmente, no presenta todas las propiedades de una función de verosimilitud genuina. Por ejemplo, la esperanza de la función escore perfilada generalmente es diferente de cero. Por tanto, los estimadores obtenidos vía función de verosimilitud perfilada pueden no ser consistentes. Por tanto, es necesario hacer ajustes en la verosimilitud perfilada para minimizar estos problemas. En la literatura, existen varias modificaciones para la función de verosimilitud perfilada propuestas por diversos autores; ver Barndorff-Nielsen (1983), Barndorff-Nielsen (1991), Cox & Reid (1987), Cox & Reid (1992) y McCullagh & Tibshirani (1990). Estas modificaciones consisten en la incorporación de un término en la verosimilitud perfilada anterior al proceso de estimación que tiene por efecto disminuir el sesgo de la función escore y de la información de Fisher esperada.

5. Conclusiones

En este trabajo presentamos y discutimos algunos métodos de estimación en presencia de parámetros de perturbación. Como existen diversas metodologías en la literatura para tratar tales modelos, enfocamos nuestra atención en técnicas de reducción de modelos a través de estadísticos con propiedades óptimas o a través de funciones de verosimilitudes canónicas y perfiladas. Ilustramos y analizamos algunos conceptos sobre ausencia de información presente en la muestra con relación a los parámetros de perturbación en ejemplos simples y recientemente discutidos en la literatura. A los interesados, dejamos las referencias para que sean consultadas posteriormente.

Agradecimentos

Durante el desarrollo de este trabajo los autores recibieron apoyo financiero del Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), de la

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil, y de la Universidad Industrial de Santander, Colombia. Los autores también expresan sus agradecimientos al profesor Dr. Heleno Bolfarine (IME-USP) por las sugerencias metodológicas, a la profesora Dra. Silvia Ferrari (IME-USP) por la motivación para escribir este trabajo, al profesor Dr. Bernardo Mayorga (UIS) por la revisión de estilo y a los dos árbitros por las valiosas sugerencias dadas para mejorar el presente documento.

[Recibido: junio de 2008 — Aceptado: marzo de 2009]

Referencias

Azzalini, A. (1985), 'A Class of Distributions which Includes the Normal Ones', *Scandinavian Journal of Statistics* **12**, 171–178.

Barndorff-Nielsen, O. (1983), 'On a Formula for the Distribution of the Maximum Likelihood Estimator', *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O. (1991), *Likelihood Theory*, Chapman and Hall, London, England.

Cordeiro, G. (1992), Introdução à Teoria de Verossimilhança, *in* '10 Simpósio Nacional de Probabilidade e Estatística', Rio de Janeiro, Brazil.

Cox, D. R. & Reid, N. (1987), 'Parameter Orthogonality and Approximate Conditional Inference (with Discussion)', *Journal The Royal Statistical Society: Series B* **49**, 1–39.

Cox, D. R. & Reid, N. (1992), 'A Note on the Difference Between Profile and Modified Profile Likelihood', *Biometrika* **79**, 408–411.

Durrans, S. R. (1992), 'Distributions of Fractional Order Statistics in Hydrology', *Water Resources Research* **28**, 1649–1655.

Fuller, W. A. (1987), *Measurement Error Models*, Wiley, New York, United States.

Halmos, P. R. & Savage, L. J. (1949), 'Application of the Radon–Nikodym Theorem to the Theory of Sufficient Statistics', *Annals of Mathematics Statistics* **20**, 225–241.

Hinde, J. & Aitkin, M. (1987), 'Canonical Likelihoods: A New Likelihood Treatment of Nuisance Parameters', *Biometrika* **74**, 45–58.

Jones, M. C. (2004), 'Families of Distributions Arising from Distributions of Order Statistics', *Test* **13**, 1–43.

Jorgensen, B. (1993), 'A Review of Conditional Inference: Is there a Universal Definition of Noinformation?', *Bulletin of International Statistical Institute* **55,2**, 323–340.

Lehmann, E. L. & Casella, G. (1998), *Theory of Point Estimation*, Springer-Verlag, New York, United States.

Lindsey, J. K. (1996), *Parametric Statistical Inference*, Clarendon Press, Oxford, England.

Mak, T. K. (1982), 'Estimation in the Presence of Incidental Parameters', *The Canadian Journal of Statistics, La Revue Canadienne de Statistique* **10-2**, 121–132.

McCullagh, P. & Tibshirani, R. (1990), 'A Simple Method for the Adjustment of Profile Likelihoods', *Journal The Royal Statistical Society: Series B* **52**, 325–344.

Neyman, J. & Scott, E. L. (1948), 'Consistent Estimates Based on Partially Consistent Observations', *Econometrica* **16-1**, 1–32.

Pace, L. & Salvan, A. (1997), *Principles of Statistical Inference*, World Scientific, Singapore, Singapore.

Patefield, W. M. (1978), 'The Unreplicated Ultrastructural Relation: Large Sample Properties', *Biometrika* **65**, 535–540.