UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Unidad Académica de los Ciclos Profesional y de Posgrado del CCH
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

CONTRASTE BAYESIANO DE HIPOTESIS PARAMETRICAS

T E S I S

Que para obtener el grado de:
Maestro en Estadística e Investigación de Operaciones
Presenta el Actuario
Eduardo A. Gutiérrez Peña

México, D. F. Agosto 1991

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

CONTRASTE BAYESIANO DE HIPOTESIS PARAMETRICAS

PREFACIO

Una de las áreas más importantes de la Inferencia Estadística se refiere al problema de contraste de hipótesis. Dicho problema ha sido ampliamente estudiado desde una perspectiva frecuentista, a raíz del importante trabajo desarrollado por Neyman y Pearson en los años 20. Por otro lado, aunque desde el punto de vista Bayesiano se cuenta con una solución general al problema, obtenida al plantear el problema de contraste de hipótesis como un Problema de Decisión, las propiedades de los procedimientos generados de esta manera dependen fundamentalmente de la asignación de una función de utilidad y de una distribución inicial. En los procedimientos Bayesianos reportados en la literatura, esas asignaciones se realizan en función del tipo de hipótesis que se desea contrastar.

En este trabajo se presenta un procedimiento Bayesiano que permite dar un tratamiento unificado al problema de contraste de hipótesis paramétricas dentro del contexto de la Teoría de Decisiones. Dicho procedimiento de basa en la consideración de una familia particular de funciones de utilidad, definida en términos de una medida de discrepancia entre modelos.

En el Capítulo 1 se plantea el problema general de contraste de hipótesis paramétricas y se presenta una breve revisión de los procedimientos clásicos. Asimismo, se exponen algunos de los conceptos básicos de la Teoría Bayesiana de Toma de Decisiones. Por otro lado, en el Capítulo 2 se revisan de manera breve los procedimientos Bayesianos más comunes para contrastar hipótesis paramétricas, tanto a nivel inferencial como en un contexto de Teoría de Decisiones. En el Capítulo 3 se propone un procedimiento Bayesiano alternativo, el cual, por un lado, elimina algunas de las desventajas de los procedimientos Bayesianos discutidos en el Capítulo 2 y, por otro lado, permite reproducir la forma de las soluciones clásicas en algunos casos. Adicionalmente, se trata con cierto detalle el caso de las familias exponenciales regulares, ampliamente utilizadas en las teoría estadística. El Capítulo 4 contiene, a manera de ilustración, la solución bayesiana al problema de contraste de la hipótesis lineal general en el modelo de regresión lineal múltiple. Finalmente, en el Capítulo 5 se presentan algunas conclusiones y comentarios finales.

Deseo agradecer profundamente al M. en C. Raúl Rueda por su paciencia y dedicación en la dirección de esta tesis. Asimismo, agradezco al Dr. Manuel Mendoza por sus valiosos comentarios. Ambos son, en gran medida, responsables por los aciertos que pueda tener este trabajo.

INDICE

1. ANTECEDENTES

1.1 El Problema General de Contraste de Hipotesis Parametricas

En casi cualquier área de la investigación científica es común que, al estudiar un fenómeno determinado, surjan ciertas suposiciones o hipótesis que deben ser verificadas. Desde el punto de vista estadístico, esta situación se plantea suponiendo que existe una variable aleatoria observable que describe adecuadamente el comportamiento de la característica de interés del fenómeno bajo estudio. De esta manera, puede decirse que una hipótesis estadística es una conjetura acerca del comportamiento de una variable aleatoria. El problema consiste entonces en verificar la validez de la hipótesis con base en la información disponible.

Sea X una variable aleatoria (posiblemente m-variada) con función de densidad **p** y suponga que existe una familia paramétrica de densidades de probabilidad

$$\mathcal{F} = \left\{ p(x|\theta) \ : \ \theta \in \Theta \right\} \qquad \Theta \subseteq \mathbb{R}^d,$$

que contiene a **p**. En otras palabras, la forma de la distribución de X se supone conocida, excepto por un número finito de parámetros. Claramente, cualquier conjetura acerca de la distribución de X es una hipótesis estadística, pero en el presente contexto dicha conjetura se convierte en una proposición sobre el valor de algún parámetro y se denomina hipótesis paramétrica.

Si bien es cierto que los procedimientos de contraste no dependen de la forma específica de $p(x|\theta)$, algunas familias paramétricas permiten un analisis más detallado y la obtención de resultados más generales. En particular, una clase importante de familias paramétricas está constituída por las llamadas familias exponenciales (*e.g.* Barndorff-Nielsen, 1978), las cuales serán revisadas en el Capítulo 3. La importancia de estas familias se debe a que las distribuciones que pertenecen a ellas son, en general, matemáticamente tratables y permiten modelar una gran variedad de fenómenos, por lo que han sido ampliamente utilizadas en la teoría estadística.

Sean $\Theta_0 \subset \Theta$ y $\Theta_1 \subset \Theta$, tales que $\Theta_0 \cap \Theta_1 = \varnothing$, y sea $Z=\{X_1,X_2,\ldots,X_n\}$ un conjunto de observaciones de la variable aleatoria X. El problema general de contraste de hipótesis paramétricas consiste entonces en decidir cual de las hipótesis $H_0:\theta \in \Theta_0$ y $H_1:\theta \in \Theta_1$ puede considerarse como cierta, con base en la información contenida en Z y en la información inicial disponible sobre el valor de θ. Observe que las hipótesis H_0 y H_1 son matemáticamente equivalentes a las hipótesis $H_0':p \in \mathcal{F}_0$ y $H_1':p \in \mathcal{F}_1$, respectivamente, donde

$$\mathcal{F}_i = \left\{ p(x|\theta) : \theta \in \Theta_i \right\} \qquad (i=0,1).$$

1.2 Teoria Clasica de Contraste de Hipotesis Parametricas

En esta sección se presentará el planteamiento clásico del problema de contraste de hipótesis paramétricas, que se basa en una interpretación frecuentista de la probabilidad. Una revisión extensa de los procedimientos clásicos puede encontrarse en Lehmann (1986), así como en Cox & Hinkley (1974).

Sea $Z=\{X_1,X_2,\ldots,X_n\}$ una muestra de una variable aleatoria X con función de densidad $p(x|\theta) \in \mathcal{F}$ y suponga que $\{\Theta_0,\Theta_1\}$ es una

partición del espacio parametral Θ. Denote por d_0 y d_1 las decisiones de aceptar y rechazar a $H_0: \theta \in \Theta_0$, respectivamente. La hipótesis H_0 es simple si la familia $\mathcal{F}_0$ contiene una sola distribución; en caso contrario H_0 es compuesta.

Cualquier procedimiento de decisión es llamado una prueba de la hipótesis en cuestión. Dicho procedimiento asigna a cada valor z de Z una de las decisiones $\{d_0, d_1\}$ y, por lo tanto, divide el espacio muestral $\mathcal{Z}$ en dos regiones complementarias A y C. La hipótesis H_0 es rechazada sólo si Z cae dentro de C, por lo que esta región es llamada región de rechazo o región crítica.

Al llevar a cabo una prueba puede cometerse cualquiera de dos posibles errores: rechazar H_0 cuando es cierta (error tipo I) o bien aceptar H_0 cuando es falsa (error tipo II). En general, las consecuencias de cada uno de estos errores son diferentes. Resulta natural intentar la construcción de una prueba de tal forma que la probabilidad de cada uno de los errores sea mínima. Sin embargo, si el tamaño de muestra es fijo no es posible controlar ambas probabilidades de error simultáneamente. Una solución consiste en asignar una cota a la probabilidad de rechazar incorrectamente H_0 e intentar entonces minimizar la probabilidad de cometer el error tipo II, dada esta condición. En este caso, debe seleccionarse un número $\alpha \in (0,1)$, llamado el nivel de significancia, e imponer la siguiente condición

$$P[z \in C|\theta] \leq \alpha \quad \text{para todo } \theta \in \Theta_0. \tag{2.1}$$

Dada esta condición, se desea minimizar

$$P[z \in A|\theta] \quad \text{para todo } \theta \in \Theta_1$$

o, equivalentemente, maximizar

$$P[z \in C|\theta] \text{ para todo } \theta \in \Theta_1. \tag{2.2}$$

El valor $sup\ \{P[z \in C|\theta] : \theta \in \Theta_0\}$ es conocido como el tamaño de la prueba. La condición (2.1), por lo tanto, implica que sólo deben considerarse aquellas pruebas cuyo tamaño no exceda el nivel de significancia dado.

Para cada $\theta \in \Theta_1$, la probabilidad de rechazar H_0 dada en (2.2) es llamada la potencia de la prueba contra la alternativa θ. Si se considera como función de θ (para todo $\theta \in \Theta$), dicha probabilidad se conoce como la función potencia de la prueba y se denota generalmente por $\pi(\theta)$.

Finalmente, se dice que una prueba es insesgada si

$$sup\ \{\pi(\theta) : \theta \in \Theta_0\} \leq inf\ \{\pi(\theta) : \theta \in \Theta_1\}.$$

1.2.1 Hipotesis Simples

Si tanto H_0 como H_1 son hipótesis simples, el problema de contraste de hipótesis queda totalmente especificado por (2.1) y (2.2). En este caso, la solución está dada por el siguiente resultado:

LEMA DE NEYMAN-PEARSON: Sea $Z = \{X_1, X_2, \ldots, X_n\}$ una muestra de observaciones de una variable aleatoria X con función de densidad $p(x|\theta)$, y suponga que $\Theta = \{\theta_0, \theta_1\}$. Sean $C^* \subset Z$ y $k^* \geq 0$ tales que

(i) $P[Z \in C^*|\theta_0] = \alpha$ con $\alpha \in (0,1)$,

(ii) $\dfrac{p(z|\theta_0)}{p(z|\theta_1)} \leq k^*$ si y sólo si $z \in C^*$,

donde $p(z|\theta)$ denota a la función de verosimilitud.

Entonces la prueba con región crítica C^* es la prueba de tamaño α más potente para probar $H_0: \theta = \theta_0$ vs. $H_1: \theta = \theta_1$. ▷

El Lema de Neyman-Pearson proporciona la prueba más potente cuando se tienen hipótesis simples. Sin embargo, en ciertas situaciones también permite hallar pruebas óptimas cuando la hipótesis alternativa es compuesta.

1.2.2 Hipotesis Compuestas

Típicamente, la prueba que maximiza la potencia (2.2) contra una alternativa particular en Θ_1 depende de esta alternativa, por lo que es necesario introducir un criterio adicional cuando la hipótesis alternativa es compuesta (por ejemplo, restringiendo la clase de las pruebas posibles). Sin embargo, existen situaciones en las que la misma prueba maximiza la potencia contra todas las alternativas simples en Θ_1. Las pruebas que tienen esta propiedad son llamadas pruebas uniformemente más potentes (PUMP). En algunos casos, argumentos del tipo Neyman-Pearson son útiles para encontrar este tipo de pruebas.

En el caso más sencillo, suponga que θ es un parámetro real y que las hipótesis que se desea probar son de la forma $H_0:\theta \leq \theta_0$ vs. $H_1:\theta > \theta_0$. Se dice que una familia de densidades

$$\left\{ p(x|\theta) \ : \ \theta \in \Theta \right\} \qquad \Theta \subseteq \mathbb{R}$$

tiene un cociente de verosimilitudes monótono si existe una estadística $T=t(Z)$ tal que, para todo $\theta'<\theta''$, el cociente $\dfrac{p(z|\theta'')}{p(z|\theta')}$ es una función no-decreciente o una función no-creciente de T (*e.g.* Lehmann, 1986).

TEOREMA: Sea $Z = \{X_1, X_2, \ldots, X_n\}$ un conjunto de observaciones de una variable aleatoria X con función de densidad $p(x|\theta)$. Suponga que la familia $\left\{ p(x|\theta) \ : \ \theta \in \Theta \right\}$ tiene un cociente de verosimilitudes monótono en la estadística $T=t(Z)$ y que Θ es un intervalo.

(a) Si el cociente de verosimilitudes monótono es no-decreciente en $t(z)$ y k es tal que $P[t(Z) \geq k|\theta_0] = \alpha$, entonces la prueba cuya región crítica es

$$C = \left\{ z \in \mathcal{Z} : t(z) \geq k \right\}$$

es una prueba uniformemente más potente de tamaño α para probar $H_0:\theta \leq \theta_0$ vs. $H_1:\theta > \theta_0$.

(b) Si el cociente de verosimilitudes monótono es no-creciente en $t(z)$ y k es tal que $P[t(Z) \leq k|\theta_0] = \alpha$, entonces la prueba cuya región crítica es

$$C = \left\{ z \in \mathcal{Z} : t(z) \leq k \right\}$$

es una prueba uniformemente más potente de tamaño α para probar $H_0:\theta \leq \theta_0$ vs. $H_1:\theta > \theta_0$. ▷

Si se desea probar las hipótesis $H_0:\theta \geq \theta_0$ vs. $H_1:\theta < \theta_0$, el teorema sigue siendo válido siempre que las desigualdades que definen a la región crítica se inviertan. Cabe mencionar que, bajo ciertas condiciones, las distribuciones que pertenecen a familias exponenciales uniparametrales satisfacen las hipótesis del teorema anterior. Si esto ocurre, entonces también es posible hallar una PUMP para hipótesis de la forma $H_0:\theta \leq \theta_1$ o $\theta \geq \theta_2$ vs. $H_1:\theta_1 < \theta < \theta_2$ $(\theta_1<\theta_2)$ (Lehmann, 1986).

Por otro lado, para una gran clase de problemas no existe una prueba uniformemente más potente. En este caso es común restringir de alguna forma la clase de pruebas posibles y buscar una prueba uniformemente más potente dentro de la clase restringida. Una manera usual de hacerlo es considerar la clase de las pruebas insesgadas. En muchos problemas para los cuales no existe una PUMP, sí es posible encontrar una prueba insesgada uniformemente

más potente, como es el caso de las hipótesis

$$H_0:\theta = \theta_0 \quad \text{vs.} \quad H_1:\theta \neq \theta_0,$$

$$H_0:\theta_1 \leq \theta \leq \theta_2 \quad \text{vs.} \quad H_1:\theta < \theta_1 \quad \text{o} \quad \theta > \theta_2.$$

Las condiciones que se requieren para construir dichas pruebas se satisfacen, en particular, si $p(x|\theta)$ pertenece a una famlia exponencial (Lehmann, 1986; Cap. 4).

Un enfoque adicional, cuando no existe una prueba uniformemente más potente, consiste en considerar sólo aquellas alternativas cercanas a los valores del parámetro definidos por H_0 y maximizar la potencia localmente, obteniéndose las llamadas pruebas localmente más potentes. Para ilustrar lo anterior, suponga que se desea contrastar las hipótesis $H_0:\theta = \theta_0$ vs $H_1:\theta > \theta_0$. Si se toma una alternativa particular $\theta_1 = \theta_0 + \varepsilon$, con $\varepsilon > 0$ pequeña, se tiene entonces que

$$log\ \frac{p(z|\theta_1)}{p(z|\theta_0)} = log\ \frac{p(z|\theta_0+\varepsilon)}{p(z|\theta_0)} = \varepsilon\ \frac{\partial\ log\ p(z|\theta_0)}{\partial\ \theta_0} + o(\varepsilon),$$

bajo ciertas condiciones de regularidad. De esta manera, para valores suficientemente pequeños de ε, la región crítica dada por el Lema de Neyman-Pearson toma la forma

$$C = \left\{ z \in Z : U(z;\theta_0) \geq k \right\},$$

donde $U(z;\theta_0) = \dfrac{\partial\ log\ p(z|\theta_0)}{\partial\ \theta_0}$ y k se elige de manera que la prueba tenga tamaño α. Esta prueba resulta conveniente por dos motivos. En primer lugar, en muchos casos la distribución de $U(z;\theta_0)$ puede aproximarse adecuadamente a través de una distribución Normal $N(0,I(\theta_0))$, donde $I(\theta)$ denota a la cantidad de información de Fisher. En segundo lugar, si existe una prueba uniformemente más potente, en particular dicha prueba debe tener

potencia local máxima siempre que la familia de alternativas incluya valores locales del parámetro y, por lo tanto, debe ser idéntica a la prueba basada en $U(z;\theta_0)$.

Por otro lado, en el caso de hipótesis de la forma $H_0:\theta = \theta_0$ vs $H_1:\theta \neq \theta_0$, el argumento puede extenderse para obtener pruebas insesgadas localmente más potentes. Sin embargo, estas pruebas presentan algunas dificultades cuando las regiones críticas que se obtienen son utilizadas para construir pruebas de significancia (Cox & Hinkley, 1974).

Finalmente, un método relativamente simple que ha llevado a pruebas razonables en una gran variedad de problemas es el método generalizado de cociente de verosimilitudes, que en cierto sentido extiende el criterio de Neyman y Pearson. Considere el problema general de contraste de hipótesis paramétricas descrito al principio de esta sección. La prueba generada por el procedimiento generalizado de cociente de verosimilitudes utiliza la región crítica C dada por

$$C = \left\{ z \in \mathcal{Z} : \frac{sup\ \{p(z|\theta) : \theta \in \Theta_0\}}{sup\ \{p(z|\theta) : \theta \in \Theta_1\}} \leq k \right\},$$

donde k es una constante elegida de manera que la prueba tenga tamaño α.

Aunque el principio generalizado de cociente de verosimilitudes no se basa en consideraciones de optimalidad claramente definidas, sí ha tenido cierto éxito al producir procedimientos satisfactorios en muchos problemas específicos. En particular, si H_0 es simple y H_1 es compuesta entonces el método generalizado de cociente de verosimilitudes produce la prueba uniformemente más potente, siempre que ésta exista. En muchos casos, las pruebas basadas en el método generalizado de cociente de verosimilitudes tienen propiedades asintóticas óptimas (*e.g.* Cox & Hinkley, 1974).

1.3 Conceptos Basicos de la Teoria Bayesiana de Toma de Decisiones

La teoría bayesiana de toma de decisiones es una teoría formal basada en una serie de axiomas o principios de coherencia. El propósito de esta sección es presentar sólo los conceptos básicos que permitan ubicar al problema de contraste de hipótesis paramétricas como un problema de decisión en ambiente de incertidumbre. Una presentación formal puede encontrarse, por ejemplo, en el trabajo de Bernardo, Ferrándiz & Smith (1985).

Todo problema de decisión en ambiente de incertidumbre involucra la elección de un curso de acción cuyas consecuencias son desconocidas, pues dependen de la ocurrencia de algún suceso incierto. De esta manera, resulta de interés considerar la elección entre distintos cursos de acción cuando:

a) las consecuencias de cualquiera de ellos dependen del estado de la naturaleza;

b) el verdadero estado de la naturaleza es aún desconocido;

c) es posible, a un cierto costo, obtener información acerca de dicho estado.

Se supone que el decisor ha eliminado ya aquellos posibles cursos de acción que no merecen mayor consideración y, por lo tanto, ha reducido su problema a la elección de una acción entre un conjunto de alternativas relevantes bien definidas. Se supone, además, que el decisor desea escoger entre estas alternativas de tal manera que la elección sea consistente con sus preferencias personales respecto a las consecuencias, así como con su juicio personal acerca del verdadero, pero desconocido, estado de la naturaleza.

La información básica que el decisor debe ser capaz de especificar, y que define su problema de decisión, es la siguiente:

$\mathbb{D}$. El espacio de acciones: el decisor desea seleccionar una acción d perteneciente a un conjunto $\mathbb{D}$ de acciones potenciales bien definidas.

Θ. El espacio de estados de la naturaleza: el decisor considera que las consecuencias de adoptar alguna acción particular dependen del estado de la naturaleza, que no puede predecir con certeza. Los eventos relevantes están contenidos en una σ-álgebra $\mathcal{A}$ definida sobre Θ.

$\mathbb{C}$. El conjunto de consecuencias: este conjunto debe contener a todas las posibles consecuencias que resulten de elegir una acción d cuando el verdadero estado de la naturaleza es θ. En general, puede considerarse que $\mathbb{C} = \mathbb{D} \times \Theta$, por lo que una consecuencia $c \in \mathbb{C}$ puede identificarse con un par ordenado (d,θ).

$\preceq$. La relación de preferencia: el decisor debe poder expresar sus preferencias personales entre distintas opciones. Una opción, denotada por $\{c_i | E_i\}$, se define como un mapeo de $\{E_i : i \in I\}$ en $\mathbb{C}$, donde $\{E_i : i \in I\}$ denota a una partición de Θ y $c_i \in \mathbb{C}$ $(i \in I)$. En particular, una consecuencia $c \in \mathbb{C}$ corresponde a la opción $\{c|\Theta\}$, por lo que la relación de preferencia $\preceq$ establece un orden en $\mathbb{C}$.

El problema de decisión queda entonces determinado por la cuarteta $(\mathbb{D},\Theta,\mathbb{C},\preceq)$. Por otra parte, el decisor tiene que cuantificar las posibles consecuencias de su elección, así como el conocimiento inicial que tenga sobre el estado de la naturaleza. Loa axiomas o principios de coherencia mencionados al principio de esta sección garantizan la existencia de una única medida de probabilidad **P** y de una única función de utilidad U compatibles

con la relación de preferencia $\preceq$. De esta manera, se supone que el decisor está en condiciones de especificar las siguientes funciones:

$P:\mathcal{A} \longrightarrow [0,1]$. Medida de probabilidad: dado que el estado de la naturaleza es incierto, el decisor tiene que describir su conocimiento inicial sobre θ a través de una medida de probabilidad, cuya función de densidad se denotará por $p(\theta)$.

$U:\mathbb{C} \longrightarrow \mathbb{R}$. Función de utilidad: el decisor debe cuantificar las consecuencias de acuerdo a sus preferencias personales, asignando una utilidad $U(d,\theta)$ a la elección de una acción d, suponiendo que el verdadero estado de la naturaleza es θ.

Una vez que el decisor ha planteado su problema de decisión, todavía es necesario cuantificar adecuadamente las posibles consecuencias de adoptar algún curso de acción determinado, pues dichas consecuencias dependen del verdadero estado de la naturaleza, aún desconocido. Con este fin se define la utilidad esperada de una acción d en $\mathbb{D}$ como

$$U_E(d) = \int U(d,\theta)\ p(\theta)\ d\theta.$$

El resultado fundamental de la teoría bayesiana de toma de decisiones implica que la decisión óptima es aquella que maximiza la utilidad esperada. En otras palabras, siempre que el decisor esté de acuerdo con los axiomas en los que se fundamenta la teoría, la única forma racional de actuar consiste en elegir d^* en $\mathbb{D}$ tal que

$$U_E(d^*) = \sup_{\mathbb{D}} \int U(d,\theta)\ p(\theta)\ d\theta.$$

Suponga ahora que se tiene información adicional contenida en un conjunto $Z=\{X_1, X_2, \ldots, X_n\}$ de observaciones de una variable aleatoria X cuya distribución depende de θ. La información contenida en Z debe usarse entonces para actualizar el conocimiento que se tenga sobre θ, a través del Teorema de Bayes

$$p(\theta|z) \propto p(\theta)\ p(z|\theta),$$

donde $p(z|\theta)$ denota a la función de verosimilitud y $p(\theta|z)$ denota a la función de probabilidad de la distribución final de θ, que describe el conocimiento que tiene el decisor después de haber observado Z. La utilidad esperada final de una acción d está dada entonces por

$$U_E(d|z) = \int U(d,\theta)\ p(\theta|z)\ d\theta$$

y la mejor decisión, por lo tanto, consistirá en elegir d^* en $\mathbb{D}$ tal que

$$U_E(d^*|z) = \sup_{\mathbb{D}} \int U(d,\theta)\ p(\theta|z)\ d\theta.$$

2. PROCEDIMIENTOS BAYESIANOS

2.1 Inferencia y Decision

Como se mencionó en la Sección 1.1, en muchas ocasiones existe el problema de contrastar hipótesis relativas al valor desconocido de algún parámetro. Habiendo establecido de manera precisa las hipótesis a comparar, la forma en que se hace esta comparación dependerá del propósito del análisis, del estado de la información inicial y de si se ha formulado explícitamente alguna función de utilidad.

De acuerdo al propósito del análisis, un conjunto de datos puede ser analizado, por ejemplo, sólo para proporcionar una revisión de las probabilidades iniciales asociadas con las hipótesis. Más aún, puede ocurrir que el investigador no tenga idea de cómo serán utilizados los resultados de su investigación, por lo que no podrá (o no le interesa) formular un problema de decisión con una función de utilidad explícita. El proceso de revisión de las probabilidades iniciales asociadas a las hipótesis no necesariamente involucra alguna decisión con respecto a éstas.

Por otra parte, en muchas ocasiones el objetivo del análisis es alcanzar una decisión con respecto a las hipótesis, digamos aceptar o rechazar. Si es posible especificar alguna función de utilidad, el procedimiento utilizado para alcanzar una decisión consiste en maximizar la utilidad esperada (Sección 1.3). Sin embargo, si no se tiene una función de utilidad explícita, puede

haber cierto grado de arbitrariedad en el análisis. Dicho de otra forma, cuando no se establece de manera precisa cuáles son las consecuencias correspondientes a las acciones, ir más allá del reporte de las probabilidades finales asociadas a las hipótesis puede involucrar algún elemento de arbitrariedad adicional.

Finalmente, un punto importante a considerar es el papel de la información inicial al contrastar hipótesis. La cantidad y el tipo de información inicial a utilizarse en un análisis particular, dependerá de lo que se conoce y de lo que se juzgue apropiado incorporar al análisis. Si se reconoce que existen situaciones en las que inicialmente se conoce muy poco, es necesario entonces contar con procedimientos que permitan contrastar hipótesis incluso si se carece de información inicial. Sin embargo, existen otras situaciones en las sí se cuenta con este tipo de información (por ejemplo, proveniente de análisis previos) y se desea incorporarla en la comparación de las hipótesis. En este caso, la Estadística Bayesiana permite incorporar la información inicial de una manera formal, a través de una distribución de probabilidad que describa el conocimiento inicial que se tenga sobre el valor del parámetro.

En este capítulo se presenta una breve revisión de los procedimientos bayesianos propuestos en la literatura para contrastar hipótesis paramétricas. Las fuentes consultadas incluyen los trabajos de DeGroot (1970), Zellner (1971), Lindley (1972), Winkler (1972), Berger (1980) y Press (1989), entre otros. Al igual que en los procedimientos clásicos revisados en el capítulo anterior, en los procedimientos bayesianos revisados en este capítulo se supone que $\Theta = \Theta_0 \cup \Theta_1$, de manera que $\{\Theta_0, \Theta_1\}$ forma una partición del espacio parametral.

2.2 Comparacion de Hipotesis

A lo largo de esta sección se presentarán algunos de los procedimientos bayesianos más comunes para atacar el problema de contraste de hipótesis paramétricas. Los procedimientos presentados en esta sección se basan esencialmente en la comparación de las probabilidades asociadas a cada una de las hipótesis y, por lo tanto, no requieren de la especificación de una función de utilidad explícita.

Sea X una variable aleatoria con función de densidad $p(x|\theta)$ y suponga que se desea contrastar las hipótesis

$$H_0:\theta \in \Theta_0 \quad \text{vs.} \quad H_1:\theta \in \Theta_1$$

con base en un conjunto $Z = \{X_1, X_2, \ldots, X_n\}$ de observaciones de X. Asimismo, sea $p(\theta)$ la función de densidad de la distribución inicial de θ y denote por $p(\theta|z)$ a la densidad de la correspondiente distribución final. Una manera de contrastar estas hipótesis consiste en comparar las probabilidades finales correspondientes a cada una de ellas. Dicho de otra forma, si

$$P[H_0|z] := P[\theta \in \Theta_0|z] = \int_{\Theta_0} p(\theta|z)\, d\theta$$

$$P[H_1|z] := P[\theta \in \Theta_1|z] = \int_{\Theta_1} p(\theta|z)\, d\theta$$

entonces H_1 es más verosímil (o más creíble) que H_0 siempre que $P[H_0|z] < P[H_1|z]$ o, equivalentemente, siempre que

$$\frac{P[H_0|z]}{P[H_1|z]} < 1. \qquad (2.1)$$

Observe que, al suponer que $\{\Theta_0,\Theta_1\}$ es una partición del espacio parametral Θ, entonces $P[H_1|z] = 1 - P[H_0|z]$. En este caso, al cociente en (2.1) se le conoce como los momios finales en favor de H_0 y se denota por M_{01}.

Cabe mencionar que si se desea comparar más de dos hipótesis, este procedimiento puede extenderse de manera natural: simplemente se tendría mayor evidencia a favor de aquella hipótesis cuya probabilidad final fuese la más alta.

Si no se desea incorporar al análisis el conocimiento inicial sobre el valor de θ, debe utilizarse una distribución inicial de referencia o distribución inicial no informativa. Dichas distribuciones pueden obtenerse a través de diversos procedimientos, como la Regla de Jeffreys (Jeffreys, 1961), el método de Familias Conjugadas (DeGroot, 1970), o recurriendo a nociones de la Teoría de la Información (Bernardo, 1979a).

Resulta de interés distinguir tres casos, de acuerdo a la forma de las hipótesis que se desee contrastar, ya que frecuentemente se requieren tipos especiales de distribuciones iniciales si las dimensiones de Θ_0 y Θ_1 son distintas.

CASO 1. $H_0:\theta = \theta_0$ vs. $H_1:\theta = \theta_1$.

En este caso $\Theta = \{\theta_0,\theta_1\}$, por lo que $p(\theta)$ toma la forma

$$p(\theta) = \begin{cases} p & \text{si } \theta = \theta_0 \\ 1-p & \text{si } \theta = \theta_1 \end{cases} \qquad (0 < p < 1) \qquad (2.2)$$

y, por lo tanto,

$$M_{01} = \frac{p}{1-p} \quad \frac{p(z|\theta_0)}{p(z|\theta_1)} \ .$$

En otras palabras, los momios finales en favor de H_0 son el producto de los mimios iniciales en favor de H_0 por el cociente de verosimilitudes. De acuerdo con (2.1), H_1 es más verosímil que H_0 si

$$\frac{p(z|\theta_0)}{p(z|\theta_1)} < \frac{1-p}{p} \ .$$

Compare esta regla con la regla de rechazo dada por el Lema de Neyman-Pearson (Sección 1.2).

Por otro lado, la distribución de referencia comúnmente aceptada en este caso está dada por

$$\pi(\theta) = \begin{cases} 1/2 & \text{si } \theta = \theta_0 \\ 1/2 & \text{si } \theta = \theta_1 \end{cases}$$

por lo que, si no se cuenta con información inicial, entonces se tiene mayor evidencia a favor de H_1 cuando

$$\frac{p(z|\theta_0)}{p(z|\theta_1)} < 1.$$

CASO 2. $H_0:\theta = \theta_0$ vs. $H_1:\theta \neq \theta_0$.

Este caso presenta una pequeña dificultad: si se supone que la distribución inicial de θ es continua, entonces $P[\theta = \theta_0] = 0$ (y, por lo tanto, $P[\theta = \theta_0|z] = 0$ para todo z), por lo que nunca se tendría suficiente evidencia a favor de H_0. Esta dificultad se resuelve comúnmente definiendo a $p(\theta)$ de la siguiente manera:

$$p(\theta) = \begin{cases} p & \text{si } \theta = \theta_0 \\ (1-p)\ f(\theta) & \text{si } \theta \neq \theta_0 \end{cases} \quad (\ 0 < p < 1\ )\ , \qquad (2.3)$$

donde $f(\theta)$ denota a una función de densidad definida sobre $\Theta_1 = \Theta - \{\theta_0\}$. De esta manera, se asigna una probabilidad positiva p al valor θ_0 de θ y la probabilidad restante (1-p) se distribuye sobre Θ_1.

En estas condiciones, los momios finales en favor de H_0 están dados por

$$M_{01} = \frac{p}{1-p}\ \frac{p(z|\theta_0)}{\int_{\Theta_1} f(\theta)\ p(z|\theta)\ d\theta}\ ,$$

de manera que H_1 es más verosímil que H_0 si

$$\frac{p(z|\theta_0)}{\int_{\Theta_1} f(\theta)\ p(z|\theta)\ d\theta} < \frac{1-p}{p}\ .$$

Este caso resulta problemático también en lo que respecta a la asignación de una distribución de referencia. Una posible solución (Bernardo, 1980) consiste en maximizar la cantidad de información faltante (*missing information*), la cual está dada en este caso por

$$\mathcal{I}[p(\theta)] = H(p) + (1-p)\ I^n[f(\theta)],$$

donde

$$H(p) = -\ p\ log\ p - (1-p)\ log\ (1-p)$$

y

$$I^n[f(\theta)] = \int p(z) \int p(\theta|z) \log \frac{p(\theta|z)}{p(\theta)} \, d\theta \, dz.$$

De esta forma, para $f(\theta)$ fija, se tiene que la probabilidad inicial π que maximiza la cantidad de información faltante es tal que

$$\frac{1-\pi}{\pi} = exp\{I^n[p(\theta)]\},$$

(Bernardo, 1980) por lo que, en ausencia de información inicial, se tiene mayor evidencia a favor de H_1 si

$$\frac{p(z|\theta_0)}{\int_{\Theta_1} f(\theta) \, p(z|\theta) \, d\theta} < exp\{I^n[p(\theta)]\}.$$

Esta solución, sin embargo, no es del todo adecuada ya que la probabilidad inicial de referencia, π, depende del tamaño de la muestra. El problema consiste en que una distribución inicial no puede depender de la muestra, ya que debe describir el conocimiento inicial que se tiene sobre el valor de θ antes de que se observe la muestra.

CASO 3. $H_0: \theta \in \Theta_0$ vs. $H_1: \theta \in \Theta_1$.

Si tanto Θ_0 como Θ_1 contienen un continuo de puntos y son de la misma dimensión, entonces es posible definir una densidad inicial $p(\theta)$ sobre el espacio parametral Θ. En este caso los momios finales en favor de H_0 están dados por

$$M_{01} = \frac{\int_{\Theta_0} p(\theta|z)\ d\theta}{\int_{\Theta_1} p(\theta|z)\ d\theta} = \frac{\int_{\Theta_0} p(\theta)\ p(z|\theta)\ d\theta}{\int_{\Theta_1} p(\theta)\ p(z|\theta)\ d\theta},$$

por lo que se tiene mayor evidencia a favor de H_1 siempre que

$$\frac{\int_{\Theta_0} p(\theta)\ p(z|\theta)\ d\theta}{\int_{\Theta_1} p(\theta)\ p(z|\theta)\ d\theta} < 1.$$

Dicho de otra forma, H_1 es más verosímil que H_0 si el cociente de las verosimilitudes ponderadas es menor que 1. Compare esta regla con la regla de rechazo obtenida por el método generalizado de cociente de verosimilitudes (Sección 1.2).

Finalmente, si no se desea incorporar la información inicial sobre el valor de θ, puede usarse una distribución inicial de referencia $\pi(\theta)$. Dicha distribución puede obtenerse, por ejemplo, a través del procedimiento propuesto por Bernardo (1979a).

En este caso existe mayor evidencia a favor de H_1 siempre que

$$\frac{\int_{\Theta_0} \pi(\theta)\ p(z|\theta)\ d\theta}{\int_{\Theta_1} \pi(\theta)\ p(z|\theta)\ d\theta} < 1.$$

2.3 Planteamiento del Problema de Contraste de Hipotesis Parametricas como un Problema de Decision

El problema general de contraste de hipótesis paramétricas descrito en la Sección 1.1 puede plantearse como un problema de toma de decisiones en ambiente de incertidumbre, cuya solución general fue presentada en la Sección 1.3. Dependiendo de la función de utililidad particular que se asigne, pueden generarse distintos procedimientos para contrastar las hipótesis que sean de interés.

Sea X una variable aleatoria con función de densidad $p(x|\theta)$ y suponga que se desea contrastar las hipótesis

$$H_0:\theta \in \Theta_0 \quad \text{vs.} \quad H_1:\theta \in \Theta_1$$

con base en un conjunto $Z = \{X_1, X_2, \ldots, X_n\}$ de observaciones de X.

Los elementos del problema de decisión correspondiente son los siguientes:

$\mathbb{D} = \{d_0, d_1\}$, el espacio de decisiones, donde d_i denota la decisión de elegir a la hipótesis H_i como cierta (i=0,1).

$\mathbb{E} = \Theta$, el espacio de estados de la naturaleza. Dado que en este caso la incertidumbre corresponde al valor del parámetro, debe considerarse al espacio parametral como el conjunto de los estados de la naturaleza.

$U(d,\theta)$, la función de utilidad, que permite cuantificar las consecuencias de tomar la decisión d cuando el verdadero valor del parámetro es θ, siempre de acuerdo a las preferencias del decisor.

$p(\theta)$, la distribución inicial de θ, que describe el conocimiento previo que el decisor tiene sobre el valor del parámetro.

Debe recordarse que la información contenida en Z permite actualizar el conocimiento que se tiene sobre el valor de θ, a través del Teorema de Bayes, obteniéndose una distribución final $p(\theta|z)$ para θ. De acuerdo a lo expuesto en la Sección 1.3, la solución general de este problema de decisión consiste en elegir H_1 si y sólo si

$$U_E(d_0|z) < U_E(d_1|z), \tag{3.1}$$

donde $U_E(d_i|z)$ denota a la utilidad esperada final de la decisión d_i $(i=0,1)$.

Recuerde que $\{\Theta_0,\Theta_1\}$ es una partición del espacio parametral. Una posible asignación de la función de utilidad, frecuente en la literatura, es la siguiente:

$$U(d_0,\theta) = \begin{cases} a_0(\theta) & \text{si } \theta \in \Theta_0 \\ 0 & \text{si } \theta \in \Theta_1 \end{cases}$$

$$U(d_1,\theta) = \begin{cases} a_1(\theta) & \text{si } \theta \in \Theta_1 \\ 0 & \text{si } \theta \in \Theta_0 \end{cases} \tag{3.2}$$

donde a_0 y a_1 son funciones no negativas de θ.

En este caso, la utilidad esperada final de cada una de las decisiones está dada por

$$U_E(d_0|z) = \int_{\Theta_0} a_0(\theta)\ p(\theta|z)\ d\theta$$

$$U_E(d_1|z) = \int_{\Theta_1} a_1(\theta)\ p(\theta|z)\ d\theta,$$

por lo cual debe rechazarse H_0 si y sólo si

$$\frac{\int_{\Theta_0} a_0(\theta)\ p(\theta|z)\ d\theta}{\int_{\Theta_1} a_1(\theta)\ p(\theta|z)\ d\theta} < 1. \qquad (3.3)$$

Aquí, como en la sección anterior, pueden distinguirse tres casos de acuerdo a la forma de las hipótesis que se desea contrastar, ya que se requiere considerar casos particulares de (3.2) dependiendo de las dimensiones de Θ_0 y Θ_1.

CASO 1. $H_0:\theta = \theta_0$ vs. $H_1:\theta = \theta_1$.

En este caso, la función de utilidad (3.2) toma la forma

$$U(d_0,\theta) = \begin{cases} a_0 & \text{si } \theta = \theta_0 \\ 0 & \text{si } \theta = \theta_1 \end{cases}$$

$$U(d_1,\theta) = \begin{cases} a_1 & \text{si } \theta = \theta_1 \\ 0 & \text{si } \theta = \theta_0 \end{cases}$$

donde a_0 y a_1 son dos constantes no negativas. Si se hace uso de la distribución inicial $p(\theta)$ dada por (2.2), entonces se elige d_1 (se rechaza H_0) siempre que

$$\frac{p(z|\theta_0)}{p(z|\theta_1)} < \frac{a_1\ (1-p)}{a_0\ p}\ .$$

CASO 2. $H_0:\theta = \theta_0$ vs. $H_1:\theta \neq \theta_0$.

La función de utilidad (3.2) está dada en este caso por

$$U(d_0,\theta) = \begin{cases} a_0 & \text{si } \theta = \theta_0 \\ 0 & \text{si } \theta \neq \theta_0 \end{cases}$$

$$U(d_1,\theta) = \begin{cases} a_1(\theta) & \text{si } \theta \neq \theta_0 \\ 0 & \text{si } \theta = \theta_0 \end{cases}$$

de manera que, cuando $p(\theta)$ es la distribución inicial definida por (2.3), entonces se elige d_1 (se rechaza H_0) si y sólo si

$$\frac{p(z|\theta_0)}{\int_{\Theta_1} a_1(\theta)\ f(\theta)\ p(z|\theta)\ d\theta} < \frac{1-p}{a_0\ p} ,$$

donde, por supuesto, $\Theta_1 = \Theta - \{\theta_0\}$.

CASO 3. $H_0:\theta \in \Theta_0$ vs. $H_1:\theta \in \Theta_1$.

En este caso la función de utilidad está definida directamente por (3.2). Si, como en la sección anterior, $p(\theta)$ denota a una densidad definida sobre Θ, entonces la regla de decisión consiste en rechazar la hipótesis H_0 si y sólo si se cumple la condición dada por (3.3).

Cabe mencionar que, en cualquiera de estos tres casos, pueden reproducirse las reglas de rechazo obtenidas en la sección anterior si se utiliza el siguiente caso particular de la función de utilidad (3.2):

$$U(d_0,\theta) = \begin{cases} 1 & \text{si } \theta \in \Theta_0 \\ 0 & \text{si } \theta \in \Theta_1 \end{cases}$$

$$U(d_1,\theta) = \begin{cases} 1 & \text{si } \theta \in \Theta_1 \\ 0 & \text{si } \theta \in \Theta_0 \end{cases}$$

Finalmente, si no se tiene información inicial sobre el valor de θ, entonces pueden utilizarse distribuciones de referencia en cada uno de los casos.

2.4 Parametros de Ruido

Una situación que surge frecuentemente en la práctica se obtiene cuando el parámetro es de la forma (θ,ω), donde θ es el parámetro de interés y ω es un parámetro de ruido (que no es de interés). De esta manera, el espacio parametral puede denotarse como $\Theta \times \Omega$.

Sea $Z = \{X_1, X_2, \ldots, X_n\}$ un conjunto de observaciones de una variable aleatoria X con función de densidad $p(x|\theta,\omega)$ y suponga que, independientemente del valor de ω, se desea contrastar las hipótesis

$$H_0: \theta \in \Theta_0 \quad \text{vs.} \quad H_1: \theta \in \Theta_1,$$

donde $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ y $\Theta_0 \cap \Theta_1 = \emptyset$.

Sea $p(\theta,\omega)$ la distribución inicial de (θ,ω). Por el Teorema de Bayes, la distribución final correspondiente está dada por

$$p(\theta,\omega|z) \propto p(\theta,\omega)\ p(z|\theta,\omega), \qquad (4.1)$$

donde $p(z|\theta,\omega)$ denota ahora a la función de verosimilitud. La distribución marginal final de θ puede obtenerse a través de la relación

$$p(\theta|z) = \int_{\Omega} p(\theta,\omega|z)\ d\omega. \qquad (4.2)$$

Por otro lado, la utilidad esperada final de una decisión d puede escribirse en este caso como

$$U_E(d|z) = \int\int U(d;\theta,\omega)\ p(\theta,\omega|z)\ d\theta\ d\omega, \qquad (4.3)$$

donde $U(d;\theta,\omega)$ denota a la utilidad de la decisión d cuando θ es el valor del parámetro de interés y ω es el valor del parámetro de ruido.

Observe que si la función de utilidad no depende de ω, entonces (4.3) puede escribirse como

$$\begin{aligned} U_E(d|z) &= \int\int U(d;\theta)\ p(\theta,\omega|z)\ d\theta\ d\omega \\ &= \int U(d;\theta) \int p(\theta,\omega|z)\ d\omega\ d\theta \\ &= \int U(d;\theta)\ p(\theta|z)\ d\theta, \end{aligned}$$

de manera que el problema se reduce al caso en el que no existen parámetros de ruido.

Por lo tanto, los procedimientos presentados en la Sección 2.2 pueden utilizarse tomando en cuenta la expresión (4.2) para el cálculo de las probabilidades de cada una de las hipótesis en (4.1). Asimismo, la expresión (4.3) permite extender directamente los procedimientos presentados en la Sección 2.3 al caso en el que existen parámetros de ruido.

3. UNA SOLUCION ALTERNATIVA

En este capítulo se propone un procedimiento bayesiano que permite dar un tratamiento unificado al problema de contraste de hipótesis paramétricas, dentro del contexto de la Teoría de Decisiones discutido en la Sección 2.3. Dicho procedimiento se basa en la consideración de una familia particular de funciones de utilidad, definida en términos de una medida de la discrepancia entre modelos, de acuerdo con una propuesta de Bernardo (1984) para el caso de hipótesis de la forma $H_0:\theta = \theta_0$ vs. $H_1:\theta \neq \theta_0$.

3.1 MOTIVACION

Suponga que se desea contrastar las hipótesis $H_0:\theta = \theta_0$ vs. $H_1:\theta \neq \theta_0$ con base en un conjunto $Z = \{X_1, X_2, \ldots, X_n\}$ de observaciones de una variable aleatoria X con función de densidad $p(x|\theta)$. De acuerdo a lo expuesto en la Sección 1.3, la acción d_1 (rechazar H_0) será preferible a la acción d_0 (aceptar H_0) si y sólo si $U_E(d_1|z) > U_E(d_0|z)$, es decir, si y sólo si

$$\int U(d_1,\theta)\ p(\theta|z)\ d\theta > \int U(d_0,\theta)\ p(\theta|z)\ d\theta$$

o, equivalentemente, siempre y cuando

$$\int \{U(d_1,\theta) - U(d_0,\theta)\}\ p(\theta|z)\ d\theta > 0. \qquad (1.1)$$

En este caso, por lo tanto, basta especificar la diferencia de las utilidades. Bernardo (1984) sugiere que la ventaja de elegir d_1 debe depender solamente de la discrepancia entre $p(x|\theta)$ y $p(x|\theta_0)$, ya que, por el principio de verosimilitud, si ambos modelos son iguales entonces las inferencias que se hagan con base en $p(z|\theta_0)$ serán las mismas que se harían al basarse en $p(z|\theta)$. Dicha ventaja deberá aumentar conforme crezca la discrepancia entre los modelos involucrados. En vista de esto, si se denota por $\delta(\theta:\theta_0)$ a una medida de la discrepancia entre $p(x|\theta)$ y $p(x|\theta_0)$, puede considerarse a la diferencia $\{U(d_1,\theta) - U(d_0,\theta)\}$ como una función no decreciente de $\delta(\theta:\theta_0)$. En particular, Bernardo propone considerar

$$U(d_1,\theta) - U(d_0,\theta) = A\,\delta(\theta:\theta_0) + B \qquad A \in \mathbb{R}^+,\ B \in \mathbb{R}.$$

La expresión (1.1), por lo tanto, implica que la decisión óptima es d_1 si y sólo si

$$E[\delta(\theta:\theta_0)|z] > \delta^*,$$

donde $\delta^* = -B/A$ y

$$E[\delta(\theta:\theta_0)|z] = \int \delta(\theta:\theta_0)\ p(\theta|z)\ d\theta.$$

Las propiedades de este procedimiento, para elecciones particulares de $\delta(\theta:\theta_0)$ y δ^*, han sido estudiadas por Bernardo (1984). Por otra parte, Gutiérrez (1989) analiza el comportamiento de esta solución particular en algunas de las familias paramétricas univariadas más comunes.

Suponga ahora que se desea contrastar las hipótesis $H_0:\theta = \theta_0$ vs. $H_1:\theta = \theta_1$. Utilizando nuevamente la noción de discrepancia entre modelos, suponga que $\delta(\theta:\theta_0) > \delta(\theta:\theta_1)$. Resulta natural suponer entonces que $p(x|\theta_1)$ proporciona una mejor aproximación que $p(x|\theta_0)$ al modelo indexado por el verdadero valor del parámetro, es decir, $p(x|\theta)$. Claramente, a medida que la

discrepancia entre $p(x|\theta_1)$ y $p(x|\theta)$ disminuya, la utilidad de la acción correspondiente, d_1, debe incrementarse. Estas consideraciones permiten definir una familia de funciones de utilidad en términos de la discrepancia entre los modelos, de la siguiente manera:

$$U(d_0,\theta) = B_0 - A\,\delta(\theta:\theta_0)$$

$$U(d_1,\theta) = B_1 - A\,\delta(\theta:\theta_1)$$

donde $A \in \mathbb{R}^+$ y $B_i \in \mathbb{R}$ $(i=0,1)$.

En estas condiciones, la mejor decisión es d_1 si y sólo si

$$E[\delta(\theta:\theta_0)|z] - E[\delta(\theta:\theta_1)|z] > \delta^*,$$

donde ahora $\delta^* = (B_0-B_1)/A$ (Rueda & Gutiérrez, 1990).

Observe que este procedimiento extiende en cierta forma el procedimiento propuesto por Bernardo, ya que esencialmente se elige la acción correspondiente al modelo cuya discrepancia esperada con respecto a $p(x|\theta)$ es menor.

Las ideas expuestas hasta el momento pueden extenderse para atacar el problema general de contraste de hipótesis descrito en la Sección 1.1, como se verá en la siguiente sección.

3.2 Planteamiento

Sea X una variable aleatoria m-variada con función de densidad $p(x|\theta)$, perteneciente a una familia paramétrica

$$\mathcal{F} = \left\{ p(x|\theta) : \theta \in \Theta \right\} \qquad \Theta \subseteq \mathbb{R}^d$$

y sea $Z = \{X_1, X_2, \ldots, X_n\}$ un conjunto de observaciones de X.

Suponga que se desea contrastar las hipótesis

$$H_0:\theta \in \Theta_0 \quad \text{vs.} \quad H_1:\theta \in \Theta_1$$

donde $\Theta_0 \subset \Theta$ y $\Theta_1 \subset \Theta$ son tales que $\Theta_0 \cap \Theta_1 = \emptyset$. Recuerde que, al plantear este problema como un problema de decisión, a cada una de las hipótesis le corresponde un curso de acción determinado, denotado por d_0 o d_1 respectivamente.

Como se mencionó en la Sección 1.1, estas hipótesis son matemáticamente equivalentes a las hipótesis

$$H_0':p(x|\theta) \in \mathcal{F}_0 \quad \text{vs.} \quad H_1':p(x|\theta) \in \mathcal{F}_1,$$

donde

$$\mathcal{F}_i = \left\{ p(x|\theta) \ : \ \theta \in \Theta_i \right\} \qquad (i=0,1).$$

De esta manera, por ejemplo, al elegir la acción d_1 se está considerando que $p(x|\theta)$ puede aproximarse más adecuadamente a través de un elemento de $\mathcal{F}_1$ que a través de un elemento de $\mathcal{F}_0$. Es por esto que se propone, como parte de la asignación de la función de utilidad, que la utilidad de cada una de las acciones esté en función de la discrepancia entre los modelos involucrados.

Extendiendo la propuesta para contrastar las hipótesis $H_0:\theta = \theta_0$ vs. $H_1:\theta = \theta_1$ (discutida en la sección anterior), para el caso general $H_0:\theta \in \Theta_0$ vs. $H_1:\theta \in \Theta_1$ se propone una función de utilidad de la forma

$$U(d_1,\theta) = B_1 - A\,\delta(\theta:\theta_1^*) \qquad A \in \mathbb{R}^+,\ B_1 \in \mathbb{R}, \qquad (2.1)$$

donde $\theta_1^* \in \Theta_1$ es un valor *representativo* de Θ_1 $(i=0,1)$.

La elección de $\theta_i^* \in \Theta_i$ (i=0,1) debe considerarse a su vez como un problema de decisión en ambiente de incertidumbre. De acuerdo con la discusión de la Sección 3.4, bajo ciertas condiciones el valor $\theta_i^* \in \Theta_i$ puede elegirse de manera que

$$E[\delta(\theta:\theta_i^*)|z] = inf\left\{ E[\delta(\theta:\theta_i)|z] : \theta_i \in \Theta_i \right\} \quad (i=0,1). \qquad (2.2)$$

En estas condiciones, la acción d_1 será preferible a la acción d_0 si y sólo si

$$U_E(d_1|z) > U_E(d_0|z)$$

o, equivalentemente, si y sólo si

$$E[\delta(\theta:\theta_0^*)|z] - E[\delta(\theta:\theta_1^*)|z] > \delta^*, \qquad (2.3)$$

donde $\delta^* = (B_0-B_1)/A$. Observe que, de acuerdo con (2.2), la expresión (2.3) puede reescribirse como

$$inf\left\{E[\delta(\theta:\theta_0)|z] : \theta_0 \in \Theta_0\right\} - inf\left\{E[\delta(\theta:\theta_1)|z] : \theta_1 \in \Theta_1\right\} > \delta^*.$$

En resumen, el procedimiento propuesto en esta sección consiste esencialmente en elegir primero un valor *representativo* $\theta_i^* \in \Theta_i$ (i=0,1) y entonces elegir la decisión correspondiente al valor cuya discrepancia esperada con respecto a θ sea mínima.

3.3 Divergencia Logaritmica de Kullback-Leibler como Medida de Discrepancia

Las propiedades del procedimiento descrito en la sección anterior dependen de la elección de la medida de discrepancia, $\delta(\theta:\theta_0)$. Una medida que ha resultado adecuada en este contexto es la llamada divergencia logarítmica de Kullback-Leibler, dada por

$$\delta(\theta:\theta_0) = \int p(x|\theta)\ log\ \frac{p(x|\theta)}{p(x|\theta_0)}\ dx, \qquad (3.1)$$

la cual puede interpretarse como la información media para discriminar en contra de $p(x|\theta_0)$ (y, por lo tanto, en favor de $p(x|\theta)$) cuando $p(x|\theta)$ es la verdadera distribución de X (Kullback, 1959). La divergencia logarítimica mide la discrepancia entre dos distribuciones: entre más grande sea $\delta(\theta:\theta_0)$ más grande será discrepancia entre $p(x|\theta)$ y $p(x|\theta_0)$.

Esta medida surge al plantear el problema de aproximación de distribuciones como un problema de toma de decisiones. Bajo ciertas condiciones se tiene que, al aproximar una distribución de probabilidad p a través de algún elemento de una familia de distribuciones determinada, debe elegirse aquel elemento de la familia cuya divergencia logarítmica con respecto a p sea mínima (Bernardo, 1979b). Cabe mencionar, por otra parte, que la divergencia logarítmica de Kullback-Leibler ha sido propuesta por Bernardo (1984) como una medida de discrepancia en el procedimiento de contraste para las hipótesis $H_0:\theta = \theta_0$ vs. $H_1:\theta \neq \theta_0$, revisado en la Sección 3.1.

Si se sustituye (3.1) en la regla de decisión dada por la expresión (2.2), se sigue que la decisión óptima es d_1 si y sólo si

$$E\left[\int p(x|\theta)\ log\ \frac{p(x|\theta_1^*)}{p(x|\theta_0^*)}\ dx \;\middle|\; z \right] > \delta^*,$$

donde el valor esperado se toma con respecto a la distribución final de θ y θ_i^* es el valor de θ_i que minimiza $E[\delta(\theta:\theta_i)|z]$ en θ_i (i=0,1).

3.4 El Problema de Estimacion

La elección del valor $\theta_i^* \in \Theta_i$ (i=0,1), introducido en la expresión (2.1), puede considerarse como un problema de estimación con restricciones y debe plantearse como un problema de decisión en ambiente de incertidumbre. En general, el problema de estimación puede plantearse como un problema de decisión donde el espacio $\mathbb{D}$ de las posibles decisiones coincide con el espacio parametral Θ, siempre que este espacio no se restrinja. El espacio de estados de la naturaleza corresponde también al espacio parametral, pues la incertidumbre reside aún en el valor del parámetro.

En esencia, la elección de $\theta_i^* \in \Theta_i$ es equivalente a la elección de un modelo *representativo* de cada una de las familias $\mathcal{F}_i$ (i=0,1), por lo que de nueva cuenta resulta natural el uso de una función de utilidad definida en términos de una medida de discrepancia. En particular, suponga que

$$U'(\hat{\theta},\theta) = -\delta(\theta:\hat{\theta}) \qquad \hat{\theta} \in \Theta,\ \theta \in \Theta \qquad (4.1)$$

donde $\delta(\theta:\hat{\theta})$ denota a la medida de discrepancia utilizada en (2.1) para el problema de contraste de hipótesis. Se tiene entonces que el valor $\hat{\theta}^*$ de $\hat{\theta}$ que maximiza la utilidad esperada final es tal que

$$\Delta(\hat{\theta}^*) = inf \left\{ \Delta(\hat{\theta}) : \hat{\theta} \in \Theta \right\}, \qquad (4.2)$$

donde

$$\Delta(\hat{\theta}) = E[\delta(\theta:\hat{\theta})|z] = \int \delta(\theta:\hat{\theta})\ p(\theta|z)\ d\theta.$$

La expresión (4.2) permite justificar la elección del valor $\theta_i^* \in \Theta_i$ (i=0,1), definido en (2.1), necesario en la asignación de la función de utilidad para el problema de contraste de hipótesis.

Una propiedad importante del estimador obtenido a través de este procedimiento es que resulta invariante ante transformaciones uno a uno. En efecto, suponga que no se está interesado en estimar el valor de θ, sino de una transformación de θ. Mas aún, suponga que $\varphi = g(\theta)$ es una transformación uno a uno y diferenciable, tal que g^{-1} es también diferenciable. Si $p(\theta|z)$ denota la distribución final de θ, entonces la distribución final de φ está dada por

$$p_1(\varphi|z) = p(g^{-1}(\varphi)|z)\, J(\varphi),$$

donde $J(\varphi) = \left| \dfrac{\partial\, g^{-1}(\varphi)}{\partial\, \varphi} \right|$ denota al Jacobiano de la transformación. Sea $\delta_1(\varphi:\hat{\varphi})$ la discrepancia entre $p(x|\varphi)$ y $p(x|\hat{\varphi})$, obtenida al parametrizar la distribución de X en términos de φ. Se tiene entonces que

$$\Delta_1(\hat{\varphi}) = E[\delta_1(\varphi:\hat{\varphi})|z] = \int \delta_1(\varphi:\hat{\varphi})\; p_1(\varphi|z)\; d\varphi$$

$$= \int \delta(g^{-1}(\varphi):g^{-1}(\hat{\varphi}))\; p(g^{-1}(\varphi)|z)\; J(\varphi)\; d\varphi$$

$$= \int \delta(\theta:g^{-1}(\hat{\varphi}))\; p(\theta|z)\; d\theta$$

$$= \Delta(g^{-1}(\hat{\varphi})),$$

de donde

$$\Delta(\hat{\theta}) = \Delta_1(g(\hat{\theta})). \qquad (4.3)$$

De acuerdo con la expresión (4.2) se tiene que

$$\Delta(\hat{\theta}^*) \leq \Delta(\hat{\theta}) \qquad \text{para todo } \hat{\theta},$$

de manera que

$$\Delta_1(g(\hat{\theta}^*)) \leq \Delta_1(g(\hat{\theta})) \qquad \text{para todo } \hat{\theta},$$

lo cual implica que

$$\Delta_1(g(\hat{\theta}^*)) \leq \Delta_1(\hat{\varphi}) \qquad \text{para todo } \hat{\varphi}.$$

Por lo tanto, el estimador dado por $\hat{\theta}^*$ es invariante ante transformaciones uno a uno, ya que $\Delta_1(\hat{\varphi})$ se minimiza en $\hat{\varphi}^* = g(\hat{\theta}^*)$.

Suponga ahora que $\delta(\theta:\theta_0)$ es la divergencia logarítmica de Kullback-Leibler, introducida en la sección anterior. En este caso, de la expresión (4.1) se sigue que

$$U'(\hat{\theta},\theta) = -\int p(x|\theta)\ log\ \frac{p(x|\theta)}{p(x|\hat{\theta})}\ dx$$

$$= \int p(x|\theta)\ log\ p(x|\hat{\theta})\ dx - \int p(x|\theta)\ log\ p(x|\theta)\ dx,$$

de donde

$$U'_E(\hat{\theta}) = E\left[\int p(x|\theta)\ log\ p(x|\hat{\theta})\ dx|z\right] - E\left[\int p(x|\theta)\ log\ p(x|\theta)\ dx|z\right]$$

Por lo tanto, maximizar la utilidad esperada final con respecto a $\hat{\theta}$ es equivalente a maximizar

$$E\left[\int p(x|\theta)\ log\ p(x|\hat{\theta})\ dx \mid z\right].$$

Suponga ahora que se desea minimizar la divergencia logarítmica entre $p(x|\hat{\theta})$ y $p(x|z)$, donde $p(x|z)$ denota a la distribución predictiva final de X. En otras palabras, se desea minimizar

$$\delta\left\{p(x|z):p(x|\hat{\theta})\right\} = \int p(x|z)\ log\ \frac{p(x|z)}{p(x|\hat{\theta})}\ dx$$

$$= \int p(x|z)\ log\ p(x|z)\ dx - \int p(x|z)\ log\ p(x|\hat{\theta})\ dx.$$

De esta manera, minimizar $\delta\left\{p(x|z):p(x|\hat{\theta})\right\}$ con respecto a $\hat{\theta}$ es equivalente a maximizar

$$\int p(x|z)\ log\ p(x|\hat{\theta})\ dx.$$

Pero

$$E\left[\int p(x|\theta)\ log\ p(x|\hat{\theta})\ dx\ |z\right] = \int p(\theta|z)\ \left\{\int p(x|\theta)\ log\ p(x|\hat{\theta})\ dx\right\}\ d\theta$$

$$= \int log\ p(x|\hat{\theta})\ \left\{\int p(x|\theta)\ log\ p(\theta|z)\ d\theta\right\}\ dx$$

$$= \int p(x|z)\ log\ p(x|\hat{\theta})\ dx.$$

Por lo tanto, $p(x|\hat{\theta}^{*})$ puede considerarse como la mejor aproximación, entre todos los modelos en la familia $\mathcal{F}$, a la distribución predictiva final $p(x|z)$ (ver Sección 3.3).

En esta sección se propone definir a la función de utilidad para el problema de estimación en términos de una medida de discrepancia entre modelos. Dada la naturaleza secuencial del procedimiento de contraste de hipótesis descrito en la Sección 3.2, es posible utilizar distintas funciones de utilidad en cada

uno de los problemas de decisión involucrados: la elección de $\theta_1^* \in \Theta_1$ (estimación) y la elección de una hipótesis que pueda considerarse como cierta (contraste de hipótesis).

3.5 Familias Exponenciales

Como una aplicación importante, en esta sección se tratará con detalle el caso en el que la familia paramétrica $\mathcal{F}$ es una familia exponencial regular. Asimismo, se hará uso de distribuciones iniciales conjugadas. Una referencia importante en esta sección es el trabajo de Diaconis & Ylvisaker (1979).

3.5.1 Divergencia Logaritmica Esperada para Familias Exponenciales

Una familia de distribuciones cuya función de densidad puede expresarse de la forma

$$p(x|\lambda) = a_*(\lambda)\, b(x)\, exp\, \{c(\lambda)'t(x)\} \qquad x \in \mathbb{R}^m,\ \lambda \in \mathbb{R}^d \qquad (5.1)$$

con $c(\lambda)'=(c_1(\lambda),c_2(\lambda),\ldots,c_d(\lambda))$, $t(x)=(t_1(x),t_2(x),\ldots,t_d(x))'$ y $c(\lambda)'t(x) = \sum_{j=1}^{d} c_j(\lambda)t_j(x)$, es llamada una familia exponencial. El vector $\theta = c(\lambda)$ es llamado parámetro natural o parámetro canónico.

En ocasiones resulta conveniente trabajar con una parametrización más natural, en términos del parámetro canónico. Sea

$$p(x|\theta) = a(\theta)\, b(x)\, exp\, \{\theta't(x)\} \qquad x \in \mathbb{R}^m \qquad (5.2)$$

con $\theta' = (\theta_1,\theta_2,\ldots,\theta_d)$, $t(x) = (t_1(x),t_2(x),\ldots,t_d(x))'$ y $\theta't(x) = \sum_{j=1}^{d} \theta_j t_j(x)$. La función $p(x|\theta)$ es una densidad con

respecto a alguna medida σ-finita μ definida sobre los Borelianos de $\mathbb{R}^m$ (típicamente μ es la medida de Lebesgue o una medida de conteo).

Cabe mencionar que las familias exponenciales son esencialmente los únicos modelos que permiten una reducción suficiente de los datos, es decir, bajo condiciones de regularidad siempre existe una estadística suficiente de dimensión finita independientemente del tamaño de la muestra (Barndorff-Nielsen & Pedersen, 1968).

Sea $\Theta = \{ \theta \in \mathbb{R}^d : M(\theta) < +\infty \}$, donde

$$M(\theta) = log \int b(x)\ exp\ \{\theta' t(x)\}\ d\mu(x) = -log\ a(\theta). \qquad (5.3)$$

Una familia $\mathcal{F}$ de distribuciones con función de densidad de la forma (5.2) y $\theta \in \Theta$ es llamada una familia exponencial regular si Θ es un conjunto abierto (no vacío) en $\mathbb{R}^d$.

Denote por $\mathcal{C}$ a la familia de distribuciones propias, definidas sobre los Borelianos de Θ, cuya densidad con respecto a la medida de Lebesgue es de la forma

$$p(\theta) = p(\theta; n_0, t_0) = H(n_0, t_0)\ a^{n_0}(\theta)\ exp\ \{\theta' t_0\} \qquad (5.4)$$

donde $n_0 \in \mathbb{R}^+$, $t_0 \in \mathbb{R}^d$ y

$$H(n_0, t_0) = \left\{ \int a^{n_0}(\theta)\ exp\ \{\theta' t_0\}\ d\theta \right\}^{-1},$$

siempre que la integral exista. La familia $\mathcal{C}$ es llamada entonces una familia conjugada de la familia exponencial $\mathcal{F}$. Claramente, $\mathcal{C}$ es también una familia exponencial y es cerrada bajo muestreo.

Sea

$$\nabla M(\theta) = \left[\frac{\partial M(\theta)}{\partial \theta_1}, \ldots, \frac{\partial M(\theta)}{\partial \theta_d}\right]' = \left[M_1(\theta), \ldots, M_d(\theta)\right]'.$$

Entonces, diferenciando la identidad $\int p(x|\theta)\, dx = 1$ con respecto a θ, e intercambiando la integral y la diferencial, se tiene que

$$E[t(X)|\theta] = \nabla M(\theta). \tag{5.5}$$

En estas condiciones, se tiene el siguiente resultado.

TEOREMA 1. (Diaconis & Ylvisaker, 1979). Suponga que Θ es un conjunto abierto en $\mathbb{R}^d$. Si θ tiene una distribución $p(\theta; n_0, t_0)$ en $\mathcal{C}$ y $n_0 > 0$, entonces

$$E[\nabla M(\theta)] = \frac{t_0}{n_0}. \quad \triangleright \tag{5.6}$$

Ahora, sea $W(n,t) = -log\ H(n,t)$. Defínase

$$\nabla W(n,t) = \left[\frac{\partial W(n,t)}{\partial t_1}, \ldots, \frac{\partial W(n,t)}{\partial t_d}\right]' = \left[W_1(n,t), \ldots, W_d(n,t)\right]'$$

y

$$W_0(n,t) = \frac{\partial W(n,t)}{\partial n}.$$

El siguiente resultado muestra que el valor esperado de la divergencia logarítmica de Kullback-Leibler, definida en la Sección 3.3, toma una forma simple en estas condiciones.

TEOREMA 2. (Gutiérrez, 1991). Bajo las condiciones del Teorema 1,

$$E[\delta(\theta:\theta_0)] = W_0(n_0,t_0) + M(\theta_0) + \frac{1}{n_0}\left\{ d + \left[\nabla W(n_0,t_0) - \theta_0\right]' t_0 \right\} \tag{5.7}$$

DEMOSTRACION. Si $p(x|\theta)$ es una distribución en $\mathcal{F}$, entonces la divergencia logarítmica está dada por

$$\delta(\theta:\theta_0) = \int p(x|\theta)\ log\ \frac{p(x|\theta)}{p(x|\theta_0)}\ dx$$

$$= M(\theta_0) - M(\theta) + (\theta - \theta_0)'E[t(x)|\theta]$$

$$= M(\theta_0) - M(\theta) + (\theta - \theta_0)'\nabla M(\theta)$$

por (5.5). En vista de esto,

$$E[\delta(\theta:\theta_0)] = M(\theta_0) - E[M(\theta)] + E[\theta'\nabla M(\theta)] - \theta_0' E[\nabla M(\theta)]. \tag{5.8}$$

Recuerde que, de acuerdo al Teorema 1,

$$E[\nabla M(\theta)] = \frac{t_0}{n_0}\ .$$

Por otra parte, al diferenciar la identidad

$$log \int a^n(\theta)\ exp\ \{\theta' t\}\ d\theta = W(n,t)$$

en n, se tiene

$$\int log\ a(\theta)\ H(n,t)\ a^n(\theta)\ exp\ \{\theta' t\}\ d\theta = W_0(n,t),$$

por lo que

$$E[M(\theta)] = - W_0(n_0, t_0). \qquad (5.9)$$

Ahora, de la ecuación (5.6),

$$log \int M_i(\theta)\ a^n(\theta)\ exp\ \{\theta' t\}\ d\theta = log\ t_i - log\ H(n,t) - log\ n$$

para i=1,2,...,d. Si esta identidad es diferenciada en t_i y se intercambian la diferencial y la integral, se sigue que

$$\int \theta_i\ M_i(\theta)\ H(n,t)\ a^n(\theta)\ exp\ \{\theta' t\}\ d\theta = \frac{1}{n}\left[\ 1 + W_i(n,t)\ t_i\ \right]$$

(i=1,2,...,d). Por lo tanto

$$E[\theta' \nabla M(\theta)] = \frac{1}{n_0}\left\{\ d + \nabla W(n_0, t_0)' t_0\ \right\}, \qquad (5.10)$$

y el resultado se obtiene al sustituir (5.7), (5.9) y (5.10) en la expresión (5.8). ▷

3.5.2 Estimacion

Como se mencionó en la Sección 3.4, el problema de estimación puede ser visto como un problema de decisión en el que el espacio de decisiones y el espacio paramétrico coinciden, siempre que el espacio paramétrico no se restrinja. Las funciones de utilidad más comunes incluyen, cuando se estima un parámetro real θ, a la utilidad cuadrática

$$U_c(\hat{\theta}, \theta) = - (\hat{\theta} - \theta)^2$$

y a la utilidad del error absoluto

$$U_a(\hat{\theta}, \theta) = - |\hat{\theta} - \theta| .$$

El uso de la utilidad cuadrática produce como estimación óptima a la media de la distribución de θ, si dicho valor está en el soporte de $p(\theta)$. Por otra parte, una mediana de la distribución de θ resulta ser la mejor estimación cuando se utiliza la utilidad del error absoluto (*e.g.* DeGroot, 1970).

Sea $Z = \{X_1, X_2, \ldots, X_n\}$ un conjunto de observaciones de la variable aleatoria X. Si la distribución inicial de θ pertenece a la familia conjugada $\mathfrak{C}$, entonces la distribución final estará dada por

$$p(\theta|z) = p(\theta; n_1, t_1) = H(n_1, t_1)\, a^{n_1}(\theta)\, exp\{\theta' t_1\}, \qquad (5.11)$$

donde $n_1 = n_0 + n$ y $t_1 = t_0 + \sum_{i=1}^{n} t(x_i)$.

De acuerdo con lo expuesto en la Sección 3.4, considere la función de utilidad

$$U'(\hat{\theta}, \theta) = -\delta(\theta : \hat{\theta}),$$

donde $\delta(\theta:\hat{\theta})$ denota ahora a la divergencia logarítmica de Kullback-Leibler. Entonces, por el Teorema 2,

$$\begin{aligned} \Delta(\hat{\theta}) &= E[\delta(\theta:\hat{\theta})|z] \\ &= W_0(n_1, t_1) + M(\hat{\theta}) + \frac{1}{n_1}\left\{ d + \left[\nabla W(n_1, t_1) - \hat{\theta}\right]' t_1 \right\} \end{aligned}$$

Puede demostrarse que $\Delta(\hat{\theta})$ es una función convexa de $\hat{\theta}$. Por otro lado, puede comprobarse fácilmente que el valor $\hat{\theta}^*$ de $\hat{\theta}$ que minimiza $\Delta(\hat{\theta})$ se encuentra resolviendo la ecuación

$$\nabla M(\hat{\theta}) = \frac{t_1}{n_1}. \qquad (5.12)$$

Compare éste con el estimador de máxima verosimilitud para este caso, $\hat{\theta}_{ML}$, que se obtiene al resolver la ecuación

$$\nabla M(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} t(x_i)$$

(Cox & Hinkley, 1974; Capítulo 9).

Por otro lado,

$$\frac{\partial \log p(\theta|z)}{\partial \theta} = t_1 - n_1 \nabla M(\theta),$$

de donde se sigue que la moda de la distribución final de θ también puede encontrarse al resolver la ecuación (5.12). En otras palabras, $\hat{\theta}^*$ y la moda de la distribución final de θ coinciden. El estimador $\hat{\theta}^*$ es llamado también el estimador de máxima verosimilitud generalizado (DeGroot, 1970; Sección 11.4). Finalmente, observe que si $n_0 = t_0 = 0$, entonces $\hat{\theta}_{ML}$ y $\hat{\theta}^*$ son iguales.

3.5.3 Contraste de Hipotesis

Si la distribución de X pertenece a una familia exponencial, el procedimiento general de contraste de hipótesis paramétricas propuesto en la Sección 3.2 produce una regla de decisión cuya forma es sencilla. Como se verá en esta sección, a nivel operativo dicha regla es análoga a la regla producida por el método generalizado de cociente de verosimilitudes.

Sea X una variable aleatoria m-variada con función de densidad $p(x|\theta)$, perteneciente a una familia exponencial regular $\mathcal{F}$ y sea $Z = \{X_1, X_2, \ldots, X_n\}$ un conjunto de observaciones de X.

Suponga que se desea contrastar las hipótesis

$$H_0:\theta \in \Theta_0 \quad \text{vs.} \quad H_1:\theta \in \Theta_1$$

donde $\Theta_0 \subset \Theta$ y $\Theta_1 \subset \Theta$ son tales que $\Theta_0 \cap \Theta_1 = \emptyset$.

De acuerdo a lo expuesto en la Sección 3.2, la acción d_1 (aceptar H_1) será preferible a la acción d_0 (aceptar H_0) si y sólo si

$$E[\delta(\theta:\theta_0^*)|z] - E[\delta(\theta:\theta_1^*)|z] > \delta^*,$$

donde $\delta^* = (B_0-B_1)/A$, $\theta_i^* \in \Theta_i$ es tal que

$$E[\delta(\theta:\theta_i^*)|z] = \inf \left\{ E[\delta(\theta:\theta_i)|z] : \theta_i \in \Theta_i \right\} \quad (i=0,1)$$

y δ denota en este caso a la divergencia logarítmica de Kullback-Leibler.

Por el Teorema 2, se tiene entonces que debe elegirse d_1 si y sólo si

$$M(\theta_0^*) - M(\theta_1^*) + \frac{1}{n_1} \left[\theta_1^* - \theta_0^* \right]' t_1 > \delta^*. \qquad (5.13)$$

Por otra parte, suponga ahora que se elige d_1 si y sólo si

$$\frac{\sup \{p(\theta|z) : \theta \in \Theta_0\}}{\sup \{p(\theta|z) : \theta \in \Theta_1\}} < \exp\{-n_1\delta^*\}. \qquad (5.14)$$

Como se mencionó en la sección anterior, en el caso de las familias exponenciales regulares minimizar la divergencia logarítmica esperada con respecto a θ equivale a maximizar $p(\theta|z)$. Este argumento permite demostrar fácilmente que la regla de decisión dada por la desigualdad anterior equivale a elegir d_1 si y sólo si

$$log\ p(\theta_1^*|z) - log\ p(\theta_0^*|z) > n_1\delta^*$$

o, equivalentemente, siempre y cuando

$$n_1\,(M(\theta_0^*) - M(\theta_1^*)) + (\theta_1^* - \theta_0^*)'t_1 > n_1\delta^*.$$

Por lo tanto, como puede observarse, la regla de decisión dada por (5.14) coincide con la regla (5.13), producida por el procedimiento propuesto en la Sección 3.2.

Compare esta regla de decisión con la regla generada a través del método generalizado de cociente de verosimilitudes (Sección 1.2), cuya región crítica para este caso es de la forma

$$C = \left\{ z \in Z : M(\hat{\theta}_0) - M(\hat{\theta}_1) + \frac{1}{n_1}\left[\hat{\theta}_1 - \hat{\theta}_0\right]'t_1 > k^* \right\},$$

donde $\hat{\theta}_0$ y $\hat{\theta}_1$ son los estimadores de máxima verosimilitud restringidos a Θ_0 y Θ_1, respectivamente, y k^* es una constante.

Claramente, si $n_0 = t_0 = 0$ entonces la regla de decisión dada por (5.13) es equivalente a la regla que proporciona el método generalizado de cociente de verosimilitudes, siempre que $\delta^* = k^*$.

3.6 Parametros de Ruido

Como se mencionó en la Sección 2.4, una situación común en la práctica surge cuando el parámetro es de la forma (θ,ω), donde θ denota al parámetro de interés y ω es un parámetro de ruido. El procedimiento de contraste desarrollado en la Sección 3.2, sin embargo, supone que todos los parámetros en el modelo son de interés, de manera que no considera la existencia de parámetros de ruido. Por lo tanto, en esta sección se describirá la forma de instrumentar dicho procedimiento cuando existen parámetros de

ruido, extendiéndose algunos de los resultados obtenidos en la sección anterior. A lo largo de esta sección el espacio parametral es de la forma $\Theta \times \Omega$, donde Θ y Ω son conjuntos abiertos de $\mathbb{R}^d$ y $\mathbb{R}^r$, respectivamente.

Sea X una variable aleatoria m-variada con función de densidad $p(x|\theta,\omega)$, perteneciente a una familia paramétrica

$$\mathcal{F} = \left\{ p(x|\theta,\omega) : \theta \in \Theta,\ \omega \in \Omega \right\} \qquad \Theta \subseteq \mathbb{R}^d,\quad \Omega \subseteq \mathbb{R}^r$$

y sea $Z = \{X_1, X_2, \ldots, X_n\}$ un conjunto de observaciones de X. Suponga que, independientemente del valor de ω, se desea contrastar las hipótesis

$$H_0: \theta \in \Theta_0 \quad \text{vs.} \quad H_1: \theta \in \Theta_1,$$

donde $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ y $\Theta_0 \cap \Theta_1 = \varnothing$.

Estas hipótesis son matemáticamente equivalentes a las hipótesis

$$H_0': p(x|\theta,\omega) \in \mathcal{F}_0 \quad \text{vs.} \quad H_1': p(x|\theta,\omega) \in \mathcal{F}_1,$$

donde ahora

$$\mathcal{F}_i = \left\{ p(x|\theta,\omega) : \theta \in \Theta_i,\ \omega \in \Omega \right\} \qquad (i=0,1).$$

Denote por $p(\theta,\omega)$ a la distribución inicial de (θ,ω). Por el Teorema de Bayes, la distribución final correspondiente está dada entonces por

$$p(\theta,\omega|z) \propto p(\theta,\omega)\ p(z|\theta,\omega),$$

donde ahora $p(z|\theta,\omega)$ denota a la función de verosimilitud.

Como antes, sea d_i la decisión correspondiente a la aceptación de la hipótesis H_i $(i=0,1)$ y denote por $U(d_i;\theta,\omega)$ a la utilidad de la decisión d_i cuando (θ,ω) es el valor del parámetro. En este caso, por lo tanto, la utilidad esperada final de d_i es

$$U_E(d_i|z) = \int\int U(d_i;\theta,\omega)\ p(\theta,\omega|z)\ d\theta\ d\omega \qquad (i=0,1). \qquad (6.1)$$

Ahora, para un valor fijo θ^* de θ, denote por $\delta(\theta,\omega:\theta^*,\omega)$ a una medida de la discrepancia entre $p(x|\theta,\omega)$ y $p(x|\theta^*,\omega)$. Entonces

$$E[\delta(\theta,\omega:\theta^*,\omega)|z] = \int\int \delta(\theta,\omega:\theta^*,\omega)\ p(\theta,\omega|z)\ d\theta\ d\omega.$$

De acuerdo a las ideas expuestas en la Sección 3.2, se propone para este caso una función de utilidad de la forma

$$U(d_i;\theta,\omega) = B_i - A\ \delta(\theta,\omega:\theta_i^*,\omega) \qquad A \in \mathbb{R}^+,\ B_i \in \mathbb{R}, \qquad (6.2)$$

donde $\theta_i^* \in \Theta_i$ es tal que

$$E[\delta(\theta,\omega:\theta_i^*,\omega)|z] = inf\left\{ E[\delta(\theta,\omega:\theta_i,\omega)|z] : \theta_i \in \Theta_i \right\} \qquad (6.3)$$

$(i=0,1)$. De esta manera, se tiene que la mejor decisión es d_i si y sólo si

$$E[\delta(\theta,\omega:\theta_0^*,\omega)|z] - E[\delta(\theta,\omega:\theta_1^*,\omega)|z] > \delta^*, \qquad (6.4)$$

donde $\delta^* = (B_0 - B_1)/A$.

Recuerde que, al plantear el problema de estimación de θ como un problema de decisión, resulta conveniente utilizar la función de utilidad

$$U'(\hat{\theta};\theta,\omega) = -\delta(\theta,\omega:\hat{\theta},\omega),$$

de manera que el valor $\hat{\theta}^{\bullet}$ de $\hat{\theta}$ que maximiza la utilidad esperada final $U'_E(\hat{\theta}|z)$ es tal que

$$E[\delta(\theta,\omega:\hat{\theta}^{\bullet},\omega)|z] = \inf \left\{ E[\delta(\theta,\omega:\hat{\theta},\omega)|z] : \hat{\theta} \in \Theta \right\}.$$

Esta última afirmación, tal como en la Sección 3.4, permite justificar la elección del valor $\theta_i^{\bullet} \in \Theta_i$ $(i=0,1)$ definido en la expresión (6.3). Cabe mencionar que, aunque en este caso se tienen parámetros de ruido, el estimador $\hat{\theta}^{\bullet}$ obtenido de esta forma también es invariante ante transformaciones uno a uno de θ.

Finalmente, y para ilustrar lo que se ha expuesto a lo largo de esta sección, se revisará el caso en el que la distribución $p(x|\theta,\omega)$ de X pertenece a una familia exponencial regular y $\delta(\theta,\omega:\theta_0,\omega)$ es la divergencia logarítmica de Kullback-Leibler.

Sea X es una variable aleatoria m-variada con función de densidad de la forma

$$p(x|\theta,\omega) = a(\theta,\omega)\, b(x)\, exp\, \{\theta' t(x) + \omega' s(x)\}. \qquad (6.5)$$

y suponga que el espacio parametral

$$\Pi = \left\{ (\theta,\omega) \in \mathbb{R}^{d+r} : M(\theta,\omega) < +\infty \right\}$$

es un conjunto abierto (no vacío) en $\mathbb{R}^{d+r}$, donde

$$M(\theta,\omega) = - log\, a(\theta,\omega). \qquad (6.6)$$

Por lo tanto, la familia $\mathcal{F}$ de las distribuciones cuya función de densidad es de la forma (6.5), con $(\theta,\omega) \in \Theta \times \Omega$, es una familia exponencial regular. Como se mencionó al principio de esta sección, se supondrá que Π puede expresarse como $\Pi = \Theta \times \Omega$, donde Θ y Ω son conjuntos abiertos de $\mathbb{R}^d$ y $\mathbb{R}^r$ respectivamente.

Ahora, sean

$$\nabla M_1(\theta,\omega) = \frac{\partial M(\theta,\omega)}{\partial \theta}$$

$$\nabla M_2(\theta,\omega) = \frac{\partial M(\theta,\omega)}{\partial \omega} \quad .$$

Entonces, por la expresión (5.5), se tiene que

$$E[t(X)|\theta,\omega] = \nabla M_1(\theta,\omega)$$

$$E[s(X)|\theta,\omega] = \nabla M_2(\theta,\omega) \quad .$$

Si $p(x|\theta,\omega)$ pertenece a la familia $\mathcal{F}$, entonces la divergencia logarítmica entre $p(x|\theta,\omega)$ y $p(x|\theta_0,\omega)$ está dada por

$$\begin{aligned}\delta(\theta,\omega:\theta_0,\omega) &= \int p(x|\theta,\omega)\ log\ \frac{p(x|\theta,\omega)}{p(x|\theta_0,\omega)}\ dx \\ &= M(\theta_0,\omega) - M(\theta,\omega) + (\theta - \theta_0)'E[t(X)|\theta,\omega] \\ &= M(\theta_0,\omega) - M(\theta,\omega) + (\theta - \theta_0)'\nabla M_1(\theta,\omega).\end{aligned} \tag{6.7}$$

Por otro lado, denote por $\mathcal{C}$ a la familia de las distribuciones propias definidas sobre los Borelianos de $\Theta \times \Omega$ cuya densidad con respecto a la medida de Lebesgue es de la forma

$$p(\theta,\omega) = p(\theta,\omega;n_0,t_0,s_0) = H(n_0,t_0,s_0)\ a^{n_0}(\theta,\omega)\ exp\ \{\theta't_0 + \omega's_0\}$$

donde $n_0 \in \mathbb{R}^+$, $t_0 \in \mathbb{R}^d$, $s_0 \in \mathbb{R}^r$ y

$$H(n_0,t_0,s_0) = \left\{ \int\int a^{n_0}(\theta,\omega)\ exp\ \{\theta't_0 + \omega's_0\}\ d\theta\ d\omega \right\}^{-1},$$

si la integral existe. Como antes, la familia $\mathcal{C}$ es una familia conjugada de la familia exponencial $\mathcal{F}$ y es también una familia exponencial.

Ahora, por el Teorema 1 de la sección anterior, si (θ,ω) tiene una distribución $p(\theta,\omega;n_0,t_0,s_0)$ en $\mathcal{C}$ y $n_0 > 0$, entonces

$$E[\nabla M_1(\theta,\omega)] = \frac{t_0}{n_0}$$

$$E[\nabla M_2(\theta,\omega)] = \frac{s_0}{n_0} \quad . \tag{6.8}$$

Al igual que en la Sección 3.5, este resultado permite encontrar la forma del valor esperado de la divergencia logarítmica de Kullback-Leibler para este caso.

Sea $W(n,t,s) = -\log H(n,t,s)$ y defina

$$\nabla W_1(n,t,s) = \frac{\partial W(n,t,s)}{\partial t}$$

$$\nabla W_2(n,t,s) = \frac{\partial W(n,t,s)}{\partial s}$$

y

$$W_0(n,t,s) = \frac{\partial W(n,t,s)}{\partial n} \quad .$$

TEOREMA 2'. Suponga que $\Theta \times \Omega$ es un conjunto abierto en $\mathbb{R}^{d+r}$. Si (θ,ω) tiene una distribución $p(\theta,\omega;n_0,t_0,s_0)$ en $\mathcal{C}$ y $n_0 > 0$, entonces

$$E[\delta(\theta,\omega:\theta_0,\omega)] = W_0(n_0,t_0,s_0) + R(\theta_0;n_0,t_0,s_0)$$

$$+ \frac{1}{n_0}\left\{ d + \left[\nabla W_1(n_0,t_0,s_0) - \theta_0\right]' t_0 \right\} , \tag{6.9}$$

donde

$$R(\theta_0;n_0,t_0,s_0) = E[M(\theta_0,\omega)]$$

$$= \int M(\theta_0,\omega)\ \frac{H(n_0,t_0,s_0)}{G(n_0,t_0,\omega)}\ \ exp\ \{\omega' s_0\}\ d\omega$$

con

$$G(n_0,t_0,\omega) = \left\{ \int a^{n_0}(\theta,\omega)\ exp\ \{\theta' t_0\}\ d\theta \right\}^{-1}.$$

DEMOSTRACION. De acuerdo con (6.7),

$$E[\delta(\theta,\omega:\theta_0,\omega)] = E[M(\theta_0,\omega)] - E[M(\theta,\omega)]$$

$$+ E[\theta' \nabla M_1(\theta,\omega)] - \theta_0' E[\nabla M_1(\theta,\omega)] \qquad (6.10)$$

De la expresión (5.9) se tiene que

$$E[M(\theta,\omega)] = - W_0(n_0,t_0,s_0), \qquad (6.11)$$

mientras que, de la expresión (5.10),

$$E[\theta' \nabla M_1(\theta,\omega)] = \frac{1}{n_0} \left\{ d + \nabla W_1(n_0,t_0,s_0)' t_0 \right\}. \qquad (6.12)$$

Finalmente

$$E[M(\theta_0,\omega)] = R(\theta_0;n_0,t_0,s_0) = \int M(\theta_0,\omega)\ p(\omega)\ d\omega, \qquad (6.13)$$

donde $p(\omega)$ denota a la densidad marginal de ω. Observe que

$$p(\omega) = \int p(\theta,\omega)\ d\omega$$

$$= \int H(n_0,t_0,s_0)\ a^{n_0}(\theta,\omega)\ exp\ \{\theta' t_0 + \omega' s_0\}\ d\theta$$

$$= H(n_0,t_0,s_0)\ exp\ \{\omega' s_0\} \int a^{n_0}(\theta,\omega)\ exp\ \{\theta' t_0\}\ d\theta$$

$$= \frac{H(n_0,t_0,s_0)}{G(n_0,t_0,\omega)}\ exp\ \{\omega' s_0\}\ .$$

Por lo tanto

$$R(\theta_0;n_0,t_0,s_0) = \int M(\theta_0,\omega)\ \frac{H(n_0,t_0,s_0)}{G(n_0,t_0,\omega)}\ exp\ \{\omega' s_0\}\ d\omega\ ,$$

de manera que la expresión (6.9) se obtiene al sustituir las expresiones (6.8), (6.11), (6.12) y (6.13) en la expresión (6.10). ▷

Observe que, dada una muestra $Z = \{X_1, X_2, \ldots, X_n\}$ de $p(x|\theta,\omega)$, la distribución final de (θ,ω) esta dada por

$$p(\theta,\omega|z) = p(\theta,\omega;n_1,t_1,s_1)$$

$$= H(n_1,t_1,s_1)\ a^{n_1}(\theta,\omega)\ exp\ \{\theta' t_1 + \omega' s_1\}$$

donde $n_1 = n_0 + n$, $t_1 = t_0 + \sum_{i=1}^{n} t(x_i)$ y $s_1 = s_0 + \sum_{i=1}^{n} s(x_i)$.

En estas condiciones, el estimador $\hat{\theta}^*$ de $\hat{\theta}$, obtenido al minimizar la divergencia logarítmica esperada final se encuetra resolviendo la ecuación

$$R(\hat{\theta};n_1,t_1,s_1) = \frac{t_1}{n_1}\ ,$$

mientras que el estimador de máxima verosimilitud para este caso se encuentra al resolver el sistema de ecuaciones

$$\nabla M_1(\hat{\theta},\hat{\omega}) = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$$

$$\nabla M_2(\hat{\theta},\hat{\omega}) = \frac{1}{n}\sum_{i=1}^{n} s(x_i)$$

obteniéndose, además, un estimador del parámetro de ruido ω.

Finalmente, si se desea contrastar las hipótesis

$$H_0:\theta \in \Theta_0 \quad \text{vs.} \quad H_1:\theta \in \Theta_1,$$

la expresión (6.4) indica que debe elegirse d_1 (aceptar H_1) si y solo si

$$E[\delta(\theta,\omega:\theta_0^*,\omega)|z] - E[\delta(\theta,\omega:\theta_1^*,\omega)|z] > \delta^*,$$

donde $\theta_i^* \in \Theta_i$ es tal que

$$E[\delta(\theta,\omega:\theta_i^*,\omega)|z] = inf\left\{ E[\delta(\theta,\omega:\theta_i,\omega)|z] : \theta_i \in \Theta_i \right\}$$

(i=0,1). En otras palabras, en este caso debe elegirse d_1 siempre y cuando

$$R(\theta_0^*;n_1,t_1,s_1) - R(\theta_1^*;n_1,t_1,s_1) + \frac{1}{n_1}\left[\theta_1^* - \theta_0^*\right]' t_1 > \delta^*.$$

Observe que, a diferencia del caso en el que no se tienen parámetros de ruido, la forma de la solución bayesiana no coincide con la forma de la solución clásica correspondiente, tanto en el problema de estimación como en el de contraste de hipótesis. Esto se debe a que el tratamiento que se le da a los parámetros de

ruido en cada uno de estos enfoques es distinto. En efecto, mientras en el enfoque frecuentista la solución consiste en estimar el valor del parámetro de ruido ω (generalmente utilizando el método de máxima verosimilitud), en el enfoque bayesiano basta integrar con respecto a la distribución marginal de ω.

4. APLICACION

Como una aplicación importante, en este capítulo se analizará el problema de inferencia en el modelo de regresión lineal múltiple, utilizando el procedimiento propuesto en el capítulo anterior en la estimación de los coeficientes, así como en el contraste de la hipótesis lineal general. Una exposición completa del análisis clásico de este problema puede encontrarse, por ejemplo, en el libro de Seber (1977).

4.1 PLANTEAMIENTO

Sea Y un vector aleatorio en $\mathbb{R}^n$ con media μ y matriz de precisión Σ. Suponga que Y sigue una distribución Normal Multivariada con

$$\mu = X\beta \qquad y \qquad \Sigma = hI_n \ ,$$

donde X es una matriz (nxk) de rango k (formada por constantes conocidas), I_n denota a la matriz identidad, β es vector de dimensión k (formado por constantes desconocidas) y $h>0$ es una constante desconocida. Se tiene entonces que la función de densidad de Y está dada por

$$p(y|\beta,h) = \left[\frac{h}{2\pi} \right]^{n/2} exp \left\{ - \frac{h}{2} (y - X\beta)'(y - X\beta) \right\} \ ,$$

que puede reescribirse de manera más conveniente como

$$p(y|\beta,h) = \left[\frac{h}{2\pi}\right]^{n/2} exp\left\{-\frac{h}{2}\left[\frac{n-k}{\hat{h}} + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]\right\},$$

donde

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{h} = (n-k)\ \{(y - \hat{y})'(y - \hat{y})\}^{-1}, \qquad (1.1)$$

y $\hat{y} = X\hat{\beta}$.

Dado que el parámetro de interés en este caso es β, es necesario calcular la divergencia logarítmica entre $p(y|\beta,h)$ y $p(y|\beta_0,h)$, considerando a h como un parámetro de ruido.

RESULTADO.

$$\delta(\beta,h:\beta_0,h) = \int p(y|\beta,h)\ log\ \frac{p(y|\beta,h)}{p(y|\beta_0,h)}\ dy$$

$$= \frac{h}{2}(\beta - \beta_0)'X'X(\beta - \beta_0) \qquad (1.2)$$

DEMOSTRACION.

$$log\ p(y|\beta,h) = \frac{n}{2}\ log\ \frac{h}{2} - \frac{h}{2}\left[\frac{n-k}{\hat{h}} + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]$$

Por lo tanto,

$$log\ \frac{p(y|\beta,h)}{p(y|\beta_0,h)} = \frac{h}{2}\left[(\beta_0 - \hat{\beta})'X'X(\beta_0 - \hat{\beta}) - (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]$$

$$= \frac{h}{2}\left[\beta_0'X'X\beta_0 - \beta'X'X\beta + 2(\beta - \beta_0)'X'y\right],$$

de donde

$$\delta(\beta,h:\beta_0,h) = \frac{h}{2}\left[\beta_0'X'X\beta_0 - \beta'X'X\beta + 2(\beta-\beta_0)'X'E[y|\beta,h]\right]$$

$$= \frac{h}{2}\left[\beta_0'X'X\beta_0 - \beta'X'X\beta + 2(\beta-\beta_0)'X'X\beta\right]$$

$$= \frac{h}{2}\left[\beta'X'X\beta - 2\beta_0'X'X\beta + \beta_0'X'X\beta_0\right]$$

$$= \frac{h}{2}(\beta-\beta_0)'X'X(\beta-\beta_0). \quad \triangleright$$

Ahora, para poder calcular el valor esperado de $\delta(\beta,h:\beta_0,h)$, es necesario asignar una distribución inicial para (β,h). La distribución conjugada natural para este caso es

$$p(\beta,h) = N_k(\beta|\mu,h\Sigma)\ Ga(h|\tfrac{a+2}{2},\tfrac{b}{2}), \qquad (1.3)$$

donde $N_k(\beta|\mu,h\Sigma)$ denota a una densidad Normal k-variada con media μ y matriz de precisión $h\Sigma$ y $Ga(h|\frac{a+2}{2},\frac{b}{2})$ denota a una densidad Gama con media $\frac{a+2}{b}$ y varianza $\frac{2}{b}(\frac{a+2}{b})$. Por lo tanto, la distribución final para (β,h) está dada por

$$p(\beta,h|y) = N_k(\beta|\mu_1,h\Sigma_1)\ Ga(h|\tfrac{a1+2}{2},\tfrac{b1}{2}), \qquad (1.4)$$

donde

$$\mu_1 = (X'X+\Sigma)^{-1}(X'y+\Sigma\mu)$$

$$\Sigma_1 = X'X+\Sigma \qquad (1.5)$$

$$a_1 = n+a$$

y

$$b_1 = b + \mu'\Sigma\mu + y'y - (X'y + \Sigma\mu)'(X'X + \Sigma)^{-1}(X'y + \Sigma\mu)$$

(Broemeling, 1985; Capítulo 8). Cabe mencionar que b_1 puede reescribirse de manera más conveniente como

$$b_1 = b + y'[I_n - X(X'X + \Sigma)^{-1}X']y + \mu'[\Sigma + \Sigma(X'X + \Sigma)^{-1}\Sigma]\mu - 2\mu\Sigma\mu_1 \qquad (1.6)$$

$$= b + y'[I_n - X(X'X + \Sigma)^{-1}X']y + R(y,\mu,\Sigma).$$

Para facilitar el cálculo del valor esperado de $\delta(\beta,h:\beta_0,h)$, observe que

$$\delta(\beta,h:\beta_0,h) = \frac{h}{2}\left[(\beta - \mu_1 + \mu_1 - \beta_0)'X'X(\beta - \mu_1 + \mu_1 - \beta_0)\right]$$

$$= \frac{h}{2}\left[(\beta - \mu_1)'X'X(\beta - \mu_1) + (\mu_1 - \beta_0)'X'X(\mu_1 - \beta_0)\right.$$

$$\left. + 2(\beta - \mu_1)'X'X(\mu_1 - \beta_0)\right]$$

Por lo tanto

$$E[\ \delta(\beta,h:\beta_0,h)\ |\ y\ ] = \frac{1}{2}E[\ h\ (\beta - \mu_1)'X'X(\beta - \mu_1)\ |\ y\ ]$$

$$+ \frac{1}{2}E[\ h\ |\ y\ ]\ (\mu_1 - \beta_0)'X'X(\mu_1 - \beta_0) \qquad (1.7)$$

$$= \frac{k}{2} + \frac{a_1+2}{2b_1}(\mu_1 - \beta_0)'X'X(\mu_1 - \beta_0).$$

4.2 Estimacion de los Coeficientes de Regresion

Sea

$$\Delta(\tilde{\beta}) = E[\ \delta(\beta,h:\tilde{\beta},h)\ |\ y\ ]$$

$$= \frac{k}{2} + \frac{a_1+2}{2b_1}(\mu_1 - \tilde{\beta})'X'X(\mu_1 - \tilde{\beta}).$$

Recuerde que al utilizar $U'(\tilde{\beta};\beta,h) = -\delta(\beta,h:\tilde{\beta},h)$ como función de utilidad en el problema de estimación, la decisión óptima es el valor $\tilde{\beta}^*$ de $\tilde{\beta}$ que minimiza a $\Delta(\tilde{\beta})$.

Observe que

$$\frac{\partial\,\Delta(\tilde{\beta})}{\partial\,\tilde{\beta}} = -\frac{a_1+2}{b_1}X'X(\mu_1 - \tilde{\beta}),$$

por lo que $\dfrac{\partial\,\Delta(\tilde{\beta})}{\partial\,\tilde{\beta}}$ se anula en

$$\tilde{\beta}^* = \mu_1. \qquad (2.1)$$

Por lo tanto

$$\Delta(\mu_1) = inf\ \{\ \Delta(\tilde{\beta})\ :\ \tilde{\beta} \in \mathbb{R}^k\ \}.$$

4.3 Contraste de la Hipotesis Lineal General

Sea C una matriz (rxk) de rango r ($1 \le r \le k$) y sea $\gamma \in \mathbb{R}^r$, formados por constantes conocidas. Suponga que se desea contrastar las hipótesis

$$H_0: C\beta = \gamma \quad \text{vs.} \quad H_1: C\beta \neq \gamma.$$

De acuerdo con lo expuesto en el capítulo anterior, debe rechazarse H_0 si y sólo si

$$\Delta(\tilde{\beta}_0^*) - \Delta(\tilde{\beta}_1^*) > \delta^*,$$

donde

$$\Delta(\tilde{\beta}_0^*) = inf \{ \Delta(\tilde{\beta}_0) : C\tilde{\beta}_0 = \gamma \}$$

y

$$\Delta(\tilde{\beta}_1^*) = inf \{ \Delta(\tilde{\beta}_1) : C\tilde{\beta}_1 \neq \gamma \},$$

con δ^* una constante.

Para hallar $\tilde{\beta}_0^*$ se utilizarán multiplicadores de Lagrange. Sea $\lambda \in \mathbb{R}^r$ y defina

$$L(\tilde{\beta},\lambda) = \Delta(\tilde{\beta}) + \lambda(C\tilde{\beta} - \gamma).$$

Se tiene entonces que

$$\frac{\partial L(\tilde{\beta},\lambda)}{\partial \tilde{\beta}} = - \frac{a_1+2}{b_1} X'X(\mu_1 - \tilde{\beta}) + C'\lambda$$

$$\frac{\partial L(\tilde{\beta},\lambda)}{\partial \lambda} = C\beta - \gamma .$$

Al igualar a cero estas expresiones, se tiene

$$X'X\tilde{\beta}_0^* + C'\lambda^* = X'X\mu_1 \qquad (3.1)$$

$$C\tilde{\beta}_0^* = \gamma .$$

Puede verse fácilmente que la solución a este sistema de ecuaciones está dada por

$$\tilde{\beta}_0^* = \mu_1 - (X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}[C\mu_1 - \gamma]$$

$$\lambda^* = C'[C(X'X)^{-1}C']^{-1}[C\mu_1 - \gamma] \ .$$

Por lo tanto

$$\begin{aligned} \Delta(\tilde{\beta}_0^*) &= \frac{k}{2} + \frac{a_1+2}{2b_1}(\mu_1 - \tilde{\beta}_0^*)'X'X(\mu_1 - \tilde{\beta}_0^*) \\ &= \frac{k}{2} + \frac{a_1+2}{2b_1}[C\mu_1 - \gamma]'[C(X'X)^{-1}C']^{-1}[C\mu_1 - \gamma] \ . \end{aligned} \tag{3.2}$$

Por otro lado, para determinar $\tilde{\beta}_1^*$ observe que

$$inf\ \{\ \Delta(\tilde{\beta}_1) : C\tilde{\beta}_1 \neq \gamma\ \} = inf\ \{\ \Delta(\tilde{\beta}) : \tilde{\beta} \in \mathbb{R}^k\ \} \ ,$$

de manera que, por (2.1),

$$\tilde{\beta}_1^* = \tilde{\beta}^* = \mu_1 \ .$$

Finalmente

$$\Delta(\tilde{\beta}_1^*) = \frac{k}{2} \ . \tag{3.3}$$

Se tiene entonces que

$$\Delta(\tilde{\beta}_0^*) - \Delta(\tilde{\beta}_1^*) = \frac{a_1+2}{2b_1}[C\mu_1 - \gamma]'[C(X'X)^{-1}C']^{-1}[C\mu_1 - \gamma]$$

$$= \frac{n+a+2}{2} \quad \frac{[C\mu_1 - \gamma]'[C(X'X)^{-1}C']^{-1}[C\mu_1 - \gamma]}{b + y'[I_n - X(X'X + \Sigma)^{-1}X']y + R(y,\mu,\Sigma)}$$

por (1.6). Por lo tanto, la hipótesis H_0 debe rechazarse si y sólo si

$$\frac{n+a+2}{2} \quad \frac{[C\mu_1 - \gamma]'[C(X'X)^{-1}C']^{-1}[C\mu_1 - \gamma]}{b + y'[I_n - X(X'X + \Sigma)^{-1}X']y + R(y,\mu,\Sigma)} > \delta^*. \qquad (3.4)$$

4.4 Analisis de Referencia

Si se utiliza la distribución inicial de referencia

$$\pi(\beta,h) \propto h^{-1},$$

entonces la distribución final de referencia para (β,h) está dada por

$$\pi(\beta,h|y) = N_k(\beta|\hat{\beta},hX'X)\ Ga(h|\tfrac{n-k}{2}\ ,\tfrac{n-k}{2}\ \hat{h}^{-1})$$

(Box & Tiao, 1973; Sección 1.3).

Observe que si en la expresiones (1.5) y (1.6) se asignan los valores

$$\Sigma = 0, \quad a = -(k+2) \quad y \quad b = 0,$$

entonces

$$\mu_1 = (X'X)^{-1}X'y = \hat{\beta}$$

$$\Sigma_1 = X'X$$

$$a_1 = n-k-2$$

y

$$b_1 = y'[I_n - X(X'X + \Sigma)^{-1}X']y = (n-k)\ \hat{h}^{-1},$$

por lo que, en estas condiciones, la distribución $p(\beta,h|y)$ dada por (1.4) es precisamente $\pi(\beta,h|y)$. De la expresión (1.7) se tiene entonces que

$$E[\ \delta(\beta,h:\beta_0,h)\ |\ y\ ] = \frac{k}{2} + \frac{\hat{h}}{2}(\hat{\beta} - \beta_0)'X'X(\hat{\beta} - \beta_0).$$

Por otro lado, de (2.1) se sigue que el estimador de β que se obtiene al minimizar la divergencia logarítmica esperada final coincide con el estimador de mínimos cuadrados, es decir,

$$\tilde{\beta}^* = \hat{\beta}.$$

Finalmente, al contrastar la hipótesis lineal general

$$H_0: C\beta = \gamma \quad \text{vs.} \quad H_1: C\beta \neq \gamma,$$

la regla de decisión dada por la expresión (3.4) equivale a rechazar la hipótesis H_0 si y sólo si

$$\frac{n-k}{2}\ \frac{[C\hat{\beta} - \gamma]'[C(X'X)^{-1}C']^{-1}[C\hat{\beta} - \gamma]}{y'[I_n - X(X'X)^{-1}X']y} > \delta^*,$$

o, equivalentemente, siempre y cuando

$$\frac{r}{2}\ F(y) > \delta^*,$$

donde

$$F(y) = \frac{n-k}{r} \quad \frac{[C\hat{\beta} - \gamma]'[C(X'X)^{-1}C']^{-1}[C\hat{\beta} - \gamma]}{y'[I_n - X(X'X)^{-1}X']y} .$$

Observe que $F(y)$ es la estadística utilizada en el procedimiento clásico, cuya región de rechazo tiene la forma

$$C = \left\{ y \in \mathbb{R}^n : F(y) > c^* \right\}$$

con c^* una constante (*e.g.* Seber, 1977).

5. COMENTARIOS FINALES

Los procedimientos clásicos de contraste de hipótesis paramétricas se basan fundamentalmente en el concepto de potencia. Como se expuso en la Sección 1.2, cuando no es posible hallar una prueba uniformemente más potente se requieren criterios adicionales para discriminar entre las diversas pruebas que generalmente están disponibles para contrastar una misma hipótesis. De cualquier manera, el procedimiento que se utiliza más comúnmente es el método generalizado de cociente de verosimilitudes. Este procedimiento ha producido resultados razonables en muchos casos, aunque no se basa en criterios de optimalidad bien definidos.

Por otro lado, en los procedimientos Bayesianos el único criterio de optimalidad que se requiere consiste en la maximización de la utilidad esperada. Sin embargo, la elección de distintas funciones de utilidad da lugar a una variedad de criterios concretos de contraste.

Como se mostró en el Capítulo 3, cuando la distribución de X pertenece a una familia exponencial regular y no se tienen parámetros de ruido, el procedimiento frecuentista vía el método generalizado de cociente de verosimilitudes resulta un caso particular de la solución bayesiana propuesta, con la divergencia logarítmica como medida de discrepancia, si se hace uso de una distribución inicial no informativa y si el valor de δ^* se asigna de manera adecuada.

El caso, más común en las aplicaciones, en el que existen parámetros de ruido, merece mayor atención. La equivalencia de las soluciones clásica y Bayesiana no es evidente en este caso, aún si se consideran solamente distribuciones que pertenezcan a una familia exponencial. Sin embargo, como puede apreciarse en el Capítulo 4, el procedimiento Bayesiano propuesto para contrastar la hipótesis lineal general en el modelo de regresión lineal múltiple genera una regla de decisión esencialmente equivalente a la regla producida por el procedimiento clásico usual. Asimismo, en una gran variedad de ejemplos se ha observado una coincidencia análoga. Esto resulta interesante, ya que el tratamiento que se le da al parámetro de ruido en cada uno de estos enfoques es conceptualmente distinto.

Por otro lado, debe observarse que tanto los procedimientos clásicos como los procedimientos Bayesianos más comunes restringen el espacio parametral al conjunto $\Theta_0 \cup \Theta_1$, lo que supone que se cuenta con cierta información inicial que, en todo caso, podría describirse de una manera más adecuada a través de una distribución inicial definida sobre la totalidad del espacio parametral Θ. En particular, esto implica que la asignación de la distribución inicial depende de la forma de las hipótesis que se desea contrastar, como puede verse en el Capítulo 2.

En contraste, en el procedimiento propuesto en el Capítulo 3 se postula que el espacio parametral relevante es Θ, independientemente de las hipótesis a contrastar. Por lo tanto, sin importar la forma de la hipótesis que se desea contrastar, este procedimiento sólo requiere de la asignación de una única distribución inicial sobre Θ, lo que, además, reduce el problema de la asignación de distribuciones iniciales no informativas.

A diferencia de los procedimientos Bayesianos revisados en el Capítulo 2, el procedimiento general propuesto en este trabajo contiene como caso particular , bajo ciertas condiciones descritas en el Capítulo 3, a la solución frecuentista comúnmente utilizada. Cabe mencionar que, aunque la solución presentada se basa en una propuesta de Bernardo para contrastar las hipótesis $H_0:\theta = \theta_0$ vs. $H_1:\theta \neq \theta_0$, los procedimientos no son equivalentes en este caso.

BIBLIOGRAFIA

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, Chichester.

BARNDORFF-NIELSEN, O. & PEDERSEN, K. (1968). Sufficient Data Reduction and Exponential Families. *Math. Scand.* **2**, 197-202.

BERGER, J. O. (1980). *Statistical Decision Theory: Foundations, Concepts and Methods.* Springer-Verlag, New York.

BERNARDO, J. M. (1979a). Reference Posterior Distributions for Bayesian Inference. *J. R. Statist. Soc. B* **41**, 113-147.

BERNARDO, J. M. (1979b). An Information-Theoretical Approach to Approximations in Statistics. *12th European Meeting on Statistics*, Varna, Bulgaria.

BERNARDO, J. M. (1980). A Bayesian Analysis of Classical Hypothesis Testing. *Trab. Estadist.* **31**, 605-618.

BERNARDO, J. M. (1984). Análisis Bayesiano de los Contrastes de Hipótesis Paramétricos. *Pub. Depto. Bioestadist.* **6**, Universidad de Valencia.

BERNARDO, J. M., FERRANDIZ, J. R. & SMITH, A. F. .M. (1985). The Foundations of Decision Theory: an Intuitive, Operational Approach with Mathematical Extensions. *Theory and Decision* **19**, 127-150.

BOX, G. E. P. & TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Mass.

BROEMELING, L. D. (1985). *Bayesian Analysis of Linear Models.* Marcel Dekker, New York.

COX, D. R. & HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

DEGROOT, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

DIACONIS, P. & YLVISAKER, D. (1979). Conjugate Priors for Exponential Families. *Ann. Statist. 7,* 269-281.

GUTIERREZ, E. (1989). Análisis Bayesiano de Algunos Contrastes de Hipótesis Paramétricas. *Tesis de Licenciatura,* Facultad de Ciencias, UNAM.

GUTIERREZ, E. (1991). Expected Logarithmic Divergence for Exponential Families. *Fourth Valencia International Meeting on Bayesian Statistics.* Peñíscola, Spain.

JEFFREYS, H. (1961). *Theory of Probability.* 3rd. Ed., Clarendon Press, Oxford.

KULLBACK, S. (1959). *Information Theory and Statistics.* Wiley, New York.

LEHMANN, E. (1986). *Testing Statistical Hypotheses.* 2nd. Ed., Wiley, New York.

LINDLEY, D. V. (1972). *Bayesian Statistics: A Review.* SIAM, Philadelphia.

PRESS, S. J. (1989). *Bayesian Statistics: Principles Models and Applications.* Wiley, New York.

RUEDA, R. & GUTIERREZ, E. (1990). Un Procedimiento Bayesiano de Contraste de Hipótesis Paramétricas. *Aportaciones Matematicas* **8**, 257-263.

SEBER, G. A. F. (1977). *Linear Regression Analysis.* Wiley, New York.

WINKLER, R. L. (1972). *Introduction to Bayesian Inference and Decision.* Holt, Rinehart and Winston, New York.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* Wiley, New York.