

03061



**UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO**

UNIDAD ACADÉMICA DE LOS CICLOS PROFESIONALES
Y DE POSGRADO DEL COLEGIO DE CIENCIAS
Y HUMANIDADES
(UACYP-CCH)

SELECCION BAYESIANA DE MODELOS:
APLICACION A LOS MODELOS DE REGRESION
LINEAL HETEROSCEDASTICOS USANDO FACTORES
DE BAYES INTRINSECOS.

T E S I S

PARA OBTENER EL GRADO DE
MAESTRA EN ESTADISTICA E INVESTIGACION
DE OPERACIONES
P R E S E N T A :
LIC. NORMA EDITH ALAMILLA LOPEZ

ASESOR DE TESIS
DR. JUAN ANTONIO CANO LOPEZ



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

ESTA TESIS NO SALE
DE LA BIBLIOTECA

Selección bayesiana de modelos: Una aplicación a los
modelos de regresión lineal heteroscedásticos, utilizando
los factores de Bayes intrínsecos.

Norma Edith Alamilla López.

Autorizo a la Dirección General de Bibliotecas de la
UNAM a difundir en formato electrónico e impreso el
contenido de mi trabajo recepcional.

NOMBRE: Norma Edith

Alamilla López

FECHA: 6 Agosto de 2004

FIRMA: 

Índice general

. Dedicatoria y agradecimientos	5
. Introducción	7
1. Conceptos básicos del análisis bayesiano	9
1.1. Información a priori y probabilidad subjetiva	10
1.1.1. Probabilidad subjetiva	10
1.1.2. Distribución a priori no informativa	10
1.2. Análisis bayesiano	13
1.2.1. Distribución a posteriori	13
1.2.2. Inferencia bayesiana	16
2. Análisis bayesiano del modelo lineal	25
2.1. Introducción al modelo de regresión	25
2.2. Modelo lineal general	26
2.2.1. Análisis del modelo lineal	26
2.2.2. Varianza σ^2 conocida	29
2.2.3. Varianza σ^2 desconocida	30
2.2.4. Conjuntos creíbles HPD para el modelo lineal	34
2.2.5. Introducción al contraste de hipótesis como selección de modelos	35
3. Factores de Bayes intrínsecos.	37
3.1. Preliminares	38
3.2. Otros métodos bayesianos	43
3.2.1. Distribuciones a priori convencionales	44
3.2.2. Métodos asintóticos y criterio de información de Bayes	45
3.2.3. Distribuciones a priori no informativas convencionales	45
3.2.4. Muestras de entrenamiento	46
3.3. Factores de Bayes intrínsecos.	46
3.3.1. Modelos anidados	46
3.3.2. El factor de Bayes intrínseco	47
3.3.3. El factor de Bayes intrínseco esperado	52
3.3.4. Comparaciones	54

3.4.	Distribuciones a priori intrínsecas	59
3.4.1.	Definición y motivación	59
3.4.2.	Distribuciones a priori intrínsecas para modelos anidados.	61
4.	Pruebas de igualdad de coeficientes de regresión	65
4.1.	El estadístico de Chow	65
4.2.	Definición y motivación	66
4.3.	Distribuciones a priori intrínsecas y factor de Bayes	70
4.3.1.	Distribuciones a priori intrínsecas.	70
4.3.2.	Distribuciones a priori intrínsecas para modelos de regresión heteroscedásticos.	71
4.3.3.	Consideraciones de diseño.	77
4.3.4.	El factor de Bayes para modelos de regresión heteroscedásticos.	78
4.4.	Comparaciones	81
4.4.1.	Prueba de homoscedasticidad.	81
4.4.2.	Prueba de igualdad de regresiones bajo homoscedasticidad.	85
4.4.3.	Una prueba incondicional	86
4.5.	Ejemplos.	86
.	Conclusiones	91
.	Notación	93
.	Apéndice	95

Dedicatoria y agradecimientos

Esta tesis se la dedico a mis padres, Jorge y María Matea, por su amor y apoyo incondicional que me dieron para estudiar y realizar esta tesis.

A mi esposo José Angel por su comprensión, apoyo, consejos y correcciones que me hizo durante la realización de este trabajo.

A mi hijo Francisco Mateo por haber sufrido todas las consecuencias de las horas de trabajo dedicadas a la presente.

A mis maestros por haberme enseñado a querer la estadística, y sobre todo al Dr. Eduardo Gutiérrez Peña catedrático de la UNAM, quien me encauzó por el camino de la estadística bayesiana, a quien se lo agradezco enormemente.

A mis amigas y amigos de la UNAM, Lety, Oly, Adriana, Gerardo Núñez, Adalberto, y de la Universidad Miguel Hernández de Elche (España) a Inma y Vero, quienes con sus ánimos y consejos fueron una clave importante en el desarrollo de este trabajo, y a todos aquellos compañeros que de alguna u otra forma siempre estuvieron a mi lado apoyándome.

Agradezco sus valiosos consejos y el apoyo proporcionado para realizar la presente tesis, al Dr. Juan Antonio Cano Sánchez, catedrático de la Universidad de Murcia (España), y al Dr. Elías Moreno Bas, catedrático de la Universidad de Granada (España), por haberme proporcionado varios de sus artículos, sobre los cuales esta basada una parte de este trabajo.

A mis sinodales Dr. Federico O'Reilly, Dr. Raul Rueda, M. en C. Salvador Zamora, de la Universidad Nacional Autónoma de México, y al Dr. Luis Enrique Nieto Barajas, del Instituto Tecnológico Autónomo de México, por todas sus observaciones realizadas respecto a esta tesis. Así como al Dr. Jorge Bustamente por sus buenos consejos respecto a la misma.

Por supuesto, no puedo dejar de agradecer al CONACYT por haberme dado el apoyo necesario para realizar el estudio de maestría en la UNAM.

Y por último agradezco a la Universidad Tecnológica de la Mixteca el haberme facilitado la tramitación y presentación de esta tesis.

Introducción

Actualmente es reconocida la importancia que tiene la estadística aplicada a diversas áreas del conocimiento. Cada vez son más las diferentes disciplinas que usan como herramienta los métodos estadísticos, para el análisis y la interpretación de datos, todo esto con el fin de mejorar la toma de decisiones.

La estadística bayesiana, es un enfoque alternativo (en la resolución de problemas estadísticos) a la estadística clásica y frecuentista. Utiliza el teorema de Bayes combinando la información a priori y la de los datos (función de verosimilitud) para obtener la distribución a posteriori. De la distribución a posteriori es de donde se derivan los estimadores y, cuya interpretación, difiere radicalmente de la interpretación proporcionada por la inferencia clásica.

La estadística bayesiana proporciona un completo paradigma de inferencia estadística y teoría de la decisión con incertidumbre. Está firmemente basada en fundamentos matemáticos estrictos y una metodología coherente con la que es posible incorporar información inicial relevante. La inferencia sobre una cantidad de interés queda descrita mediante modificaciones en la incertidumbre a la luz de la evidencia, y el teorema de Bayes especifica cómo estas modificaciones deben llevarse a cabo. Los métodos bayesianos pueden ser aplicados a problemas que han sido inaccesibles a la teoría frecuentista tradicional.

Cuando se trabaja con distribuciones iniciales “no informativas”, los resultados de los métodos bayesianos coinciden en los resultados numéricos presentados por los métodos clásicos, aunque como se dijo anteriormente, las interpretaciones de estos resultados difieren radicalmente de los obtenidos a través de los métodos bayesianos.

El presente trabajo es una revisión bibliográfica. Haremos un recuento de algunos conceptos relacionados con el análisis bayesiano. Como es bien sabido, en el enfoque bayesiano para problemas de selección de modelos y contrastes de hipótesis la herramienta principal son los factores de Bayes. En el caso en que al definir los factores de Bayes, se usa una distribución a priori impropia, éstos no están bien definidos. Los “factores de Bayes intrínsecos”, definidos por Berger y Pericchi (1996), es un método interesante para resolver dicho problema, por lo que se hace un análisis exhaustivo del método.

Se hace una revisión exhaustiva de un método propuesto por Moreno, Torres y

Casella (2002), para el problema que se presenta cuando se quiere comprobar si los vectores de coeficientes de regresión son iguales, para dos modelos de regresión normales independientes, cuando no se conoce la varianza de los errores.

Por lo que el presente trabajo se encuentra estructurado de la siguiente manera:

En el primer capítulo, se presenta una introducción a la teoría de la estadística bayesiana, así como también se definen algunos conceptos básicos del análisis bayesiano, comenzando con la definición de lo que es la probabilidad subjetiva, hasta tratar los problemas típicamente estadísticos, que son: estimación, contraste de hipótesis y predicción.

En el segundo capítulo, se describe el modelo de regresión lineal normal, y las propiedades de los estimadores relacionados con los parámetros del modelo. Se hace un repaso de los conceptos básicos de los modelos de regresión lineal normales desde el punto de vista bayesiano.

En el tercer capítulo se hace una revisión exhaustiva a la metodología intrínseca, para solventar el problema del uso de distribuciones a priori impropias en los problemas de selección de modelos y de contraste de hipótesis, como herramienta “automática”, diseñada para, a través de distribuciones a priori no informativas estándar, proporcionar distribuciones a priori propias con las que pueda calcularse un factor de Bayes, asintóticamente verdadero. Se muestran algunas comparaciones con algunos otros métodos bayesianos, y se definen las distribuciones a priori intrínsecas.

Por último, en el capítulo cuatro, se hace una revisión exhaustiva al método propuesto por Moreno, Torres y Casella (2002), para solventar el problema que se presenta cuando se quiere comprobar si los coeficientes de regresión son iguales para dos modelos de regresión normales independientes, cuando no se conocen las varianzas de los errores. Estos autores generan una distribución a priori propia (distribución intrínseca) para la cual el factor de Bayes y la probabilidad a posteriori del modelo, están bien definidos; y usan a la probabilidad a posteriori del modelo como una herramienta de selección.

El paquete computacional usado para las soluciones numéricas en este trabajo, es el Mathematica.

Capítulo 1

Conceptos básicos del análisis bayesiano

La estadística bayesiana constituye una alternativa a la estadística clásica o frecuentista, en la solución de problemas habituales estadísticos, como son, estimación, prueba de hipótesis y predicción. En el enfoque bayesiano el parámetro no es simplemente un número, sino que se considera una variable aleatoria. La inferencia respecto a los posibles valores del parámetro, se obtiene aplicando el cálculo de probabilidades (teorema de Bayes) para obtener su distribución condicionada a la información disponible. Una vez obtenida la distribución de probabilidad del parámetro, los problemas habituales de inferencia quedan resueltos con la distribución a posteriori.

La estadística bayesiana también puede plantearse como un problema de decisión estadística. La estructura de un problema de decisión en un ambiente de incertidumbre está compuesto por un espacio \mathcal{D} de acciones, una σ -álgebra, ξ , de subconjuntos de un conjunto Θ , y un conjunto \mathcal{C} de consecuencias posibles.

El espacio \mathcal{D} de acciones debe definirse de manera que sea exhaustivo, y sus elementos mutuamente excluyentes, los elementos de la σ -álgebra, ξ , son los eventos relevantes al problema de decisión, y el conjunto \mathcal{C} de consecuencias posibles, describe las consecuencias de elegir una acción del conjunto \mathcal{D} , cuando ocurre un evento en ξ .

La información que el estadístico tiene sobre la verosimilitud de los distintos eventos relevantes al problema de decisión, debe ser cuantificada a través de una medida de probabilidad sobre el espacio Θ . De la misma manera, es posible cuantificar las preferencias del estadístico entre las distintas consecuencias, esto se puede hacer mediante el uso de una función de utilidad, donde la solución óptima de el problema de decisión consiste en elegir aquella acción en \mathcal{D} que la maximice.

El enfoque bayesiano tiene dos ventajas principales. La primera es su generalidad y coherencia, conceptualmente todos los problemas de inferencia se resuelven con los principios del cálculo de probabilidades y teoría de la decisión. La segunda, es la capacidad de incorporar la información a priori adicional a la muestra que se tenga del parámetro, sobre el cual se desea inferir.

Esta fortaleza, es sin embargo, también su debilidad, porque exige siempre representar la información inicial respecto al vector de parámetros mediante una distribución a priori, $\pi(\theta)$. Éste es el aspecto más controvertido de esta teoría, ya que algunos científicos rechazan que la información inicial - que puede incluir los prejuicios del investigador - se incluya en un proceso de inferencia científica; aunque en principio, esto podría evitarse estableciendo una distribución a priori no informativa o de referencia.

El proceso intelectual asociado a la inferencia bayesiana es mucho más coherente con el pensamiento usual del científico que el que ofrece el paradigma frecuentista. Los procedimientos bayesianos constituyen una tecnología emergente de procesamiento y análisis de información para la que cabe esperar una presencia cada vez más intensa en el campo de la aplicación de la estadística a la investigación en un futuro próximo.

A lo largo de este capítulo, haremos una revisión de todos los conceptos básicos de la estadística bayesiana, desde la definición de lo que es la probabilidad subjetiva hasta tratar los problemas típicamente estadísticos, que son: estimación, contraste de hipótesis y predicción.

1.1. Información a priori y probabilidad subjetiva

1.1.1. Probabilidad subjetiva

La estadística bayesiana, basada en una interpretación subjetiva de la probabilidad, establece que la probabilidad que un estadístico asigna a cada uno de los posibles resultados de un proceso, representa su propio juicio sobre la verosimilitud de que se obtenga algún resultado. Este juicio estará basado en opiniones e información acerca del proceso. Además, esta interpretación subjetiva de la probabilidad puede ser formalizada (Degroot, (1982)).

En general, si los juicios de un estadístico acerca de las verosimilitudes relativas a diversas combinaciones de resultados, satisfacen ciertas condiciones de consistencia, entonces puede demostrarse que sus probabilidades subjetivas para los diferentes sucesos posibles, pueden determinarse en forma única.

Un elemento importante en las decisiones de muchos problemas es la información a priori que se tenga acerca del parámetro de interés θ . La información que el estadístico tiene sobre la verosimilitud de dichos sucesos debe ser cuantificada a través de una medida de probabilidad sobre el espacio paramétrico Θ .

1.1.2. Distribución a priori no informativa

Sea $\mathcal{F} = \{f(x | \theta) : \theta \in \Theta\}$, una familia paramétrica de distribuciones de probabilidad. Un modelo estadístico bayesiano está formado por un modelo paramétrico, $f(x | \theta)$, y una distribución a priori sobre los parámetros, $\pi(\theta)$.

En ciertos problemas, el conocimiento a priori sobre el verdadero valor del parámetro θ puede ser muy débil o vago, ésto ha llevado a generar un tipo de distribuciones

iniciales llamada *distribuciones a priori no informativas*, las cuales reflejan un estado de ignorancia inicial. Por ejemplo, en el contraste de dos hipótesis simples, la distribución a priori que asigna una probabilidad de $\frac{1}{2}$ a cada hipótesis, podría considerarse no informativa. Un ejemplo más complejo sería el siguiente.

Ejemplo 1 *Supóngase que el parámetro de interés es la media, θ , de una distribución normal, y a priori no se dispone de ningún dato o no se quiere evidenciar nada sobre ella. Puesto que su espacio paramétrico es $\Theta = (-\infty, \infty)$, si necesitamos una distribución a priori no informativa, parece razonable dar igual peso a todos los posibles valores de θ . Desafortunadamente, si elegimos $\pi(\theta) = c > 0$, entonces π asigna “masa” infinita a \mathfrak{R} es decir, $\int \pi(\theta) d\theta = \infty$, o sea, π es una distribución impropia. Sin embargo, con tal π , puede trabajarse satisfactoriamente. La elección de c es irrelevante, así que típicamente la distribución a priori no informativa para este problema es $\pi(\theta) = 1$.*

Como en el ejemplo anterior, ocurre con frecuencia, que la distribución a priori no informativa natural ($\pi(\theta) = 1$), es una distribución a priori impropia; es decir, que tiene “masa” infinita.

No obstante, la situación más simple es cuando el espacio paramétrico Θ es un conjunto finito de n elementos. La distribución a priori no informativa natural es entonces, la que asigna a cada elemento de Θ probabilidad $1/n$. Uno puede generalizar ésto a Θ infinito, dando a cada $\theta \in \Theta$ igual densidad, teniendo así una distribución a priori no informativa $\pi(\theta) \equiv c$. Esta distribución fue inicialmente introducida por Laplace (1812), y fue duramente criticado debido a una falta de invarianza bajo transformación (ver Jaynes, 1983).

Por ejemplo, en lugar de considerar el parámetro θ , supóngase que el problema ha sido parametrizado en términos de $\eta = \exp(\theta)$. Esto es, una transformación uno a uno, y por lo tanto no debería afectar la respuesta final. Pero, si $\pi(\theta)$ es la densidad para θ , entonces la correspondiente densidad para η es (notando que el jacobiano de la transformación es $d\theta/d\eta = d \log \eta/d\eta = \eta^{-1}$)

$$\pi^*(\eta) = \eta^{-1} \pi(\log \eta).$$

Por lo tanto, si la densidad a priori no informativa para θ se elige como una constante, entonces se debería elegir la densidad a priori no informativa para η proporcional a η^{-1} para mantener consistencia (y llegar a la misma respuesta en ambas parametrizaciones). Así, llegamos a la conclusión de que no podemos elegir densidades a priori no informativas para θ y η constantes y mantener la consistencia.

La falta de invarianza de las distribuciones a priori constantes ha llevado a la búsqueda de distribuciones a priori no informativas que sean apropiadamente invariantes bajo transformaciones.

Los esfuerzos para derivar distribuciones a priori no informativas considerando los problemas de invarianza bajo transformación comenzaron con Jeffreys (1947).

Ejemplo 2 *(Berger, (1985))(Parámetros de localización). Supongamos que Ω y Θ (espacios paramétricos), son subconjuntos de \mathfrak{R}^P , y que la densidad de \mathbf{X} es de la forma*

$f(\mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{x} - \boldsymbol{\theta})$ (es decir, depende solamente de $(\mathbf{x} - \boldsymbol{\theta})$). Para determinar una distribución a priori no informativa en esta situación, supongamos que, en lugar de observar \mathbf{X} , observamos la variable aleatoria $\mathbf{Y} = \mathbf{X} + \mathbf{c}$ ($\mathbf{c} \in \mathbb{R}^P$). Si definimos $\boldsymbol{\eta} = \boldsymbol{\theta} + \mathbf{c}$, es fácil mostrar que \mathbf{Y} tiene densidad $f(\mathbf{y} - \boldsymbol{\eta})$. Si $\Omega = \Theta = \mathbb{R}^P$, entonces el espacio muestral y el espacio paramétrico para $(\mathbf{Y}, \boldsymbol{\eta})$ también están en \mathbb{R}^P . Sean π y π^* distribuciones a priori no informativas para $(\mathbf{X}, \boldsymbol{\theta})$ y $(\mathbf{Y}, \boldsymbol{\eta})$ respectivamente, el argumento anterior implica que π y π^* deberían ser iguales, es decir

$$P_{\pi}(\boldsymbol{\theta} \in \mathbf{A}) = P_{\pi^*}(\boldsymbol{\eta} \in \mathbf{A}), \quad (1.1)$$

para cualquier $\mathbf{A} \in \mathbb{R}^P$. Puesto que $\boldsymbol{\eta} = \boldsymbol{\theta} + \mathbf{c}$, es cierto que

$$P_{\pi^*}(\boldsymbol{\eta} \in \mathbf{A}) = P_{\pi}(\boldsymbol{\eta} = \boldsymbol{\theta} + \mathbf{c} \in \mathbf{A}) = P_{\pi}(\boldsymbol{\theta} \in \mathbf{A} - \mathbf{c}), \quad (1.2)$$

donde $\mathbf{A} - \mathbf{c} = \{\mathbf{z} - \mathbf{c}, \mathbf{z} \in \mathbf{A}\}$. Combinando (1.1) y (1.2) se tiene que

$$P_{\pi}(\boldsymbol{\theta} \in \mathbf{A}) = P_{\pi}(\boldsymbol{\theta} \in \mathbf{A} - \mathbf{c}). \quad (1.3)$$

Además, este argumento se aplica para $\mathbf{c} \in \mathbb{R}^P$, así que (1.3) se cumple $\forall \mathbf{c} \in \mathbb{R}^P$. Cualquier distribución π que satisface esta relación se dice que es una distribución a priori invariante para cambios de localización. Asumiendo que la distribución a priori tiene una densidad, podemos escribir (1.3) como

$$\int_{\mathbf{A}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\mathbf{A} - \mathbf{c}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\mathbf{A}} \pi(\boldsymbol{\theta} - \mathbf{c}) d\boldsymbol{\theta}.$$

Si esto se cumple $\forall \mathbf{A}$, entonces puede demostrarse que

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} - \mathbf{c}), \quad \forall \boldsymbol{\theta}.$$

Tomando $\boldsymbol{\theta} = \mathbf{c}$, se tiene que $\pi(\mathbf{c}) = \pi(\mathbf{0})$. Recordemos, que esto se cumple $\forall \mathbf{c} \in \mathbb{R}^P$. La conclusión es que π puede ser una función constante. Así, es conveniente elegir una constante igual a 1, por lo que una densidad a priori no informativa para un parámetro de localización es $\pi(\boldsymbol{\theta}) = 1$.

Para problemas más generales, varias sugerencias han sido propuestas para determinar distribuciones a priori no informativas. El método más usado es el de Jeffreys (1947), el cual propone

$$\pi(\boldsymbol{\theta}) = [I(\boldsymbol{\theta})]^{1/2}, \quad \boldsymbol{\theta} \in \mathfrak{R}$$

como una distribución a priori no informativa, donde $I(\boldsymbol{\theta})$ es la información esperada de Fisher, que bajo suposiciones que comúnmente se satisfacen, (ver Lehmann (1983)) está dada por

$$I(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}} \left[\frac{\partial^2 \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right].$$

Si $\theta = (\theta_1, \theta_2, \dots, \theta_p)'$, Jeffreys (1961) sugirió el uso de

$$\pi(\theta) = [\det \mathbf{I}(\theta)]^{1/2}, \quad \theta \in \mathfrak{R}^P \quad (1.4)$$

donde $\mathbf{I}(\theta)$ es la matriz de información esperada de Fisher, que bajo condiciones de regularidad, es la matriz con elemento (i, j)

$$I_{ij}(\theta) = -E_{\theta} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\mathbf{X}|\theta) \right].$$

Otra distribución a priori no informativa también de mucha relevancia, es la *distribución a priori de referencia* $(\pi_i^R(\theta_i))$, desarrollada en Bernardo (1979) y modificada para problemas multiparamétricos en Berger y Bernardo (1992). Esta distribución no puede describirse de manera simple, pero puede entenderse como una generalización de la distribución a priori de Jeffreys distinguiendo entre los parámetros de ruido y los de interés.

1.2. Análisis bayesiano

1.2.1. Distribución a posteriori

El análisis bayesiano se hace combinando la información a priori $\pi(\theta)$ y la información muestral x en lo que se denomina la distribución a posteriori de θ dado x , sobre la que se basan todas las decisiones e inferencias.

Definición 1 La distribución a posteriori de θ dado x , denotada por $\pi(\theta|x)$ se define como la distribución condicional de θ dado x y está definida mediante el teorema de Bayes por

$$\pi(\theta | x) = \frac{f(x | \theta)\pi(\theta)}{\int f(x | \theta)\pi(\theta)d\theta}.$$

El teorema de Bayes combina la creencia a priori que se tenga acerca de θ con la información sobre θ contenida en la muestra, x , para dar una imagen completa de la creencia final sobre θ . En otras palabras, la distribución a posteriori combina la creencia a priori que se tenga acerca de θ con la información que se tenga de θ contenida en la muestra x , para dar una imagen compuesta sobre la creencia final acerca de θ .

Teniendo las distribuciones $f(x | \theta)$ y $\pi(\theta)$, además de obtener la distribución a posteriori de θ , también se pueden obtener, la distribución conjunta de (x, θ)

$$h(x, \theta) = \pi(\theta)f(x | \theta),$$

y la distribución marginal de x ,

$$m(x) = \int h(x, \theta)d\theta = \int f(x | \theta)\pi(\theta)d\theta,$$

es claro que (con la condición de que $m(x) \neq 0$)

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)}.$$

Tal como la distribución a priori refleja nuestra creencia acerca de θ antes de la experimentación, así $\pi(\theta|x)$ refleja la actualización de nuestra creencia acerca de θ después (posterior a) de observar la muestra x .

Ejemplo 3 *Supóngase que $X \sim \mathcal{N}(\theta, \sigma^2)$, donde θ es desconocido pero σ^2 es conocido. Sea $\pi(\theta) = \mathcal{N}(\mu, \tau^2)$, donde μ y τ^2 son conocidas. Entonces*

$$h(x, \theta) = \pi(\theta) f(x|\theta) = (2\pi\sigma\tau)^{-1} \exp \left\{ -\frac{1}{2} \left[\frac{(\theta - \mu)^2}{\tau^2} + \frac{(x - \theta)^2}{\sigma^2} \right] \right\}.$$

Para encontrar $m(x)$, nótese que definiendo

$$\rho = \tau^{-2} + \sigma^{-2} = \frac{\tau^2 + \sigma^2}{\tau^2\sigma^2},$$

y completando cuadrados obtenemos que

$$\begin{aligned} \frac{1}{2} \left[\frac{(\theta - \mu)^2}{\tau^2} + \frac{(x - \theta)^2}{\sigma^2} \right] &= \frac{1}{2} \left[\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right) \theta^2 - 2 \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \theta + \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \right] \\ &= \frac{1}{2} \rho \left[\theta^2 - \frac{2}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \theta \right] + \frac{1}{2} \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \\ &= \frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 + \frac{1}{2\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right)^2 + \frac{1}{2} \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2} \right) \\ &= \frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 + \frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)}. \end{aligned}$$

Por lo tanto,

$$h(x, \theta) = (2\pi\sigma\tau)^{-1} \exp \left\{ -\frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 \right\} \exp \left\{ -\frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)} \right\},$$

y la distribución marginal de x es

$$\begin{aligned} m(x) &= \int_{-\infty}^{\infty} h(x, \theta) d\theta \\ &= \int_{-\infty}^{\infty} (2\pi\sigma\tau)^{-1} \exp \left\{ -\frac{1}{2} \rho \left[\theta - \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2} \right) \right]^2 \right\} \exp \left\{ -\frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)} \right\} d\theta \\ &= (2\pi\rho)^{-1/2} (\sigma\tau)^{-1} \exp \left\{ -\frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)} \right\}, \end{aligned}$$

de aquí se sigue que la distribución a posteriori de θ es

$$\begin{aligned}\pi(\theta|x) &= \frac{h(x, \theta)}{m(x)} = \frac{(2\pi\sigma\tau)^{-1} \exp\left\{-\frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2\right\} \exp\left\{-\frac{(\mu-x)^2}{2(\sigma^2+\tau^2)}\right\}}{(2\pi\rho)^{-1/2} (\sigma\tau)^{-1} \exp\left\{-\frac{(\mu-x)^2}{2(\sigma^2+\tau^2)}\right\}} \\ &= \left(\frac{\rho}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2\right\}.\end{aligned}$$

Nótese que la distribución marginal de X es una distribución normal $\mathcal{N}(\mu, (\sigma^2 + \tau^2))$ y que la distribución a posteriori de θ dado x también es una distribución normal $\mathcal{N}(\mu(x), \rho^{-1})$, donde

$$\begin{aligned}\mu(x) &= \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right) = \frac{\sigma^2}{\sigma^2 + \tau^2} \mu + \frac{\tau^2}{\sigma^2 + \tau^2} x \\ &= \frac{\sigma^2}{\sigma^2 + \tau^2} (\mu + x) = x - \frac{\sigma^2}{\sigma^2 + \tau^2} (x - \mu).\end{aligned}$$

Como un ejemplo concreto, consideremos la situación en la que a un individuo se le aplica una prueba de inteligencia y se desea estimar su coeficiente intelectual. Supongamos también que la variable aleatoria, X , que representa la inteligencia del individuo, se distribuye como una $\mathcal{N}(\theta, 100)$, donde θ representa el verdadero CI (coeficiente de inteligencia) del individuo. Asumamos también que θ , en la población, se distribuye de acuerdo a la distribución normal $\mathcal{N}(100, 225)$. Usando las ecuaciones obtenidas anteriormente, se sigue que, marginalmente, X sigue una distribución $\mathcal{N}(100, 325)$, mientras que la distribución a posteriori de θ dado x es también normal con media

$$\mu(x) = \frac{(100)(100) + (225)x}{(100 + 225)} = \frac{400 + 9x}{13},$$

y varianza

$$\rho^{-1} = \frac{(100)(225)}{(100 + 225)} = \frac{900}{13} = 69,23.$$

Así, si el individuo tiene un resultado de 115 puntos en la prueba de inteligencia, su CI verdadero tiene una distribución a posteriori $\mathcal{N}(110,39, 69,23)$.

Familia Conjugada.

En general, $m(x)$ y $\pi(\theta|x)$ no son fáciles de calcular. Si por ejemplo, X es una distribución normal $\mathcal{N}(\theta, \sigma^2)$ y θ es una distribución Cauchy $\mathcal{C}(\mu, \theta)$, entonces $\pi(\theta|x)$ sólo puede evaluarse numéricamente. Una gran parte de la literatura bayesiana se dedica a encontrar distribuciones a priori para las que $\pi(\theta|x)$ pueda calcularse fácilmente. Un tipo de distribuciones para las que se cumple ésto son las llamadas *distribuciones a priori conjugadas*, que han sido extensamente desarrolladas en Raiffa y Schlaifer (1961).

Definición 2 Sea \mathcal{F} la clase de funciones de densidad $f(x|\theta)$. Una clase \wp de distribuciones se dice que es una familia conjugada para \mathcal{F} si $\pi(\theta|x)$ está en la clase \wp para toda $f \in \mathcal{F}$ y $\pi \in \wp$.

Para una clase de densidades \mathcal{F} , una familia conjugada frecuentemente puede determinarse examinando la función de verosimilitud $l_x(\theta) = f(x|\theta)$, y eligiendo, como una familia conjugada, la clase de distribuciones de la misma forma funcional. Las distribuciones a priori resultantes son frecuentemente llamadas *distribuciones a priori conjugadas naturales*.

Cuando tratamos con distribuciones a priori conjugadas, generalmente no hay necesidad de calcular explícitamente $m(x)$. La razón es que como

$$\pi(\theta|x) = h(x, \theta) / m(x),$$

entonces los factores que involucran a θ en $\pi(\theta|x)$ deben ser los mismos que los factores que involucran a θ en $h(x, \theta)$. Por lo tanto, sólo es necesario mirar los factores que involucran a θ en $h(x, \theta)$, y ver si se puede reconocer que pertenezca a un tipo de distribución en particular, si es así, entonces $\pi(\theta|x)$ es esa distribución. La densidad marginal $m(x)$ puede entonces determinarse, si se desea, dividiendo $h(x, \theta)$ entre $\pi(\theta|x)$.

1.2.2. Inferencia bayesiana

Los principales problemas de la inferencia estadística son estimación puntual, estimación por intervalos y contraste de hipótesis. Desde el punto de vista bayesiano, la inferencia sobre θ , debe depender estrictamente de la distribución a posteriori $\pi(\theta|x)$, por lo que puede usarse para describir las propiedades de θ .

Estimación puntual

El uso inferencial más simple de la distribución a posteriori es hacer una estimación puntual para θ , con una medida de precisión asociada. En el caso de la estimación puntual pueden usarse, por ejemplo, la media, la moda, o la mediana como estimadores de θ .

Para estimar θ , las técnicas clásicas pueden aplicarse a la distribución a posteriori. La técnica clásica más común es la estimación de máxima verosimilitud, que elige como estimador de θ el valor $\hat{\theta}$ que maximiza la función de verosimilitud $l(\theta) = f(x|\theta)$. La estimación bayesiana de θ se define como sigue.

Definición 3 La probabilidad máxima generalizada estimada de θ se define como la moda más grande, $\hat{\theta}$, de $\pi(\theta|x)$ (es decir, el valor $\hat{\theta}$ que maximiza $\pi(\theta|x)$). (Berger, (1985))

Otras estimaciones bayesianas de θ son la media y la mediana de $\pi(\theta|x)$. La media y la mediana (y la moda) son relativamente fáciles de encontrar cuando la distribución

a priori, y por lo tanto la distribución a posteriori, pertenecen a una *familia conjugada de distribuciones*.

La media y la mediana de la distribución a posteriori son frecuentemente (Berger, 1985), mejores estimadores de θ que la moda. Aquí probablemente merecería la pena calcular y comparar las tres estimaciones en un estudio bayesiano, sobre todo con respecto a su robustez a cambios en la distribución a priori.

Cuando se presenta una estimación estadística, es necesario indicar la precisión de la estimación. La medida bayesiana usual, de la precisión de una estimación es, (en una dimensión) la varianza a posteriori de la estimación, definida como sigue.

Definición 4 Si θ es un parámetro real con distribución a posteriori $\pi(\theta|x)$, y δ es el estimador de θ , entonces la varianza a posteriori de δ es

$$V_{\delta}^{\pi}(x) \equiv E^{\pi(\theta|x)} [(\theta - \delta)^2].$$

Cuando δ es la media a posteriori

$$\mu^{\pi}(x) \equiv E^{\pi(\theta|x)} [\theta],$$

entonces

$$V_{\mu^{\pi}}^{\pi}(x) \equiv V^{\pi}(x),$$

es la varianza a posteriori (y esto es la varianza de θ para la distribución a posteriori $\pi(\theta|x)$). La desviación estándar a posteriori es $\sqrt{V^{\pi}(x)}$. Nótese que

$$\begin{aligned} V_{\delta}^{\pi}(x) &= E^{\pi(\theta|x)} [(\theta - \delta)^2] = E [(\theta - \mu^{\pi}(x) + \mu^{\pi}(x) - \delta)^2] & (1.5) \\ &= E [(\theta - \mu^{\pi}(x))^2] + E [2(\theta - \mu^{\pi}(x))(\mu^{\pi}(x) - \delta)] + E [(\mu^{\pi}(x) - \delta)^2] \\ &= V^{\pi}(x) + 2(\mu^{\pi}(x) - \delta)(E[\theta] - \mu^{\pi}(x)) + (\mu^{\pi}(x) - \delta)^2 \\ &= V^{\pi}(x) + (\mu^{\pi}(x) - \delta)^2. \end{aligned}$$

La probabilidad máxima generalizada estimada (la moda a posteriori) es a menudo un estimador razonable, aunque seguramente se encuentren dificultades de existencia y unicidad en el caso multivariado.

La media a posteriori es

$$\boldsymbol{\mu}^{\pi}(x) = (\mu_1^{\pi}(x), \dots, \mu_p^{\pi}(x))^t = E^{\pi(\theta|x)} [\boldsymbol{\theta}],$$

que es un estimador bayesiano muy simple, y su precisión puede escribirse a través de la matriz de covarianzas a posteriori,

$$\mathbf{V}^{\pi}(x) = E^{\pi(\theta|x)} [(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(x))(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(x))^t].$$

Por ejemplo, el error estándar del estimador $\mu_i^{\pi}(x)$ de θ_i sería $\sqrt{V_{ii}^{\pi}(x)}$, donde $V_{ii}^{\pi}(x)$ es el elemento (i, i) de $\mathbf{V}^{\pi}(x)$.

El análogo a (1.5), para una estimación general δ de θ , puede demostrarse que es

$$\mathbf{V}_{\delta}^{\pi}(x) = E^{\pi(\theta|x)} [(\theta - \delta)(\theta - \delta)^t] = \mathbf{V}^{\pi}(x) + (\mu^{\pi}(x) - \delta)(\mu^{\pi}(x) - \delta)^t,$$

otra vez, es claro que la media a posteriori minimiza $\mathbf{V}_{\delta}^{\pi}(x)$. (Berger, (1985)).

Otra aproximación común en inferencia es presentar un intervalo de confianza para θ , el análogo bayesiano de un intervalo de confianza clásico es llamado *conjunto creíble*, el cual se define a continuación.

Conjuntos Creíbles

Definición 5 *Un conjunto creíble al 100(1- α)% es un subconjunto C de Θ tal que*

$$1 - \alpha \leq P(C|x) = \int_C dF^{\pi(\theta|x)}(\theta) = \begin{cases} \int_C \pi(\theta|x) d\theta & (\text{caso continuo}), \\ \sum_{\theta \in C} \pi(\theta|x) & (\text{caso discreto}), \end{cases}$$

Puesto que la distribución a posteriori es una distribución de probabilidad sobre Θ , uno puede hablar de la probabilidad de que θ esté en C . Esto está en contraste con los procedimientos de confianza clásicos, los cuales solamente son interpretados, clásicamente, en términos de probabilidades de cobertura, (es decir, la probabilidad de que la variable aleatoria X sea tal que el conjunto de confianza $C(X)$ contenga a θ).

En la elección de un conjunto creíble para θ , es usualmente deseable intentar minimizar su tamaño. Para hacer ésto, hay que incluir en el conjunto solamente aquellos puntos con la densidad a posteriori más grande, es decir, los valores “más probables” de θ .

Definición 6 *El conjunto creíble de máxima densidad a un nivel de confianza 100(1- α)% para θ , es el subconjunto C de Θ de la forma*

$$C = \{\theta \in \Theta : \pi(\theta|x) \geq k(\alpha)\},$$

donde $k(\alpha)$ es la constante más grande tal que

$$P(C|x) \geq 1 - \alpha.$$

Los conjuntos creíbles bayesianos son usualmente mucho más fáciles de calcular que los intervalos de confianza clásicos, particularmente en situaciones donde un estadístico suficiente simple no existe.

Para θ multivariado, la definición de un conjunto creíble de máxima densidad permanece igual. En el caso multivariado, el uso de la aproximación normal para la distribución a posteriori es a menudo especialmente valioso, debido a las dificultades aumentadas del cálculo, pero con las herramientas computacionales, hoy en día esto es más factible.

Contraste de hipótesis

En el contraste de hipótesis clásico, una hipótesis nula $H_0 : \theta \in \Theta_0$ y una hipótesis alternativa $H_1 : \theta \in \Theta_1$ se especifican. Un procedimiento se evalúa en términos de las probabilidades de error de tipo I y error de tipo II. Si rechazamos una hipótesis nula cuando debería ser aceptada, diremos que se ha cometido un *error de tipo I*. Por otra parte, si aceptamos una hipótesis nula que debería ser rechazada, diremos que se ha cometido un *error de tipo II*. En ambos casos decimos que se ha producido un juicio erróneo.

Para que las reglas de decisión (o contraste de hipótesis) sean buenas, deben diseñarse de tal modo que se minimicen los errores, para cualquier tamaño de muestra; un intento de disminuir un tipo de error suele ir acompañado de un crecimiento del otro tipo. En la práctica, un tipo de error puede ser más grave que el otro, y debe alcanzarse un compromiso que disminuya el error más grave.

En el análisis bayesiano, la tarea de decidir entre H_0 y H_1 es conceptualmente más fácil. Se seleccionan las distribuciones a priori $\pi_i(\theta_i)$, $i = 0, 1$, y se calculan las probabilidades a posteriori $\alpha_0 = P(\Theta_0|x)$ y $\alpha_1 = P(\Theta_1|x)$, para los parámetros de las hipótesis nula H_0 , y la hipótesis alternativa, H_1 , respectivamente. La tarea de decidir entre ambas hipótesis depende de dichos resultados. La ventaja conceptual es que α_0 y α_1 son las probabilidades (subjetivas) de las hipótesis con base a los datos y la distribución a priori.

Aunque las probabilidades a posteriori de las hipótesis sean las medidas básicas bayesianas en problemas de contraste, los siguientes conceptos relacionados también son de interés.

Definición 7 *A la razón α_0/α_1 se le denomina la razón de probabilidad a posteriori de H_0 respecto de H_1 , y π_0/π_1 la razón de probabilidad a priori, donde π_0 y π_1 denotan las probabilidades a priori de Θ_0 y Θ_1 , respectivamente. Además*

$$B = \frac{\text{razón de probabilidad a posteriori}}{\text{razón de probabilidad a priori}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0},$$

*es denominado **factor de Bayes** a favor de Θ_0 .*

Muchos investigadores prefieren utilizar “razones” que probabilidades, ya que a menudo es conveniente resumir la evidencia en términos de la razón de probabilidad a posteriori, (por ejemplo $\alpha_0/\alpha_1 = 10$ claramente nos dice que H_0 es 10 veces más probable de ser verdadero que H_1). El interés en el factor de Bayes es que algunas veces puede ser interpretado como “la razón dada por los datos para H_0 respecto de H_1 ”. Esta interpretación es claramente válida cuando las hipótesis son simples, es decir, cuando $\Theta_0 = \{\theta_0\}$ y $\Theta_1 = \{\theta_1\}$ porque entonces

$$\alpha_0 = \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}, \quad \alpha_1 = \frac{\pi_1 f(x|\theta_1)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)},$$

por lo que

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 f(x|\theta_0)}{\pi_1 f(x|\theta_1)},$$

y en consecuencia

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

En otras palabras, B representa la razón de verosimilitudes de la hipótesis nula H_0 respecto de la hipótesis alternativa H_1 , la cual se interpreta (incluso por muchos no bayesianos) como “la razón dada por los datos para H_0 respecto de H_1 ”.

En general, sin embargo, B dependerá de la distribución a priori. Para explorar esta dependencia, es conveniente escribir la distribución a priori de la siguiente forma

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \text{si } \theta \in \Theta_0 \\ \pi_1 g_1(\theta) & \text{si } \theta \in \Theta_1 \end{cases}, \quad (1.6)$$

donde g_0 y g_1 son densidades (propias) que describen cómo la “masa” a priori, se extiende sobre las dos hipótesis. (Recordemos que π_0 y π_1 son las probabilidades a priori de Θ_0 y Θ_1). Con estas representaciones podemos escribir

$$\frac{\alpha_0}{\alpha_1} = \frac{\int_{\Theta_0} dF^{\pi(\theta|x)}(\theta)}{\int_{\Theta_1} dF^{\pi(\theta|x)}(\theta)} = \frac{\int_{\Theta_0} f(x|\theta) \pi_0 dF^{g_0}(\theta) / m(x)}{\int_{\Theta_1} f(x|\theta) \pi_1 dF^{g_1}(\theta) / m(x)} = \frac{\pi_0 \int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{\pi_1 \int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)}.$$

Por lo tanto

$$B = \frac{\int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{\int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)} = \frac{\int_{\Theta_0} f(x|\theta) g_0(\theta) d\theta}{\int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta},$$

que representa la razón de verosimilitudes “ponderadas” (por g_0 y g_1) de Θ_0 respecto a Θ_1 .

Debido a la participación de g_0 y g_1 , esto no puede interpretarse como una medida del soporte relativo para las hipótesis, proporcionado únicamente por los datos. Algunas veces, sin embargo, B será relativamente insensible para elecciones razonables de g_0 y g_1 , y entonces tal interpretación es razonable.

Es interesante mostrar el uso de distribuciones a priori impropias en problemas de contraste de hipótesis. Supongamos que consideramos una distribución a priori impropia

$$\pi(\theta) = \begin{cases} c_0 g_0(\theta) & \text{si } \theta \in \Theta_0 \\ c_1 g_1(\theta) & \text{si } \theta \in \Theta_1 \end{cases},$$

donde g_0 y g_1 son funciones cuya integral diverge, por lo que las constantes normalizadoras c_0 y c_1 están indeterminadas. En este caso las probabilidades a priori π_0 y π_1 , serán infinitas, por lo que la razón a priori π_0/π_1 no estará definida y el factor de Bayes será de la forma

$$B = \frac{c_0 \int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{c_1 \int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)} = \frac{c_0 \int f(x|\theta) g_0(\theta) d\theta}{c_1 \int f(x|\theta) g_1(\theta) d\theta},$$

y estará definido salvo la constante c_0/c_1 , que es arbitraria.

Son muchos los métodos que se han estudiado en un intento de resolver el problema del cálculo del factor Bayes con distribuciones a priori impropias; en el capítulo 3 se hace un análisis de algunos de ellos, mostrando una especial atención a los llamados factores de Bayes intrínsecos.

Ejemplo 3 (continuación). Supongamos que el individuo al cual se le hace la prueba de inteligencia, es clasificado en razón de su CI, que puede ser menor o igual al CI medio (100), o mayor. Formalmente, se desea contrastar

$$H_0 : \theta \leq 100,$$

con respecto a

$$H_1 : \theta > 100.$$

Recordemos que la distribución a posteriori de θ es $\mathcal{N}(110, 39, 69, 23)$, estandarizando los resultados y usando tablas de la distribución normal estándar, obtenemos que

$$\alpha_0 = P(\theta \leq 100|x) = 0,106, \quad \alpha_1 = P(\theta > 100|x) = 0,894,$$

y, por lo tanto, la razón de probabilidades a posteriori es

$$\frac{\alpha_0}{\alpha_1} = \frac{0,106}{0,894} = \frac{1}{8,44} \approx 0,119.$$

Además, la distribución a priori es $\mathcal{N}(100, 225)$, así que $\pi_0 = P^\pi(\theta \leq 100) = 1/2 = \pi_1$ y la razón de probabilidades a priori es 1. Así, el factor de Bayes resulta

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{1}{8,44}.$$

Lo que significa que el coeficiente intelectual del individuo en cuestión es mayor o igual a la media poblacional, o lo que es lo mismo, debemos rechazar H_0 .

A continuación se definirán explícitamente las pruebas unilaterales, las de una hipótesis nula puntual, y la prueba de hipótesis múltiple, señalando los rasgos generales de las pruebas bayesianas y comparando los enfoques clásicos y bayesianos en cada caso.

Contraste de hipótesis unilateral. Sea $\{f(x, \theta), \theta \in \Theta\}$ una familia de funciones de densidad, $\Theta \subseteq \mathfrak{R}$. El contraste de hipótesis unilateral es de la forma

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0,$$

o equivalentemente

$$H_0^* : \theta \geq \theta_0 \quad \text{vs} \quad H_1^* : \theta < \theta_0.$$

Desde el punto de vista bayesiano este tipo de contraste no presenta ninguna característica extraordinaria, sin embargo, ésta es una de las pocas situaciones de contraste en la que el contraste clásico, particularmente el uso del p -value, a veces tiene una justificación bayesiana. Consideremos un ejemplo: Supóngase que X es una v.a. que sigue una distribución $\mathcal{N}(\theta, \sigma^2)$, con σ^2 conocida, y consideremos la distribución a priori no informativa uniforme, $\pi(\theta) = 1$. Supongamos que se desea realizar el contraste: $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.

En este caso se tiene que

$$h(x, \theta) = \pi(\theta) f(x | \theta) = f(x | \theta),$$

la distribución marginal de x , es

$$m(x) = \int_{-\infty}^{\infty} f(x | \theta) d\theta = (2\pi)^{-1/2} \sigma^{-1} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} d\theta = 1$$

y

$$\pi(\theta | x) = \frac{h(x, \theta)}{m(x)} = (2\pi)^{-1/2} \sigma^{-1} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

Por tanto la distribución a posteriori de θ dado x es una $\mathcal{N}(x, \sigma^2)$. Tipificando este resultado se tiene

$$\alpha_0 = \pi(\theta \leq \theta_0 | x) = \Phi((\theta_0 - x)/2\sigma^2),$$

donde Φ es la función de distribución de una $\mathcal{N}(0, 1)$.

Desde el enfoque clásico el p -value contra H_0 es la probabilidad, cuando $\theta = \theta_0$, de observar un X “más extremo” que el dato observado x . Puesto que $X \sim \mathcal{N}(\theta, \sigma^2)$, tipificando el resultado, el p -value será

$$p\text{-value} = P_{\theta_0}(X \geq x) = 1 - \Phi((x - \theta_0)/\sigma).$$

Como consecuencia de la simetría de la distribución normal, se sigue que α_0 es igual al p -value contra H_0 , lo cual nos indica que en este caso los resultados obtenidos mediante ambas inferencias coinciden.

Contraste de hipótesis nula puntual o simple. Es muy común en la estadística clásica (Zellner, 1984) plantear contrastes de la forma

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

Tal contraste llamado contraste de hipótesis nula puntual es interesante, particularmente por que, desde el punto de vista bayesiano, contiene algunas características

nuevas, pero principalmente porque las respuestas bayesianas difieren *radicalmente* de las respuestas clásicas.

Antes de discutir estos problemas, se harán algunos comentarios acerca de la naturaleza de este tipo de contraste. En primer lugar, los contrastes de hipótesis nula puntual son comúnmente realizados en situaciones inapropiadas. En la práctica nunca se dará el caso en el que se tome en consideración la posibilidad de que $\theta = \theta_0$ exactamente (ver Hodges y Lehmann, 1954, y Lehmann, 1959). Más razonable sería la hipótesis nula $\theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b)$, donde $b > 0$ es alguna constante elegida de tal forma que todo θ en Θ_0 pueda considerarse indistinguible de θ_0 .

Un ejemplo en el que podría presentarse esta situación sería el intento de analizar una sustancia química observando algún rasgo θ , de su reacción con otra sustancia química conocida. Desearíamos probar, por ejemplo, si realmente la sustancia química desconocida posee un componente específico, con una reacción fuerte, θ_0 , con una precisión b . Entonces sería razonable plantear la hipótesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$ contra $H_1 : \theta \notin (\theta_0 - b, \theta_0 + b)$. Un ejemplo similar que involucra la ciencia forense, puede encontrarse en Lindley, 1977 (ver también Shafer, 1982). De hecho, hay también muchos problemas de decisión que conducirían a hipótesis nulas del tipo del intervalo anterior con una b grande, pero tales problemas serán raramente bien aproximados, por un contraste de hipótesis nula puntual.

Puesto que realmente uno debería contrastar $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, sería necesario saber cuando es lícita la aproximación de H_0 por $H_0 : \theta = \theta_0$. Desde la perspectiva bayesiana, la aproximación resultará razonable si la probabilidad a posteriori de H_0 sea prácticamente igual en ambos contrastes. Además otra condición importante para que ésto se verifique, es que la función de probabilidad sea aproximadamente constante en $(\theta_0 - b, \theta_0 + b)$. Pese a que desde la perspectiva bayesiana resultará mas sencillo tratar directamente con la hipótesis del intervalo que estimar la efectividad de su aproximación por una hipótesis nula puntual, a continuación se describe este tipo de contraste, aunque antes nos detendremos en una puntualización importante.

Para llevar a cabo un contraste bayesiano de una hipótesis nula puntual $H_0 : \theta = \theta_0$, no es posible utilizar una densidad a priori continua, puesto que en este caso esta distribución a priori (y por supuesto la distribución a posteriori) darían una probabilidad de cero a θ_0 . Una aproximación razonable, por lo tanto, es dar a θ_0 una probabilidad positiva π_0 , y a $\theta \neq \theta_0$ la distribución a priori $\pi_1 g_1(\theta)$, donde $\pi_1 = 1 - \pi_0$ y g_1 es propia. Uno puede pensar en π_0 como la "masa" que se le asignaría a la hipótesis nula real $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, si no se hubiera aproximado por la hipótesis nula puntual.

La densidad marginal de X será

$$m(x) = \int f(x|\theta) dF^\pi(\theta) = f(x|\theta_0) \pi_0 + (1 - \pi_0) m_1(x),$$

donde

$$m_1(x) = \int_{\{\theta \neq \theta_0\}} f(x|\theta) dF^{g_1}(\theta),$$

es la densidad marginal de X con respecto a g_1 . Por lo tanto la probabilidad a posteriori de $\theta = \theta_0$ es

$$\pi(\theta_0|x) = \frac{f(x|\theta_0)\pi_0}{m(x)} = \frac{f(x|\theta_0)\pi_0}{f(x|\theta_0)\pi_0 + (1-\pi_0)m_1(x)} = \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

La razón de probabilidad a posteriori (recordando que $\pi_1 = 1 - \pi_0$) será,

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi(\theta_0|x)}{1 - \pi(\theta_0|x)} = \frac{\pi_0}{\pi_1} \cdot \frac{f(x|\theta_0)}{m_1(x)},$$

así el factor de Bayes para H_0 contra H_1 será

$$B = \frac{f(x|\theta_0)}{m_1(x)}.$$

Contraste de hipótesis múltiple La característica más interesante de este tipo de contraste es que no tiene mayor dificultad que el que tiene, desde el punto de vista bayesiano, el contraste de dos hipótesis. Pues tan sólo se trata de calcular la probabilidad a posteriori de cada hipótesis.

Ejemplo 3 (Continuación). Supongamos que el individuo al cual se le hace la prueba de inteligencia, es clasificado en razón de su CI, que está por debajo del CI medio (menor que 90), comprendido en la media (90 a 110) o por encima del CI medio (mayor que 110). Llamando a estas tres regiones Θ_1, Θ_2 y Θ_3 , respectivamente, y recordando que la distribución a posteriori es $\mathcal{N}(110, 39, 69, 23)$, estandarizando los resultados y usando tablas de la distribución normal estándar, obtenemos que para $x = 115$.

$$\begin{aligned} P(\Theta_1|x = 115) &= 0,007 \\ P(\Theta_2|x = 115) &= 0,473 \\ P(\Theta_3|x = 115) &= 0,520 \end{aligned}$$

lo cual nos indicaría que el nivel de inteligencia (CI) del individuo en cuestión, está por encima del CI medio.

Inferencia predictiva En muchas ocasiones el propósito de un análisis estadístico es predecir el valor de una observación futura, x_* , con base a la información disponible. Ahora bien, la distribución marginal $m(x)$, definida anteriormente, describe nuestro conocimiento a priori acerca de x , y es llamada distribución predictiva a priori.

Para tener conocimiento acerca de x_* , una observación futura, se puede calcular:

$$m(x_* | x) = \int_{\Theta} f(x_* | \theta) dF^{\pi(\theta|x)}(\theta) = \int_{\Theta} f(x_* | \theta) \pi(\theta | x) d\theta,$$

que se conoce como la *distribución predictiva a posteriori*.

Capítulo 2

Análisis bayesiano del modelo lineal

La regresión lineal es una de las herramientas estadísticas más ampliamente usadas. En este capítulo hacemos un análisis del modelo lineal normal, centrándonos en el caso simple de regresión lineal múltiple con distribuciones a priori uniformes.

Hacemos un repaso de los conceptos básicos de la estadística bayesiana, aplicándolos a los modelos de regresión lineal normales.

2.1. Introducción al modelo de regresión

Los modelos de regresión estudian la relación estocástica cuantitativa entre una variable de interés (variable respuesta) y un conjunto de variables explicativas. Cuando se estudia la relación entre una variable de interés, variable respuesta o variable dependiente \mathbf{y} , y un conjunto de variables regresoras (explicativas independientes) \mathbf{X} , pueden darse las siguientes situaciones:

i) Existe una relación funcional determinista entre ellas en el sentido de que el conocimiento de las variables regresoras determina completamente el valor que toma la variable respuesta, ésto es,

$$\mathbf{y} = g(\mathbf{X}).$$

ii) No exista ninguna relación entre la variable respuesta y las variables regresoras, en el sentido de que el conocimiento de éstas no proporciona ninguna información sobre el compartamiento de la otra.

iii) El caso intermedio, es que existe una relación estocástica entre la variable respuesta y las variables regresoras, en el sentido de que el conocimiento de éstas últimas permite predecir con mayor o menor exactitud el valor de la variable respuesta. Por tanto, siguen un modelo de la forma,

$$\mathbf{y} = g(\mathbf{X}) + \varepsilon,$$

donde ε una variable aleatoria de media cero.

Las relaciones estocásticas son las que ocurren en la mayoría de las situaciones y su estudio corresponde a los denominados modelos de regresión.

El objetivo básico en el estudio de un modelo de regresión es el de estimar la función de regresión, g , y el modelo probabilístico que sigue el vector de error aleatorio ε . En el caso de que esta forma funcional no pueda estimarse, $g(\mathbf{X})$ puede aproximarse mediante $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon$. La estimación de ambas funciones se hace a partir del conocimiento de una muestra de las variables en estudio.

Una vez estimadas estas funciones se tiene conocimiento de la relación funcional de la variable respuesta con las variables regresoras, dada por la función de regresión que se define como sigue,

$$g(X) = E(y_u|\theta, X) \quad \forall u = 1, \dots, n.$$

Esto permite tener una idea general del comportamiento de la variable respuesta en función de las variables regresoras. También, se puede predecir el valor de la variable respuesta de un individuo del que se conocen los valores de las variables regresoras.

2.2. Modelo lineal general

2.2.1. Análisis del modelo lineal

Consideremos el modelo lineal

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (2.1)$$

donde \mathbf{y} es un vector de observaciones de dimensión $n \times 1$, \mathbf{X} es una matriz de constantes conocidas de dimensión $n \times k$, $\boldsymbol{\theta}$ es un vector de coeficientes de regresión desconocidos de tamaño $k \times 1$, y $\boldsymbol{\varepsilon}$ es un vector de variables aleatorias de dimensión $n \times 1$ que tiene media cero (llamado a menudo vector de *errores*).

Explícitamente el modelo es

$$\begin{bmatrix} y_1 \\ \vdots \\ y_u \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{u2} & \cdots & x_{uk} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_u \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_u \\ \vdots \\ \varepsilon_k \end{bmatrix}, \quad (2.2)$$

y en particular para la u -ésima observación

$$y_u = \theta_1 x_{u1} + \theta_2 x_{u2} + \dots + \theta_k x_{uk} + \varepsilon_u. \quad (2.3)$$

En el modelo lineal, y es la variable dependiente, variable respuesta o la variable de salida, mientras que las x 's se refieren a las variables independientes o variables de

entrada. La relación entre la variable dependiente e independientes, consideradas como vectores, se obtiene escribiendo al modelo en la forma

$$\mathbf{y} = \mathbf{x}_1\theta_1 + \cdots + \mathbf{x}_k\theta_k + \boldsymbol{\varepsilon}, \quad (2.4)$$

donde

$$\mathbf{x}_i = (x_{1i}, \dots, x_{ui}, \dots, x_{ni})', \quad i = 1, \dots, k,$$

son las columnas de la matriz \mathbf{X} , llamada *matriz de diseño*.

El modelo tiene una amplia aplicación:

1. Se emplea frecuentemente en problemas de regresión, donde $(x_{u1}, x_{u2}, \dots, x_{uk})$ pueden ser los niveles, o funciones de los niveles, de cantidades como temperatura o presión, que han sido fijados u observados en una serie de experimentos.

2.- Igualmente se aplica, cuando algunas, o todas las \mathbf{x}_i 's son vectores de variables "indicadoras". Una x_{ui} , que toma solamente los valores uno o cero, puede usarse para indicar la presencia, o ausencia de una característica θ_i para la observación y_u .

Como un ejemplo simple, el modelo

$$y_u = \theta + \varepsilon_u, \quad u = 1, \dots, n, \quad (2.5)$$

puede escribirse de igual forma como

$$y_u = x_{1u}\theta + \varepsilon_u, \quad x_{1u} = 1, \quad u = 1, \dots, n,$$

o como

$$\mathbf{y} = \mathbf{x}_1\theta + \boldsymbol{\varepsilon}, \quad (2.6)$$

donde \mathbf{x}_1 es una columna de unos. Este modelo es un caso especial de (2.4).

Como un segundo ejemplo, supóngase que se espera que la variable respuesta sea una función cuadrática de una variable ξ , así que el modelo toma la forma

$$y_u = \theta_1 + \theta_2\xi_u + \theta_3\xi_u^2 + \varepsilon_u, \quad (2.7)$$

entonces tendríamos que

$$x_{u1} = 1, \quad x_{u2} = \xi_u, \quad x_{u3} = \xi_u^2, \quad u = 1, \dots, n,$$

y, por lo tanto,

$$\mathbf{y} = \mathbf{x}_1\theta_1 + \mathbf{x}_2\theta_2 + \mathbf{x}_3\theta_3 + \boldsymbol{\varepsilon}, \quad (2.8)$$

y otra vez este modelo es un caso especial de (2.4).

Como un ejemplo final, en los modelos del análisis de varianza con un criterio de clasificación, típicamente interesa la comparación de k tratamientos para los que las

respuestas medias son $\theta_1, \dots, \theta_k$. Supóngase que tenemos $n = \sum_{i=1}^k n_i$ observaciones, las primeras n_1 de éstas se hicieron con el tratamiento 1, las siguientes n_2 del tratamiento 2, etc. El modelo apropiado escrito en la forma de (2.2) sería

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ \dots \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \dots \\ \vdots \\ \dots \\ y_{n-n_k+1} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \\ 1 & 0 & \dots & 0 \\ \dots & \dots & & \\ 0 & \dots & 0 & 1 \\ \vdots & \vdots & & \\ & & 1 & \\ \dots & \dots & & \\ \vdots & & & 0 \\ \dots & \dots & & \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \\ 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \quad (2.9)$$

Así, los modelos asociados con regresión polinomial, regresión múltiple, análisis de varianza, diseños de bloques aleatorizados, diseños de bloques incompletos y diseños factoriales están todos inmersos en el modelo lineal general.

Si se asume que ε se distribuye como una distribución normal multivariada $\mathcal{N}(0, \sigma^2 \mathbf{I})$ y que el rango de la matrix \mathbf{X} es k , entonces

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta}, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right] \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} [vs^2 + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})]\right\}, \end{aligned} \quad (2.10)$$

donde,

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad v = n - k,$$

$$s^2 = \frac{(\mathbf{y} - \hat{\mathbf{y}})' (\mathbf{y} - \hat{\mathbf{y}})}{v}, \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}}.$$

De esto se sigue que (Box y Tiao (1973)):

- $\hat{\boldsymbol{\theta}}$ es un vector de estadísticas conjuntamente suficientes para $\boldsymbol{\theta}$ si σ^2 es conocido;
- $\hat{\boldsymbol{\theta}}$ y s^2 son conjuntamente suficientes para $(\boldsymbol{\theta}, \sigma^2)$;
- $\hat{\boldsymbol{\theta}}$ tiene una distribución normal multivariada $\mathcal{N}(\boldsymbol{\theta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$ y
- vs^2 se distribuye, independientemente de $\hat{\boldsymbol{\theta}}$, como una distribución $\sigma^2 \chi_v^2$,

$$\left(Ga \left(s^2 \mid \frac{\nu}{2}, \frac{\nu}{2\sigma^2} \right) \right).$$

En las siguientes subsecciones se distinguen dos situaciones: la primera en la que la varianza σ^2 se supone conocida, y la segunda donde σ^2 se supone desconocida.

2.2.2. Varianza σ^2 conocida

Si σ^2 es conocida, entonces la distribución a posteriori de $\boldsymbol{\theta}$ es (Box y Tiao (1973))

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \pi(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}, \sigma^2), \quad (2.11)$$

donde $\pi(\boldsymbol{\theta})$ es la distribución a priori y $\pi(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}, \sigma^2)$ es la densidad de una distribución normal multivariada $\mathcal{N}_k \left[\boldsymbol{\theta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right]$. Entonces, la distribución a posteriori para $\boldsymbol{\theta}$ es

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \exp \left[-\frac{1}{2\sigma^2} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \right], \quad (2.12)$$

y puesto que

$$\int_{\mathfrak{R}} \exp \left\{ -\frac{1}{2\sigma^2} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \right\} d\boldsymbol{\theta} = \frac{\sigma^k (\sqrt{2\pi})^k}{|\mathbf{X}'\mathbf{X}|^{1/2}},$$

donde \mathfrak{R} es la región $-\infty < \theta_i < \infty$, $i = 1, \dots, k$, se tiene que

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{|\mathbf{X}'\mathbf{X}|^{1/2}}{(\sqrt{2\pi}\sigma)^k} \exp \left[-\frac{1}{2\sigma^2} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \right], \quad -\infty < \theta_i < \infty, \quad i = 1, \dots, k. \quad (2.13)$$

Todas las inferencias relevantes sobre $\boldsymbol{\theta}$ pueden hacerse ahora con el conocimiento de que la distribución a posteriori de $\boldsymbol{\theta}$, es la distribución normal multivariada $\mathcal{N}_k \left[\widehat{\boldsymbol{\theta}}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right]$.

Algunas propiedades de la distribución de $\boldsymbol{\theta}$

La función de densidad $\pi(\boldsymbol{\theta}|\mathbf{y})$ es una función monótona decreciente de la forma cuadrática definida positiva

$$Q(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}). \quad (2.14)$$

De las propiedades de la distribución normal multivariada, Wilks (1962), se tiene que la forma cuadrática $Q(\boldsymbol{\theta})$ se distribuye a posteriori, como $\sigma^2 \chi_k^2$.

Denotando (haciendo una partición)

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{k-r}^r, \quad \widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{bmatrix}_{k-r}^r, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}'_1\mathbf{X}_1 & \mathbf{X}'_1\mathbf{X}_2 \\ \mathbf{X}'_2\mathbf{X}_1 & \mathbf{X}'_2\mathbf{X}_2 \end{bmatrix}_{k-r}^r \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}_{k-r}^r, \quad (2.15)$$

donde $1 \leq r \leq k$. Se sigue de las propiedades de la distribución normal multivariada que

a) La distribución marginal a posteriori de un subconjunto de $\boldsymbol{\theta}$, digamos θ_1 , es la distribución normal multivariada $\mathcal{N}(\widehat{\boldsymbol{\theta}}_1, \sigma^2 \mathbf{C}_{11})$ y, en particular, la distribución marginal de θ_i es $\mathcal{N}(\widehat{\theta}_i, \sigma^2 C_{ii})$, donde C_{ii} es el i -ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$.

b) La distribución condicional a posteriori de θ_2 , dado θ_1 , es la distribución normal multivariada $\mathcal{N}_{k-r}[\widehat{\boldsymbol{\theta}}_{2,1}, \sigma^2 (\mathbf{X}'_2\mathbf{X}_2)^{-1}]$, donde

$$\widehat{\boldsymbol{\theta}}_{2,1} = \widehat{\boldsymbol{\theta}}_2 - (\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{X}_1(\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1). \quad (2.16)$$

2.2.3. Varianza σ^2 desconocida

En muchos casos la varianza σ^2 no es conocida, y entoces se usa la información acerca de σ^2 que se obtiene de la muestra. Por el lema del apéndice A2.1, la distribución a posteriori de $(\boldsymbol{\theta}, \sigma^2)$ es

$$\pi(\boldsymbol{\theta}, \sigma^2 | \mathbf{y}) \propto \pi(\boldsymbol{\theta}, \sigma^2) \pi(s^2 | \sigma^2) \pi(\widehat{\boldsymbol{\theta}} | \boldsymbol{\theta}, \sigma^2). \quad (2.17)$$

En el modelo de regresión normal, una distribución a priori no informativa conveniente es la distribución uniforme sobre $(\boldsymbol{\theta}, \log \sigma)$ (Box y Tiao (1973))

$$\pi(\boldsymbol{\theta}, \log \sigma) \propto c.$$

Equivalentemente

$$\pi(\boldsymbol{\theta}, \sigma) \propto \sigma^{-1},$$

en consecuencia

$$\pi(\boldsymbol{\theta}, \sigma^2) \propto \sigma^{-2}. \quad (2.18)$$

Sustituyendo (2.18), en (2.17) y por el mismo argumento del apéndice A2.1, la distribución conjunta a posteriori de $(\boldsymbol{\theta}, \sigma^2)$ puede factorizarse de la siguiente manera

$$\pi(\boldsymbol{\theta}, \sigma^2 | \mathbf{y}) = \pi(\sigma^2 | s^2) \pi(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}, \sigma^2), \quad (2.19)$$

donde la distribución marginal a posteriori de σ^2 es $vs^2\chi_v^{-2}$, $v = n - k$, y la distribución a posteriori condicional de θ , dado σ^2 , es la distribución normal multivariada dada por (2.13).

La distribución marginal a posteriori de θ es

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{y}) &= \int \pi(\boldsymbol{\theta}, \sigma^2|\mathbf{y}) d\sigma^2 \\ &= \frac{\Gamma\left[\frac{1}{2}(v+k) \left|\mathbf{X}'\mathbf{X}\right|^{1/2} s^{-k}\right]}{\left[\Gamma\left(\frac{1}{2}\right)\right]^k \Gamma\left(\frac{1}{2}v\right) (\sqrt{v})^k} \left[1 + \frac{(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})}{vs^2}\right]^{-\frac{1}{2}(v+k)}, \end{aligned} \quad (2.20)$$

$$-\infty < \theta_i < \infty \quad i = 1, \dots, k,$$

que es la distribución t multivariada descubierta independientemente por Cornish (1954) y por Dunnett y Sobel (1954). Denotemos la distribución (2.20) por

$$t_k \left[\widehat{\boldsymbol{\theta}}, s^2(\mathbf{X}'\mathbf{X})^{-1}, v \right].$$

A continuación se muestran algunas de las propiedades más relevantes de la distribución marginal de θ .

Algunas propiedades de la distribución de θ .

1. La función de densidad en (2.20) es una función monótona decreciente de la forma cuadrática

$$Q(\boldsymbol{\theta}) = (\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}).$$

2. La cantidad $Q(\boldsymbol{\theta})/ks^2$ se distribuye a posteriori como una distribución F con (k, v) grados de libertad.
3. La distribución marginal de un subconjunto r -dimensional, θ_1 , tiene una distribución t multivariada

$$t_r \left(\widehat{\boldsymbol{\theta}}_1, s^2 \mathbf{C}_{11}, v \right),$$

esto es,

$$\begin{aligned} \pi(\boldsymbol{\theta}_1|\mathbf{y}) &= \\ &= \frac{\Gamma\left[\frac{1}{2}(v+r) \left|\mathbf{C}_{11}^{-1}\right|^{1/2} s^{-r}\right]}{\left[\Gamma\left(\frac{1}{2}\right)\right]^r \Gamma\left(\frac{1}{2}v\right) (\sqrt{v})^r} \left[1 + \frac{(\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1)' \mathbf{C}_{11}^{-1} (\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1)}{vs^2}\right]^{-\frac{1}{2}(v+r)}, \end{aligned} \quad (2.21)$$

$$-\infty < \theta_i < \infty, \quad i = 1, \dots, r,$$

donde la notación es la misma que la de (2.15).

En particular θ_i tiene la distribución $t\left(\hat{\theta}_i, s^2 c_{ii}, v\right)$; esto es,

$$t = \frac{\theta_i - \hat{\theta}_i}{s\sqrt{c_{ii}}} \quad (2.22)$$

tiene la distribución $t - Student$ univariada con $v = n - k$ grados de libertad.

4. La distribución a posteriori condicional de un subconjunto de dimensión $(k - r)$ de los parámetros θ_2 , dado θ_1 , es también una distribución t multivariada. Específicamente, θ_2 , tiene la distribución

$$t_{k-r} \left[\hat{\theta}_{2,1}, s_{2,1}^2 \left(\mathbf{X}'_2 \mathbf{X}_2 \right)^{-1}, v + r \right],$$

donde

$$\hat{\theta}_{2,1} = \hat{\theta}_2 - \left(\mathbf{X}'_2 \mathbf{X}_2 \right)^{-1} \mathbf{X}'_2 \mathbf{X}_1 \left(\theta_1 - \hat{\theta}_1 \right),$$

y

$$s_{2,1}^2 = (v + r)^{-1} \left[v s^2 + \left(\theta_1 - \hat{\theta}_1 \right)' \mathbf{C}_{11}^{-1} \left(\theta_1 - \hat{\theta}_1 \right) \right].$$

5. Supóngase que $\Phi = (\phi_1, \phi_2, \dots, \phi_m)'$, $m \leq k$, es un conjunto de parámetros de tal forma que

$$\Phi = \mathbf{D}\theta,$$

donde \mathbf{D} es una matriz de rango m . Entonces la distribución a posteriori de Φ es

$$\Phi \sim t_m \left[\mathbf{D}\hat{\theta}, s^2 \mathbf{D} \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{D}', v \right].$$

Ejemplo 4 *Un método microscópico para pesar objetos sumamente ligeros, se cree tiene un error que es aproximadamente normal. También se cree que sobre el rango pertinente de pesos, el error tiene media cero y una varianza constante. Se obtuvieron los siguientes datos de un experimento donde se hicieron 18 mediciones del peso de dos especímenes, A y B. Se obtuvieron 2 mediciones del peso para el espécimen A solo, 9 para el espécimen B solo y 7 mediciones del peso para los especímenes A y B juntos.*

A solo	B solo		A y B juntos	
109 ⁽⁸⁾	114 ⁽¹⁾	129 ⁽⁵⁾	217 ⁽²⁾	233 ⁽³⁾
85 ⁽⁴⁾	121 ⁽⁶⁾	98 ⁽¹¹⁾	203 ⁽⁷⁾	221 ⁽⁹⁾
	140 ⁽¹²⁾	134 ⁽¹⁴⁾	243 ⁽¹⁰⁾	221 ⁽¹³⁾
	122 ⁽¹⁵⁾	133 ⁽¹⁶⁾	229 ⁽¹⁷⁾	
	125 ⁽¹⁸⁾			

Los números entre paréntesis asociados a cada peso, se refiere al orden en el que se hicieron las mediciones de los pesos. En el presente análisis este hecho es irrelevante.

A manera de ilustrar lo planteado hasta aquí, sin el propósito de hacer un análisis estadístico completo sino sólo un análisis inferencial inicial, se propone el siguiente modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

donde \mathbf{X} , una matriz de dimensión 18×2 y el vector de observaciones \mathbf{y} se muestran a continuación.

$$\mathbf{y} = \begin{bmatrix} 109 \\ 85 \\ 114 \\ 129 \\ 121 \\ 98 \\ 140 \\ 134 \\ 122 \\ 133 \\ 125 \\ 217 \\ 233 \\ 203 \\ 221 \\ 243 \\ 221 \\ 229 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Para el análisis, tenemos que

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{bmatrix} 9 & 7 \\ 7 & 16 \end{bmatrix} & (\mathbf{X}'\mathbf{X})^{-1} &= \begin{bmatrix} 0,1684 & -0,0737 \\ -0,0737 & 0,0947 \end{bmatrix} \\ \mathbf{X}'\mathbf{y} &= \begin{bmatrix} 1761 \\ 2683 \end{bmatrix} & \hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} &= \begin{bmatrix} 98,895 \\ 124,421 \end{bmatrix} \end{aligned}$$

De aquí tenemos que

$$\sum_{i=1}^{18} y_i^2 = 510521 \quad \hat{\boldsymbol{\theta}}' \mathbf{X}' \mathbf{y} = 507976,$$

$$\sum_{i=1}^{18} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{18} y_i^2 - \hat{\boldsymbol{\theta}}' \mathbf{X}' \mathbf{y} = 2525,$$

$$s^2 = 157,8 \quad s = 12,56,$$

por lo que

$$\begin{aligned} Q(\boldsymbol{\theta}) &= (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \\ &= 9(\theta_1 - 98,9)^2 + 14(\theta_1 - 98,9)(\theta_2 - 124,4) + 16(\theta_2 - 124,4)^2. \end{aligned}$$

Dados los supuestos y los datos, todo lo que puede decirse sobre el vector de parámetros $\boldsymbol{\theta}' = (\theta_1, \theta_2)$ desde la perspectiva de la distribución a priori no informativa $\pi(\boldsymbol{\theta}, \sigma^2) \propto \sigma^{-2}$, está contenida en la distribución a posteriori conjunta de (θ_1, θ_2) , que por (2.20) es una distribución t bivariada

$$t_2 \left\{ \begin{bmatrix} 98,9 \\ 124,4 \end{bmatrix}, 157,8 \begin{bmatrix} 0,1684 & -0,0737 \\ -0,0737 & 0,0947 \end{bmatrix}, 16 \right\}$$

Las distribuciones de probabilidad marginales para θ_1 y θ_2 de acuerdo a la ecuación (2.22) son: para θ_1 la distribución

$$t(98,9, 26,57352, 16)$$

y para θ_2 la distribución

$$t(124,4, 14,94366, 16).$$

El problema para hacer inferencias acerca de $\boldsymbol{\theta}$ puede resumirse en la siguiente forma.

Dados los supuestos, todo lo que podemos conocer a cerca de $\boldsymbol{\theta} = (\theta_1, \theta_2)$ está contenido en la distribución conjunta a posteriori t bivariada.

2.2.4. Conjuntos creíbles HPD para el modelo lineal

En la subsección 2.2.3 veíamos en (2.20) que en relación a una distribución a priori no informativa en $\boldsymbol{\theta}$ y σ la distribución a posteriori de $\boldsymbol{\theta}$ es una distribución t - *multivariado*. Además, también vimos que la cantidad

$$Q(\boldsymbol{\theta}) / ks^2 = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) / ks^2,$$

se distribuye a posteriori como una distribución F con (k, v) grados de libertad. Supongamos que ahora estamos interesados en la siguiente cuestión:

¿Está el punto paramétrico $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \dots, \theta_{k0})$ en el conjunto creíble de máxima densidad al $(1 - \alpha)\%$?

Tenemos que $\pi(\boldsymbol{\theta} | \mathbf{y})$ es una función monótona decreciente de la cantidad

$$(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) / ks^2.$$

El punto particular $\boldsymbol{\theta}_0$ está incluido en el conjunto creíble HPD $(1 - \alpha)\%$ si y solamente sí

$$(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}) < ks^2 F(k, v, \alpha). \quad (2.23)$$

Ejemplo 4 (Continuación). Supóngase que el interés particular ha sido asociado con los valores del parámetro $\theta_{10} = \theta_{20} = 120$. Tal interés podría surgir debido a una teoría, que diga que ambos especímenes deberían pesar 120 microgramos. Una pregunta interesante podría ser si el punto (120, 120) ha sido excluido o no, en, por ejemplo, el conjunto creíble HPD al 95 %. Hay que considerar si la cantidad $Q(\boldsymbol{\theta}_0)/2s^2$ es mayor o menor que $F(k, v, \alpha)$. Así, tenemos lo siguiente

$$\begin{aligned} & \frac{Q(\boldsymbol{\theta}_0)}{2s^2} \\ = & \frac{9(120 - 98,9)^2 + 14(120 - 98,9)(120 - 124,4) + 16(120 - 124,4)^2}{2 \times 157,8} \\ = & \frac{3015}{315,6} = 9,56, \end{aligned}$$

que es mas grande que $F(2, 16, 0,05) = 3,63$ así que el punto (120, 120) ha sido excluido.

Para propósitos de cálculo y verificación, es útil indicar las cantidades necesarias en una tabla de análisis de varianza como se muestra a continuación. Mientras dicha tabla habitualmente se interpreta en términos de teoría muestral, ésta también tiene una función útil en un enfoque bayesiano. En particular, la pregunta de si θ_0 está, o no está en un conjunto creíble HPD se determina calculando el cociente de los cuadrados medios en la última columna de la tabla, y tomando el valor de $F(k, v, \alpha)$ de la tabla de la distribución F .

Análisis de varianza para determinar si el punto ($\theta_{10} = 120, \theta_{20} = 120$) está incluido dentro de un conjunto creíble HPD, dado.

Fuente	Suma de Cuadrados	G. de L.	C. M.	F
Dif. de parámetros	$(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) = 3,015$	2	1,507.5	
Residual	$(\mathbf{y} - \hat{\mathbf{y}})' (\mathbf{y} - \hat{\mathbf{y}}) = 2,525$	16	157.8	9,56
Total	$(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_0)' (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_0) = 5,540$	18		

2.2.5. Introducción al contraste de hipótesis como selección de modelos

En relación al ejemplo de la subsección anterior, supóngase que estamos interesados en el contraste

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Entonces, en realidad estaríamos comparando dos modelos, que serían

$$M_1: \{f(\mathbf{y}|\theta_{10}, \theta_{20}, \dots, \theta_{k0}, \sigma_1), \pi_1^N(\theta_{10}, \theta_{20}, \dots, \theta_{k0}; \sigma_1)\}$$

$$M_2: \{f(\mathbf{y}|\theta_1, \theta_2, \dots, \theta_k, \sigma_2), \pi_2^N(\theta_1, \theta_2, \dots, \theta_k; \sigma_2)\}$$

tomando distribuciones a priori de la forma

$$\pi_1^N(\theta_{10}, \theta_{20}, \dots, \theta_{k0}; \sigma_1) = \frac{c_1}{\sigma_1}, \quad \pi_2^N(\theta_1, \theta_2, \dots, \theta_k; \sigma_2) = \frac{c_2}{\sigma_2},$$

y probabilidades a priori para los modelos

$$P[M_1] = p \quad P[M_2] = 1 - p,$$

entonces la tarea de aceptar o rechazar lo que la evidencia de los datos nos dice, se reduce a calcular las probabilidades a posteriori $\alpha_0 = P(M_1|x)$, $\alpha_1 = P(M_2|x)$ y el factor de Bayes, (ver subsección 1.2.2.).

Capítulo 3

Factores de Bayes intrínsecos.

En el contexto de selección de modelos o comparación de modelos, éste puede considerarse como un problema de decisión estadístico. Existen diferentes criterios de resolver el problema. En Bernardo y Smith (1994) se plantean tres criterios de éstos. El primero, el cual es llamado criterio \mathcal{M} -cerrado, se cree que uno de los modelos $\{M_i, i \in I\}$ es el verdadero, sin el conocimiento explícito de cuál de ellos es el verdadero modelo. Desde éste criterio tiene sentido asignar probabilidades a los modelos, cuyas probabilidades son las que cree el estadístico, por su experiencia, que se tienen.

El segundo criterio, el cuál es llamado, \mathcal{M} -completo, corresponde a una actuación individual, como si $\{M_i, i \in I\}$, fuera un conjunto de modelos específicos, disponibles para comparación de modelos. Desde esta perspectiva no tiene sentido asignar probabilidades a los modelos.

El tercer criterio, llamado \mathcal{M} -abierto, también se considera que el conjunto de modelos son simplemente un rango de modelos específicos, disponibles para comparación, tampoco tiene sentido asignarles probabilidades a los modelos, puesto que no se tiene creencia que el verdadero modelo esté en el conjunto considerado.

Los problemas de decisión estadístico, envuelven éste problema, introduciendo una función de utilidad y una distribución inicial sobre el espacio de decisiones relevantes al problema. La función de utilidad cuantifica las consecuencias de una acción particular de un espacio de decisiones.

Más sin embargo en el presente trabajo se considera solamente el criterio \mathcal{M} -cerrado, y se resuelve el problema de comparación de modelos usando probabilidades a posteriori y factor de Bayes.

Por otro lado, entre los diferentes métodos bayesianos que se presentan en la literatura para la comparación de modelos o el contraste de hipótesis, destacan los procedimientos basados en los factores de Bayes. Sin embargo, estos procedimientos tienen la limitación de que si se trabaja con distribuciones a priori impropias, entonces los factores de Bayes quedan definidos salvo una constante multiplicativa. Esto ha llevado a los bayesianos a usar distribuciones a priori propias convencionales o aproximaciones ordinarias al factor de Bayes, tales como el criterio de información bayesiana de Schwarz.

El factor de Bayes intrínseco introducido por Berger y Pericchi (1993, 1996) es un método interesante para resolver esta dificultad.

En este capítulo estudiaremos algunos de los diferentes métodos propuestos en la literatura para solventar dicho problema, haciendo un análisis exhaustivo del método propuesto por Berger y Pericchi (1993, 1996), llamado método intrínseco, que es totalmente automático en el sentido de requerir solamente distribuciones a priori no informativas estándares, para generar distribuciones a priori propias (a priori intrínsecas) y que parece corresponder a un factor de Bayes verdadero, al menos asintóticamente.

3.1. Preliminares

Consideremos los modelos M_1, M_2, \dots, M_n , de tal forma que la variable aleatoria correspondiente a los datos X tiene densidad $f_i(x|\theta_i)$ bajo el modelo M_i y, los vectores de parámetros θ_i son desconocidos y de dimensión k_i .

En el enfoque bayesiano para comparación de modelos, en el enfoque \mathcal{M} -cerrado (Bernardo y Smith, (1993)), se considera que uno de los modelos a comparar es el verdadero, sin el conocimiento explícito de cuál de ellos lo es. Así, se selecciona una distribución a priori $\pi_i(\theta_i)$ para los parámetros de cada modelo, junto con probabilidades a priori p_i de que cada modelo sea el verdadero. La probabilidad a posteriori de que M_i sea el verdadero modelo sería

$$P(M_i|x) = \left(\sum_{j=1}^n \frac{p_j}{p_i} \cdot B_{ji} \right)^{-1}, \quad (3.1)$$

ya que

$$\begin{aligned} P(M_i|x) &= \frac{P(x|M_i) P(M_i)}{\sum_{j=1}^n P(x|M_j) P(M_j)} = \frac{m_i(x) p_i}{\sum_{j=1}^n m_j(x) p_j} \\ &= \left(\sum_{j=1}^n \frac{m_j(x) p_j}{m_i(x) p_i} \right)^{-1} = \left(\sum_{j=1}^n \frac{p_j}{p_i} \cdot B_{ji} \right)^{-1}, \end{aligned}$$

donde B_{ji} es el factor de Bayes de M_j frente a M_i , definido como

$$B_{ji}(x) = \frac{m_j(x)}{m_i(x)} = \frac{\int f_j(x|\theta_j) \pi_j(\theta_j) d\theta_j}{\int f_i(x|\theta_i) \pi_i(\theta_i) d\theta_i}, \quad (3.2)$$

y $m_j(x)$ es la densidad marginal de X bajo el modelo M_j .

Nótese que el factor de Bayes facilita la comparación entre varios modelos, independientemente de qué modelos estén bajo consideración. Si nosotros cambiamos la lista de posibles modelos, por ejemplo, agregando un nuevo contendiente M_{n+1} , entonces

la probabilidad a priori p_i y la probabilidad a posteriori cambiarán, pero el factor de Bayes no se alterará.

Aunque en el problema anterior se ha usado un lenguaje bayesiano estándar, nótese que uno no tiene que asumir estrictamente que uno de los modelos a comparar, sea el verdadero modelo; en particular, B_{ji} puede verse como la razón de verosimilitudes “ponderada” de M_j frente a M_i y por lo tanto puede interpretarse solamente en términos de soporte comparativo de los datos para los dos modelos. (Ver Kass y Raftery, 1995) para discusión y referencias concernientes a este punto de vista). Aunque en este capítulo sólo se discutirá formalmente el problema de selección de modelos, este desarrollo también puede aplicarse al contraste de hipótesis, que se puede plantear como un caso particular de selección de modelos, como veamos en la subsección 2.2.5.

Para el cálculo del factor de Bayes a menudo se usa como distribución a priori una distribución no informativa $\pi_i^N(\theta_i)$, por ejemplo:

- i) Distribución a priori “uniforme”, $\pi_i^U(\theta_i) = 1$.
- ii) Distribución a priori de Jeffreys, $\pi_i^J(\theta_i) = (\det(I_i(\theta_i)))^{\frac{1}{2}}$, donde $I_i(\theta_i)$ es la matriz de información esperada de Fisher, correspondiente al modelo M_i .
- iii) Distribución a priori de referencia, $\pi_i^R(\theta_i)$. Bernardo(1979) y Berger y Bernardo (1992).

Usando cualquiera de las distribuciones a priori anteriores, el factor de Bayes que se produce es

$$B_{ji}^N(x) = \frac{m_j^N(x)}{m_i^N(x)} = \frac{\int f_j(x|\theta_j) \pi_j^N(\theta_j) d\theta_j}{\int f_i(x|\theta_i) \pi_i^N(\theta_i) d\theta_i}. \quad (3.3)$$

La dificultad con esta expresión es que $\pi_i^N(\theta_i)$ son típicamente impropias, y por lo tanto, definidas solamente salvo una constante c_i . Por lo tanto B_{ji}^N está definido solamente salvo una constante

$$c_j/c_i,$$

que es arbitraria. (Berger y Pericchi; (1993), (1996)).

La inclusión de estas c_i es debatible, y esto no exime el hecho de poder considerar el valor de 1 para estas constantes y poder realizar los cálculos pertinentes.

Una solución común a este problema se basa en usar parte de los datos como una **muestra de entrenamiento**. La idea es dividir la muestra \mathbf{x} , en dos partes $\mathbf{x} = (x(l), x(n-l))$, y $x(l)$, la muestra de entrenamiento, es usada para convertir $\pi_i^N(\theta_i)$ en distribuciones a posteriori propias de la forma

$$\pi_i^N(\theta_i|x(l)) = \frac{f_i(x(l)|\theta_i) \pi_i^N(\theta_i)}{m_i^N(x(l))},$$

donde, (abusando un poco de la notación) $f_i(x(l)|\theta_i)$ es la densidad marginal de $X(l)$ bajo M_i y

$$m_i^N(x(l)) = \int f_i(x(l)|\theta_i) \pi_i^N(\theta_i) d\theta_i. \quad (3.4)$$

De aquí la idea es calcular el factor de Bayes con el resto de los datos, usando $\pi_i^N(\theta_i|x(l))$ como distribución a priori. Tomando en cuenta lo anterior, enunciemos el siguiente lema.

Lema 1 *El factor de Bayes del modelo j frente al modelo i , condicional sobre la muestra de entrenamiento $x(l)$ y asumiendo que $\pi_i^N(\theta_i|x(l))$ es propia, está dado por:*

$$B_{ji}(x(n-l)|x(l)) = B_{ji}(l) = B_{ji}^N(x) \cdot B_{ij}^N(x(l)), \quad (3.5)$$

donde

$$B_{ij}^N(x(l)) = \frac{m_i^N(x(l))}{m_j^N(x(l))}.$$

Demostración

$$B_{ji}(l) = \frac{\int f_j(x(n-l)|\theta_j, x(l)) \pi_j^N(\theta_j|x(l)) d\theta_j}{\int f_i(x(n-l)|\theta_i, x(l)) \pi_i^N(\theta_i|x(l)) d\theta_i},$$

por otro lado,

$$f_i(x(n-l)|\theta_i, x(l)) f_i(x(l)|\theta_i) = f_i(x|\theta_i),$$

y

$$\pi_i^N(\theta_i|x(l)) = \frac{f_i(x(l)|\theta_i) \pi_i^N(\theta_i)}{m_i^N(x(l))},$$

análogamente tenemos que

$$f_j(x(n-l)|\theta_j, x(l)) f_j(x(l)|\theta_j) = f_j(x|\theta_j),$$

y

$$\pi_j^N(\theta_j|x(l)) = \frac{f_j(x(l)|\theta_j) \pi_j^N(\theta_j)}{m_j^N(x(l))},$$

por lo que *sustituyendo en la ecuación de $B_{ji}(l)$ tenemos que*

$$\begin{aligned}
B_{ji}(l) &= \frac{\int f_j(x(n-l)|\theta_j, x(l)) \pi_j^N(\theta_j|x(l)) d\theta_j}{\int f_i(x(n-l)|\theta_i, x(l)) \pi_i^N(\theta_i|x(l)) d\theta_i} \\
&= \frac{\int \frac{f_j(x|\theta_j)}{f_j(x(l)|\theta_j)} \frac{f_j(x(l)|\theta_j) \pi_j^N(\theta_j)}{m_j^N(x(l))} d\theta_j}{\int \frac{f_i(x|\theta_i)}{f_i(x(l)|\theta_i)} \frac{f_i(x(l)|\theta_i) \pi_i^N(\theta_i)}{m_i^N(x(l))} d\theta_i} \\
&= \frac{\int \frac{f_j(x|\theta_j) \pi_j^N(\theta_j)}{m_j^N(x(l))} d\theta_j}{\int \frac{f_i(x|\theta_i) \pi_i^N(\theta_i)}{m_i^N(x(l))} d\theta_i} \\
&= \frac{\int f_j(x|\theta_j) \pi_j^N(\theta_j) d\theta_j}{\int f_i(x|\theta_i) \pi_i^N(\theta_i) d\theta_i} \cdot \frac{m_i^N(x(l))}{m_j^N(x(l))} \\
&= B_{ji}^N(x) \cdot B_{ij}^N(x(l)).
\end{aligned}$$

El factor de Bayes

$$B_{ji}(l) = \frac{\int f_j(x|\theta_j) \pi_j^N(\theta_j) d\theta_j}{\int f_i(x|\theta_i) \pi_i^N(\theta_i) d\theta_i} \cdot \frac{m_i^N(x(l))}{m_j^N(x(l))},$$

fue propuesto por Lempers (1971) y fue llamado factor de Bayes parcial del modelo M_j frente al modelo M_i por O'Hagan (1995). Si la muestra de entrenamiento produce distribuciones a posteriori propias, entonces este factor de Bayes está propiamente definido, aunque las distribuciones a priori $\pi_i^N(\theta_i)$ sean impropias. Considerando que en ese sentido el factor de Bayes parcial proporciona una solución al problema, ésta es poco satisfactoria, porque confía en una división arbitraria de los datos, es decir, en los datos de la muestra de entrenamiento $x(l)$ y los datos de comparación de modelos $x(n-l)$.

Claramente, en (3.5) vemos que la arbitrariedad en la elección de las constantes multiplicativas de $\pi_k^N(\theta_k)$, $k = i, j$, se elimina, ya que el cociente arbitrario c_j/c_i que multiplica a B_{ji}^N se cancela con el cociente c_i/c_j que multiplica a $B_{ij}^N(x(l))$. El factor de Bayes parcial, corrige, al factor de Bayes original B_{ji}^N con el término $B_{ij}^N(x(l))$, cancelando las constante y resolviendo el problema. Pero el uso de *muestras de entrenamiento* sólo tiene sentido si $m_i^N(x(l))$ en (3.4) es finita. Esto se formaliza en la siguiente definición.

Definición 8 Una muestra de entrenamiento $x(l)$, se llama “propia” si

$$0 < m_i^N(x(l)) < \infty$$

para todo M_i , y es minimal, si es propia y ningún subconjunto de $x(l)$ es propio.

Ejemplo 5 Supongamos que $X = (X_1, X_2, \dots, X_n)$, donde X_i son iid $\mathcal{N}(0, \sigma_1^2)$ bajo el modelo M_1 e iid $\mathcal{N}(\mu, \sigma_2^2)$ bajo el modelo M_2 . Consideremos las distribuciones a priori no informativas $\pi_1^N(\sigma_1) = 1/\sigma_1$ y $\pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$.

Aunque $m_1^N(x_i) = \frac{1}{2\sqrt{x_i}}$, es fácil demostrar que $m_2^N(x_i) = \infty$. A continuación mostramos estos resultados.

Considerando $p = 1, a = \frac{x_i}{2}$, en el apéndice A3.4 tenemos que

$$\begin{aligned} m_1^N(x_i) &= \int_{\Theta_1} f_1(x_i|\theta_1) \pi_1^N(\theta_1) d\theta_1 = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}x_i^2} \cdot \frac{1}{\sigma_1^2} d\sigma_1 \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \sigma_1^{-2} e^{-\frac{x_i^2}{2}\sigma_1^{-2}} d\sigma_1 = \frac{\frac{1}{2}\left(\frac{x_i}{2}\right)^{-1/2} \sqrt{\pi}}{\sqrt{2\pi}} = \frac{1}{2\sqrt{x_i}}, \end{aligned}$$

por otro lado tenemos que, haciendo el cambio de variable $y = (x_i - \mu)/\sigma_2$; $y^2 = (x_i - \mu)^2/\sigma_2^2$; $dy = -d\mu/\sigma_2$; resulta

$$\begin{aligned} m_2^N(x_i) &= \int_{\Theta_2} f_2(x_i|\theta_2) \pi_2^N(\theta_2) d\theta_2 \\ &= \int_{-\infty}^\infty \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}(x_i-\mu)^2} \cdot \frac{1}{\sigma_2^2} d\mu d\sigma_2 \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_2^2} \int_{-\infty}^\infty e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma_2}\right)^2} \cdot \frac{1}{\sigma_2} d\mu d\sigma_2 \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_2^2} \left[\int_{-\infty}^\infty e^{-\frac{1}{2}y^2} dy \right] d\sigma_2 \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_2^2} \cdot \sqrt{2\pi} d\sigma_2 \\ &= \int_0^\infty \sigma_2^{-2} d\sigma_2 = \left[-\frac{1}{\sigma_2} \right]_0^\infty = \infty. \end{aligned}$$

Por lo tanto una sola observación no es una “muestra de entrenamiento”. Las muestras formadas por dos o más observaciones distintas sí lo son. A continuación verificamos esta afirmación.

$$m_1^N(x(l)) = \frac{1}{2\pi(x_i^2 + x_j^2)}, \quad y, \quad m_2^N(x(l)) = \frac{1}{\sqrt{\pi}(x_i - x_j)^2}. \quad (3.6)$$

Las densidades marginales de $X(l)$ bajo los modelos M_1 y M_2 se obtuvieron de la siguiente forma.

$$m_1^N(x(l)) = \int_{\Theta_1} f_1(x(l)|\theta_1) \pi_1^N(\theta_1) d\theta_1 = \int_0^\infty \frac{1}{2\pi} e^{-\frac{1}{2\sigma_1^2}(x_i^2+x_j^2)} \cdot \frac{1}{\sigma_1^3} d\sigma_1,$$

tomando $u = -\frac{1}{2\sigma_1^2}(x_i^2 + x_j^2)$; $du = -\frac{(x_i^2+x_j^2)}{\sigma_1^3} d\sigma_1$;

$$\begin{aligned}
m_1^N(x_i, x_j) &= \int_{-\infty}^0 \frac{1}{2\pi} e^u \frac{du}{(x_i^2 + x_j^2)} = \frac{1}{2\pi (x_i^2 + x_j^2)} \int_{-\infty}^0 e^u du \\
&= \frac{1}{2\pi (x_i^2 + x_j^2)};
\end{aligned}$$

por otro lado,

$$\begin{aligned}
m_2^N(x(l)) &= \int_{\Theta_2} f_2(x(l) | \theta_2) \pi_2^N(\theta_2) d\theta_2 \\
&= \int_{-\infty}^{\infty} \int_0^{\infty} \frac{1}{2\pi\sigma_2^2} e^{-\frac{1}{2\sigma_2^2} (2([\mu - \frac{1}{2}(x_i - x_j)])^2 + \frac{1}{2}(x_i - x_j)^2)} \cdot \frac{1}{\sigma_2^2} d\sigma_2 d\mu \\
&= \int_0^{\infty} \frac{e^{-\frac{1}{4\sigma_2^2}(x_i - x_j)^2}}{2\pi} \cdot \frac{1}{\sigma_2^3} \left[\int_{-\infty}^{\infty} e^{-\frac{1}{\sigma_2^2} [\mu - \frac{1}{2}(x_i - x_j)]^2} \cdot \frac{1}{\sigma_2} d\mu \right] d\sigma_2,
\end{aligned}$$

tomando $y = (\mu - \frac{1}{2}(x_i - x_j)) / \sigma_2$; $dy = d\mu / \sigma_2$;

$$m_2^N(x(l)) = \int_0^{\infty} \frac{e^{-\frac{1}{4\sigma_2^2}(x_i - x_j)^2}}{2\pi} \cdot \frac{1}{\sigma_2^3} \cdot \sqrt{\pi} \cdot d\sigma_2 = \frac{1}{\sqrt{\pi}} \int_0^{\infty} e^{-\frac{1}{4\sigma_2^2}(x_i - x_j)^2} \cdot \frac{1}{2\sigma_2^3} d\sigma_2,$$

haciendo $u = -\frac{1}{4\sigma_2^2}(x_i - x_j)^2$; $du = \frac{(x_i - x_j)^2}{2\sigma_2^3} d\sigma_2$ se tiene que

$$m_2^N(x(l)) = \frac{1}{\sqrt{\pi}(x_i - x_j)^2} \int_{-\infty}^0 e^u du = \frac{1}{\sqrt{\pi}(x_i - x_j)^2}.$$

Lo cual demuestra el resultado (3.6).

Generalmente para una muestra de entrenamiento minimal, todos los parámetros en todos los modelos deben ser identificables y a menudo es una muestra de tamaño máx $\{k_i\}$; recordemos que k_i es la dimensión de θ_i . Sin embargo, puede ser una muestra más pequeña, especialmente si las distribuciones a priori π_i^N son propias en algunas variables. De hecho, si las distribuciones a priori π_i^N son propias, entonces el conjunto de las muestras de entrenamiento minimales es el conjunto vacío.

3.2. Otros métodos bayesianos

En los problemas de selección de modelos y contraste de hipótesis, el uso de distribuciones a priori impropias ha llevado a algunos bayesianos a utilizar distribuciones a priori propias convencionales o modificaciones de los factores de Bayes. Entre los

factores de bayes propuesto para solventar dicho problema, destacan el “criterio de información bayesiano” (CIB), propuesto por Schwarz (1978), la aproximación de Jeffreys (1961) y el método de Smith and Spiegelhalter (1980). A continuación se muestran algunos métodos bayesianos “automáticos”, para la solución de dicho problema.

3.2.1. Distribuciones a priori convencionales

Jeffreys (1961) introdujo el uso de distribuciones a priori convencionales para la selección de modelos y contraste de hipótesis. Por ejemplo, en la situación del ejemplo 1, él argumentó el uso de

$$\pi_1(\sigma_1) = \frac{1}{\sigma_1}, \quad \pi_2(\mu, \sigma_2) = \frac{1}{\sigma_2} \cdot \frac{1}{\pi\sigma_2(1 + \mu^2/\sigma_2^2)} = \frac{1}{\sigma_2} \cdot \frac{1}{\pi(\sigma_2^2 + \mu^2)}, \quad (3.7)$$

para lo que utilizó una distribución a priori no informativa estándar para el parámetro de escala, pero una densidad (propia) *Cauchy* $(0, \sigma_2)$ para la distribución a priori condicional de μ dada σ_2 . Con la elección de $\pi(\mu|\sigma_2)$ propia, la indeterminación, salvo una constante, del factor de Bayes queda eliminada, al menos en términos de μ . Además, en esta situación, Jeffreys identificó $\sigma_1^2 = \sigma_2^2 = \sigma^2$; despreocupándose de la indeterminación de $\pi(\sigma) = 1/\sigma$; ya que al aparecerle la misma constante multiplicativa en ambos modelos, entonces ésta se cancelaría.

El argumento de Jeffreys para (3.7) es bastante largo y puede o no ser convincente. Su solución, es, sin embargo, bastante razonable; es natural la elección del “tipo” de distribución a priori para μ dado σ_2 , y se sabe que las distribuciones a priori *Cauchy* son robustas en varios sentidos.

Aunque es fácil objetar el “imponer” tales elecciones sobre el análisis, es crucial recordar que no hay otra alternativa (excepto subjetividad). Cualquier método por defecto alternativo corresponde a la imposición de alguna distribución por defecto (propia) o, peor aún, termina por no corresponder a un análisis bayesiano verdadero. Este problema es bastante importante para merecer énfasis.

Principio 1.- *Los métodos que corresponden al uso de distribuciones a priori por defecto (propias) plausibles son preferibles a aquéllos que no corresponden a ningún análisis bayesiano verdadero.*

De aquí en adelante, se intentará mencionar qué métodos bayesianos por defecto son o no son consistentes con este principio. Dentro de las propuestas consistentes con este principio podemos mencionar a Zellner and Siow (1980), Poirier (1985), Stewart (1987), Mitchell and Beuachamp (1988), Albert (1990), George y McCulloch (1993), McCulloch y Rossi (1993), Raftery (1993), Madigan y Raftery (1994) y Verdinelli y Wasserman (1995). La limitación de estos enfoques es que tienden a construir distribuciones a priori por defecto para un escenario específico y aquí la meta es construir un método general y automático, consistente con el principio anterior.

3.2.2. Métodos asintóticos y criterio de información de Bayes

El método asintótico de Laplace (Haughton (1988); Kass and Raftery (1995)) produce, como una aproximación a B_{ji} ,

$$B_{ji}^L = \frac{f_j(x|\hat{\theta}_j) (\det \hat{I}_j)^{-1/2}}{f_i(x|\hat{\theta}_i) (\det \hat{I}_i)^{-1/2}} \cdot \frac{(2\pi)^{k_j/2} \pi_j(\hat{\theta}_j)}{(2\pi)^{k_i/2} \pi_i(\hat{\theta}_i)}, \quad (3.8)$$

donde \hat{I}_i y $\hat{\theta}_i$ representan la matriz de información observada de Fisher y el estimador de máxima verosimilitud bajo el modelo M_i , respectivamente.

Cuando el tamaño de la muestra tiende a infinito, el primer factor de B_{ji}^L , normalmente tiende a cero o infinito, mientras que el segundo factor tiende a una constante. El criterio de información de Bayes (CIB) propuesto por Schwarz (1978) surgió reemplazando este segundo factor por una constante apropiada.

Debido a que $\pi_j(\hat{\theta}_j)$ y $\pi_i(\hat{\theta}_i)$ en (3.8) no son constantes normalmente (es decir, independientes de los datos) en problemas de selección de modelos, entonces el uso de CIB no puede corresponder a un verdadero análisis bayesiano y por lo tanto viola el principio 1. Sin embargo, ésta violación es a menudo poco severa. Además, Kass y Wasserman (1992), argumentan que esta aproximación corresponde asintóticamente a un factor de Bayes verdadero en el estudio de modelos anidados.

La expresión asintótica (3.8) tiene mucha utilidad teórica. En la sección 3.3 se utilizará para ayudar a desarrollar distribuciones a priori propias convencionales y también resulta muy útil para evaluar propuestas en los problemas de comparación de modelos. Si el criterio propuesto (expresado en forma de factor de Bayes) no es asintóticamente equivalente a B_{ji}^L , entonces puede concluirse que no se comporta como un factor de Bayes verdadero.

Existen muchos problemas de selección de modelos para los cuales el resultado asintótico difiere de (3.8), (ver Haughton and Dudley (1992)). Es interesante ver que los factores de Bayes intrínsecos (FBI) a menudo tienen éxito, como se muestra en las siguientes subsecciones.

3.2.3. Distribuciones a priori no informativas convencionales

En la sección 3.1 observamos que el problema con las distribuciones a priori no informativas, π_1^N , es que el factor de Bayes está definido salvo constantes multiplicativas, c_i , y que el cociente de estas constantes serían a su vez un factor multiplicativo de los factores de Bayes. Se han hecho esfuerzos, para especificar de manera convencional las constantes c_i . Por ejemplo, Smith y Spiegelhalter (1980) y Spiegelhalter y Smith (1982) propusieron elegir c_i , de tal forma que $B_{ji}(x(l))$ sea igual a 1, cuando $x(l)$ se elige como la muestra de entrenamiento minimal (imaginaria), favoreciendo lo más posible al modelo más simple.

Desafortunadamente este método, bastante razonable, tiende a satisfacer el principio 1, pero sin lograrlo, ya que está predispuesto sistemáticamente en favor del modelo más complejo. Esta tendencia surge por la especificación de que $B_{ji}(x(l))$ sea igual a 1, aunque la muestra de entrenamiento (imaginaria) sea escogida para favorecer al máximo al modelo más simple.

3.2.4. Muestras de entrenamiento

La idea de la elección de muestras de entrenamiento, discutida en la sección 3.1, se ha utilizado de manera informal en muchas ocasiones. Algunos desarrollos más formales de esta idea, pueden encontrarse en Lemper (1971), Atkinson (1978), Geisser y Eddy (1979), Spiegelhalter y Smith (1982), San Martini y Spezzaferrri (1984) y Gelfand, Dey y Chang (1992), aunque no todos estos trabajos se utilizan para el cálculo de factores de Bayes ordinarios. Otras referencias, así como el comportamiento asintótico general de métodos de muestras de entrenamiento, han sido proporcionados por Gelfand y Dey (1994).

El método de Aitkin (1991) también puede considerarse como un método de muestra de entrenamiento; tomando la muestra entera x , como muestra de entrenamiento para obtener $\pi_i^N(\theta_i|x)$ y entonces usar esta distribución como una distribución a priori en (3.5) para calcular el factor de Bayes. Este doble uso de los datos no es consistente con la lógica bayesiana, y el método viola el criterio asintótico de forma severa.

O'Hagan (1995) propuso usar una parte fraccional de la verosimilitud $[f(x|\theta)]^\alpha$, en lugar de una muestra de entrenamiento. Esto tiende a producir una respuesta más estable que la producida por el uso de una muestra de entrenamiento particular, pero también falla el criterio asintótico, a menos que $\alpha \propto 1/n$ cuando el tamaño de la muestra crece. El factor de Bayes obtenido a través de este procedimiento es llamado factor de Bayes fraccional y su comportamiento ha sido bastante estudiado, particularmente, para elecciones tales como $\alpha = m_0/n$, donde m_0 es el tamaño de la muestra de entrenamiento minimal. Esta opción puede resultar atractiva en los factores de Bayes que corresponden al uso de distribuciones a priori por defecto, al menos para modelos lineales y ciertas elecciones de distribuciones a priori no informativas.

Independientemente del trabajo de Berger y Pericchi (1993), De Vos (1993) propuso un método de muestra de entrenamiento para modelos lineales que es similar al método propuesto por estos autores (ver Berger y Pericchi (1995) para discusión).

3.3. Factores de Bayes intrínsecos.

3.3.1. Modelos anidados

Se asume que M_1 está anidado en M_2 , en el sentido de que puede escribirse $\theta_2 = (\xi, \eta)$ y que f_1 y f_2 satisfacen

$$f_1(x|\theta_1) = f_2(x|\xi = \theta_1, \eta = \eta_0), \quad (3.9)$$

donde η_0 es un valor específico de η . En algunas ocasiones es conveniente identificar θ_1 con (θ_1, η_0) , para que de esta forma θ_1 y θ_2 pertenezcan al mismo espacio. También, algunas veces simplemente se escribe $\theta_2 = (\theta_1, \eta)$, aunque esto es peligroso (pero común) en la práctica. El peligro radica en que θ_1 en $f_1(x|\theta_1)$ y en $f_2(x|(\theta_1, \eta))$ pueden tener interpretaciones muy diferentes, y debido a la igualdad de símbolos, parecería lógico asignarles la misma distribución a priori (especialmente cuando se usan las distribuciones a priori por defecto). Esta es una cuestión importante, porque muchos de los esquemas para el desarrollo de distribuciones a priori convencionales, están basados en una formalización en la igualdad de tales parámetros. Los factores de Bayes intrínsecos naturalmente evitan el problema, como se verá en la subsección 3.4.2.

Supuesto 1.- Si M_1 está anidado en M_2 , entonces se asume que cuando el tamaño de la muestra tiende a infinito ($n \rightarrow \infty$), se tiene que

$$\widehat{\theta}_2 \xrightarrow{\text{bajo } M_1} \theta_2^* = (\theta_1, \eta_0), \quad (3.10)$$

donde $\widehat{\theta}_2$ es el estimador máximo verosímil de M_2 .

3.3.2. El factor de Bayes intrínseco

Para un conjunto de datos x , generalmente existen muchas muestras de entrenamiento como la que se definió en la sección 3.1.

Sea

$$\mathcal{X}_T = \{x(1), x(2), \dots, x(L)\}, \quad (3.11)$$

el conjunto de todas las muestras de entrenamiento minimales, $x(l)$. Claramente, el factor de Bayes $B_{21}(x(l))$, definido en (3.5), dependerá de la elección de la muestra de entrenamiento minimal. Para eliminar esta dependencia y aumentar la estabilidad de los factores de Bayes, una idea natural es promediar los $B_{21}(x(l))$ sobre todas las $x(l) \in \mathcal{X}_T$.

Debido a que existen diferentes tipos de medias, Berger y Pericchi (1996), definieron dos nuevos factores de Bayes, a los que denominaron factores de Bayes intrínsecos.

1.- Para la media aritmética, el denominado factor de Bayes intrínseco aritmético (FBIA) está definido por:

$$B_{21}^{IA} = \frac{1}{L} \sum_{l=1}^L B_{21}(x(l)) = B_{21}^N \cdot \frac{1}{L} \sum_{l=1}^L B_{12}^N(x(l)). \quad (3.12)$$

2.- Para la media geométrica, el denominado factor de Bayes intrínseco geométrico (FBIG) está definido por:

$$B_{21}^{IG} = \left(\prod_{l=1}^L B_{21}(x(l)) \right)^{1/L} = B_{21}^N \cdot \left(\prod_{l=1}^L B_{12}^N(x(l)) \right)^{1/L}, \quad (3.13)$$

donde el factor $B_{12}^N(x(l))$ está definido en (3.5).

Es necesario subrayar que aunque cada una de estas funciones han sido llamados por sus autores “factores de Bayes”, realmente no son factores de Bayes verdaderos. Sus comportamientos son estudiados por Berger y Pericchi (1996a, 1996b, 1997a, 1997b) en varios contextos.

Notas que deben tomarse en consideración.

Nota 1 $B_{21}^{IG} \leq B_{21}^{IA}$, puesto que la media geométrica es menor o igual a la media aritmética. Por tanto B_{21}^{IG} favorecerá al modelo anidado más simple más que B_{21}^{IA} .

Nota 2 B_{12}^{IA} se define como $1/B_{21}^{IA}$ y no como (3.12) con los índices invertidos. La asimetría surge porque M_1 está anidado en M_2 ; para B_{21}^{IG} no existe problema; invirtiendo los índices en (3.13) claramente resulta $1/B_{21}^{IG}$, ya que:

$$\begin{aligned} B_{12}^{IG} &= B_{12}^N \cdot \left(\prod_{l=1}^L B_{21}^N(x(l)) \right)^{1/L} = B_{12}^N \cdot \left(\frac{1}{\prod_{l=1}^L B_{12}^N(x(l))} \right)^{1/L} \\ &= B_{12}^N \cdot \left(\prod_{l=1}^L B_{12}^N(x(l)) \right)^{-1/L} = \frac{1}{B_{21}^{IG}}. \end{aligned}$$

Nota 3 Cuando el tamaño de muestra es pequeño, se tendrán problemas debido al uso de una parte de los datos como muestra de entrenamiento.

Ejemplo 5.-(Continuación). $X = (X_1, X_2, \dots, X_n)$ es una muestra i.i.d. de $M_1 : N(0, \sigma_1^2)$ o $M_2 : N(\mu, \sigma_2^2)$. Consideremos las distribuciones a priori no informativas $\pi_1^N(\sigma_1) = 1/\sigma_1$, y, $\pi_2^N(\mu, \sigma_2) = 1/\sigma_2^2$, haciendo los cálculos respectivos se obtiene que

$$B_{21}^N = \sqrt{\frac{2\pi}{n}} \cdot \left(1 + \frac{n\bar{x}}{s^2} \right)^{n/2}, \quad (3.14)$$

donde $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$.

Demostración

$$B_{21}^N = \frac{m_2^N}{m_1^N} = \frac{\int_{\Theta_2} f_2(x|\theta_2) \pi_2^N(\theta_2) d\theta_2}{\int_{\Theta_1} f_1(x|\theta_1) \pi_1^N(\theta_1) d\theta_1},$$

$$\begin{aligned}
m_2^N(x) &= \int_0^\infty \int_{-\infty}^\infty \frac{1}{(2\pi)^{n/2} \sigma_2^n} e^{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma_2^2}} \cdot \frac{1}{\sigma_2^2} d\mu d\sigma_2 \\
&= \int_0^\infty \int_{-\infty}^\infty \frac{1}{(2\pi)^{n/2} \sigma_2^n} e^{-\frac{1}{2\sigma_2^2} \left[n(\mu - \bar{x})^2 + \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \right]} \cdot \frac{1}{\sigma_2^2} d\mu d\sigma_2 \\
&= \int_0^\infty \frac{1}{(2\pi)^{n/2} \sigma_2^{n+1}} e^{-\frac{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)}{2\sigma_2^2}} \left[\int_{-\infty}^\infty e^{-\frac{n(\mu - \bar{x})^2}{2\sigma_2^2}} \cdot \frac{1}{\sigma_2^2} d\mu \right] d\sigma_2 \\
&= \int_0^\infty \frac{1}{(2\pi)^{n/2} \sigma_2^{n+1}} e^{-\frac{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)}{2\sigma_2^2}} \cdot \left(\frac{2\pi}{n} \right)^{1/2} d\sigma_2 \\
&= \frac{(2\pi)^{1/2}}{(2\pi)^{n/2} n^{1/2}} \int_0^\infty \sigma_2^{-(n+1)} e^{-\frac{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)}{2\sigma_2^2}} d\sigma_2 \\
&= \frac{(2\pi)^{1/2}}{(2\pi)^{n/2} n^{1/2}} \cdot \frac{1}{2} \left(\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{2} \right)^{-n/2} \Gamma\left(\frac{n}{2}\right),
\end{aligned}$$

por otro lado,

$$\begin{aligned}
m_1^N(x) &= \int_0^\infty \frac{1}{(2\pi)^{n/2} \sigma_1^n} e^{-\frac{1}{2\sigma_1^2} \sum_{i=1}^n x_i^2} \cdot \frac{1}{\sigma_1} d\sigma_1 \\
&= \frac{1}{(2\pi)^{n/2}} \int_0^\infty \sigma_1^{-(n+1)} e^{-\sigma_1^{-2} \frac{\sum_{i=1}^n x_i^2}{2}} d\sigma_1 \\
&= \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{2} \left(\frac{\sum_{i=1}^n x_i^2}{2} \right)^{-n/2} \Gamma\left(\frac{n}{2}\right).
\end{aligned}$$

Así, tenemos que

$$\begin{aligned}
B_{21}^N &= \frac{\frac{(2\pi)^{1/2}}{(2\pi)^{n/2} n^{1/2}} \cdot \frac{1}{2} \left(\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{2} \right)^{-n/2} \Gamma\left(\frac{n}{2}\right)}{\frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{2} \left(\frac{\sum_{i=1}^n x_i^2}{2} \right)^{-n/2} \Gamma\left(\frac{n}{2}\right)} \\
&= \sqrt{\frac{2\pi}{n}} \cdot \left(\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right)^{n/2},
\end{aligned}$$

o lo que es lo mismo,

$$B_{21}^N = \sqrt{\frac{2\pi}{n}} \cdot \left(1 + \frac{n\bar{x}}{s^2} \right)^{n/2},$$

donde $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$.

Nótese que \mathcal{X}_T consiste en todos los pares de observaciones diferentes $L = \binom{n}{2}$, o lo que es lo mismo $L = n(n-1)/2$, y por los resultados (3.5), (3.6), y por las definiciones (3.12) y (3.13), se sigue que

$$\begin{aligned}
B_{21}^{IA} &= B_{21}^N \cdot \frac{1}{L} \sum_{l=1}^n \frac{(x_1(l) - x_2(l))^2}{2\sqrt{\pi} [x_1^2(l) + x_2^2(l)]} \\
&= B_{21}^N \cdot \frac{1}{n(n-1)} \sum_{i < j} \frac{(x_i - x_j)^2}{\sqrt{\pi} [x_i^2 + x_j^2]},
\end{aligned} \tag{3.15}$$

y

$$B_{21}^{IG} = B_{21}^N \cdot \left(\prod_{l=1}^L \frac{(x_1(l) - x_2(l))^2}{2\sqrt{\pi} [x_1^2(l) + x_2^2(l)]} \right). \tag{3.16}$$

Nótese que B_{21}^{IA} y B_{21}^{IG} están definidos para cualesquiera modelos anidados, inclusive para aquellos que no son estándares, como el siguiente.

Ejemplo 6 *Supóngase que $X = (X_1, X_2, \dots, X_n)$ es una muestra i.i.d. de $M_1 : X_i \sim \mathcal{N}(\theta_1, 1)$ con $\theta_1 < 0$ ó $M_2 : X_i \sim \mathcal{N}(\theta_2, 1)$ con $\theta_2 \in \mathcal{R}^1$. Es importante recordar que θ_1 y θ_2 podrían ser, a priori, cantidades distintas, incluso cuando $\theta_2 < 0$. Usar*

el mismo símbolo, θ , tanto en M_1 como en M_2 , puede causar que uno asuma que, por ejemplo, $\pi_1(\theta)$ (bajo M_1) sea igual a $\pi_2(\theta|\theta < 0)$ (bajo M_2). Tales supuestos no están garantizados.

En este caso las distribuciones a priori no informativas son

$$\pi_1^N(\theta_1) = 1_{(-\infty, 0)}(\theta_1), \quad \text{y} \quad \pi_2^N(\theta_2) = 1,$$

por lo que

$$m_2^N(x_i) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta_2)^2} d\theta_2 = 1,$$

y por otra parte haciendo $u = x_i - \theta_1$; $du = -d\theta_1$;

$$\begin{aligned} m_1^N(x_i) &= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta_1)^2} d\theta_1 = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i} e^{-\frac{1}{2}u^2} du \\ &= \frac{1}{\sqrt{2\pi}} \int_{x_i}^{\infty} e^{-\frac{1}{2}u^2} du = 1 - \Phi(x_i) = \Phi(-x_i), \end{aligned}$$

donde $\Phi(\cdot)$ es la función de distribución de la normal estándar.

Puesto que $m_1^N(x_i)$, y $m_2^N(x_i)$ son ambas finitas entonces las muestras de entrenamiento minimales serán de una sola observación.

Por otro lado, se tiene que

$$\begin{aligned} B_{21}^N(x) &= \frac{m_2^N(x)}{m_1^N(x)} = \frac{\int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta_2)^2} d\theta_2}{\int_{-\infty}^0 \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta_1)^2} d\theta_1} \\ &= \frac{\int_{-\infty}^0 \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta_2)^2} d\theta_2 + \int_0^{\infty} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta_2)^2} d\theta_2}{\int_{-\infty}^0 \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta_1)^2} d\theta_1}, \end{aligned}$$

$$\text{renombrando } k = \frac{\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)\right\}}{(\sqrt{2\pi})^{n-1} \sqrt{n}},$$

$$\begin{aligned}
B_{21}^N(x) &= \frac{k \left[\int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{n}\theta_2 - \sqrt{n\bar{x}})^2} d\theta_2 + \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{n}\theta_2 - \sqrt{n\bar{x}})^2} d\theta_2 \right]}{k \left[\int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{n}\theta_1 - \sqrt{n\bar{x}})^2} d\theta_1 \right]} \\
&= \frac{\int_{-\infty}^{-\sqrt{n\bar{x}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt + \int_{-\sqrt{n\bar{x}}}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt}{\int_{-\infty}^{-\sqrt{n\bar{x}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt} = 1 + \frac{\int_{-\sqrt{n\bar{x}}}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt}{\int_{-\infty}^{-\sqrt{n\bar{x}}} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt} \\
&= 1 + \frac{1 - \Phi(-\sqrt{n\bar{x}})}{\Phi(-\sqrt{n\bar{x}})} = \frac{1}{\Phi(-\sqrt{n\bar{x}})},
\end{aligned}$$

por lo tanto tenemos que

$$B_{21}^N(x) = \frac{1}{\Phi(-\sqrt{n\bar{x}})}. \quad (3.17)$$

Así, (3.12) y (3.13) resultan

$$B_{21}^{IA} = \frac{1}{\Phi(-\sqrt{n\bar{x}})} \cdot \frac{1}{n} \sum_{i=1}^n \Phi(-x_i), \quad (3.18)$$

y

$$B_{21}^{IG} = \frac{1}{\Phi(-\sqrt{n\bar{x}})} \cdot \left[\prod_{i=1}^n \Phi(-x_i) \right]^{1/n}. \quad (3.19)$$

Esto es un ejemplo no estándar que es difícil de manejar por métodos ordinarios, en parte esto se encuentra indicado por el hecho de que la expresión asintótica B_{ji}^L (y por lo tanto el CIB), aquí no son válidos. Algunas correcciones asintóticas han sido proporcionadas por Haughton y Dudley (1992), quienes han estudiado problemas muy generales de este tipo. Para este caso, la expresión asintótica análoga a (3.8) es

$$B_{21} \cong \frac{\pi_2(\bar{x})}{\Phi(-\sqrt{n\bar{x}}) \pi_1(\min\{\bar{x}, 0\})}, \quad (3.20)$$

que es válida, si π_2 es continua y π_1 es continua y tiene límite finito, $\pi_1(0)$, en cero.

3.3.3. El factor de Bayes intrínseco esperado

Para muestras de tamaño pequeño, el promedio de las muestras de entrenamiento en (3.12) y (3.13) puede tener una gran varianza, y esto no es deseable, puesto que indica una inestabilidad de los factores de Bayes intrínsecos. También, el cálculo puede ser un problema cuando L es grande. Berger y Pericchi (1996), proponen una atractiva solución a este problema, reemplazando los promedios en (3.12) y (3.13) por sus respectivas esperanzas, evaluadas en los estimadores máximos verosímiles. Formalmente, Berger y

Pericchi (1996) definen el factor de Bayes intrínseco aritmético esperado (FBIAE) y el factor de Bayes intrínseco geométrico esperado (FBIGE) por

$$B_{21}^{IAE} = B_{21}^N \cdot \frac{1}{L} \sum_{l=1}^L \left[E_{\hat{\theta}_2}^{M_2} B_{12}^N (X(l)) \right], \quad (3.21)$$

y

$$B_{21}^{IGE} = B_{21}^N \cdot \exp \left\{ \frac{1}{L} \sum_{l=1}^L E_{\hat{\theta}_2}^{M_2} [\log B_{12}^N (X(l))] \right\}, \quad (3.22)$$

respectivamente, donde las esperanzas son bajo el modelo M_2 , con θ_2 igual al estimador máximo verosímil $\hat{\theta}_2$. Si las $X(l)$ son intercambiables, como es común, es decir,

$$f(x_1, x_2, \dots, x_n) = f(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}),$$

\forall permutación σ definida sobre el conjunto $\{1, 2, \dots, n\}$, entonces los promedios sobre L sobran.

Es obvio el hecho de que B_{21}^{IAE} y B_{21}^{IGE} están justificadas como aproximaciones de (3.12) y (3.13) para L grande y bajo M_2 . Pero éstas también son aproximaciones válidas bajo el modelo M_1 si el supuesto 1 de la subsección 3.3.1. se satisface, porque entonces bajo el modelo M_1 , $\hat{\theta}_2 \cong (\theta_1, \eta_0)$ lo cual, junto con (3.9), demuestra que las esperanzas en (3.21) y (3.22) son equivalentes a B_{21}^{IA} y B_{21}^{IG} bajo M_1 . Esta propiedad es única para modelos anidados y, desafortunadamente, impide derivar aproximaciones análogas de (3.21) y (3.22) para problemas con modelos no anidados.

Ejemplo 5.- (Continuación) Puesto que las $X(l)$ son intercambiables, por (3.6) y (3.22) se tiene que

$$B_{21}^{IAE} = B_{21}^N \cdot E_{\hat{\theta}_2}^{M_2} \left[\frac{(X_i - X_j)^2}{2\sqrt{\pi} [X_i^2 + X_j^2]} \right] = B_{21}^N \cdot \left(\frac{1 - e^{-n\bar{x}/s^2}}{2\sqrt{\pi} [n\bar{x}/s^2]} \right), \quad (3.23)$$

(ver Berger y Pericchi (1993), para el cálculo de esta esperanza). Además,

$$B_{21}^{IGE} = B_{21}^N \cdot \exp \left\{ E_{\hat{\theta}_2}^{M_2} \left[\log \left(\frac{(X_i - X_j)^2}{2\sqrt{\pi} [X_i^2 + X_j^2]} \right) \right] \right\},$$

donde $\hat{\theta} = (\bar{x}, s^2/n)$. Esta puede ser evaluada sólomente como una serie infinita (ver Berger y Pericchi (1994)); pero el cálculo numérico es directo.

Ejemplo 6.- (Continuación). Usando B_{21}^{IA} y B_{21}^{IG} y considerando la intercambiabilidad de las muestras de entrenamiento $X(l)$

$$B_{21}^{IAE} = \frac{1}{\Phi(-\sqrt{n\bar{x}})} \cdot E_{\hat{\theta}_2}^{M_2} [\Phi(-X_i)] = \frac{1}{\Phi(-\sqrt{n\bar{x}})} \cdot \Phi\left(-\frac{\bar{x}}{\sqrt{2}}\right); \quad (3.24)$$

y

$$B_{21}^{IGE} = \frac{1}{\Phi(-\sqrt{n\bar{x}})} \cdot \exp E_{\hat{\theta}_2}^{M_2} [\log \Phi(-X_i)],$$

respectivamente, donde $X_i \sim N(\theta, 1)$ bajo M_2 y $\hat{\theta}_2 = \bar{x}$. De nuevo, B_{21}^{IGE} no puede encontrarse de forma explícita.

Como ocurría con B_{12}^{IA} , Berger y Pericchi (1996) definen $B_{12}^{IAE} = 1/B_{21}^{IAE}$. En este caso no existe otra opción, porque en muchos problemas (tales como en los ejemplo 5 y 6)

$$E_{\hat{\theta}_2}^{M_2} B_{12}^N(X(l)) = \infty.$$

Esto también explica la definición de $B_{12}^{IA} = 1/B_{21}^{IA}$; aunque B_{12}^{IA} podría definirse como (3.12) con los índices invertidos, el promedio de $B_{12}^N(x(l))$ normalmente divergería cuando $L \rightarrow \infty$, resultando un factor de Bayes que violaría el principio 1 y por lo tanto no válido.

3.3.4. Comparaciones

A continuación hacemos una pausa para comparar los factores de Bayes intrínsecos con otros métodos, así como obtener una visión en estos casos simples para ver si nuestros objetivos son alcanzados. Las comparaciones que se hacen son con la expresión asintótica (3.8), la aproximación de Schwarz, y con los métodos de Jeffreys (1961) y Smith y Spiegelhalter (1980). Así como también se compararán estos factores entre sí.

Ejemplo 5. – (Continuación). Es fácil demostrar que los estimadores máximo verosímiles para ambos modelos son

$$\hat{\mu} = \bar{x}, \hat{\sigma}_1 = \left(\frac{\sum_{i=1}^n x_i^2}{n} \right)^{1/2} = \left(\bar{x} + \frac{s^2}{n} \right)^{1/2}, \hat{\sigma}_2 = \sqrt{\frac{s^2}{n}}.$$

Demostración.

Sabemos que $X_i \sim \mathcal{N}(0, \sigma_1^2)$ bajo M_1 por lo que

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}x_i^2} \Rightarrow f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \sigma_1^n} e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma_1^2}},$$

por lo que

$$\frac{\partial \log [f(x_1, x_2, \dots, x_n)]}{\partial \sigma_1} = -\frac{n}{\sigma_1} + \frac{\sum_{i=1}^n x_i^2}{\sigma_1^3} = 0 \iff \hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}},$$

y esto es igual a

$$\hat{\sigma}_1 = (\bar{x}^2 + s^2/n)^{1/2}.$$

Por otro lado, $X_i \sim \mathcal{N}(\mu, \sigma_2^2)$ bajo M_2 por lo que

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}(x_i-\mu)^2} \Rightarrow f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \sigma_2^n} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma_2^2}},$$

$$\frac{\partial \log [f(x_1, x_2, \dots, x_n)]}{\partial \mu} = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma_2^2} = 0 \iff \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x},$$

y

$$\frac{\partial \log [f(x_1, x_2, \dots, x_n)]}{\partial \sigma_2} = -\frac{n}{\sigma_2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma_2^3} = 0 \iff \hat{\sigma}_2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n},$$

que es lo mismo que

$$\hat{\sigma}_2 = \sqrt{\frac{s^2}{n}},$$

donde $s^2 = \sum_{i=1}^n (x_i - \mu)^2$.

Tomando en cuenta los resultados anteriores se tienen las siguientes aproximaciones:

Aproximación asintótica: (3.8) está dado por

$$B_{21}^L = B_{21}^N \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\mu}, \hat{\sigma}_2)}{\pi_1(\hat{\sigma}_1)}. \quad (3.25)$$

Recordemos que

$$B_{ji}^L = \frac{f_j(x|\hat{\theta}_j) (\det \hat{I}_j)^{-1/2}}{f_i(x|\hat{\theta}_i) (\det \hat{I}_i)^{-1/2}} \cdot \frac{(2\pi)^{k_j/2} \pi_j(\hat{\theta}_j)}{(2\pi)^{k_i/2} \pi_i(\hat{\theta}_i)},$$

donde \hat{I}_i y $\hat{\theta}_i$ representan la matriz de información de Fisher observada y el estimador de máxima verosimilitud bajo el modelo M_i , respectivamente.

Entonces, puesto que bajo M_1 , $X_i \sim \mathcal{N}(0, \sigma_1^2)$, se tiene que $k_1 = 1$ y la matriz de información observada de Fisher es igual a

$$\begin{aligned}
\widehat{I}_1(x) &= - \left[\frac{\partial^2 \log f(x|\sigma_1)}{\partial \sigma_1^2} \right]_{\sigma_1 = \widehat{\sigma}_1} = - \left[\frac{n}{\sigma_1^2} - 3 \frac{\sum_{i=1}^n x_i^2}{\sigma_1^4} \right]_{\sigma_1 = \widehat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}} \\
&= - \left[\frac{n}{\frac{\sum_{i=1}^n x_i^2}{n}} - \frac{3 \sum_{i=1}^n x_i^2}{\left(\frac{\sum_{i=1}^n x_i^2}{n} \right)^2} \right] = - \left[\frac{n^2}{\sum_{i=1}^n x_i^2} - \frac{3n^2}{\sum_{i=1}^n x_i^2} \right] \\
&= - \left[-\frac{2n^2}{\sum_{i=1}^n x_i^2} \right] = \frac{2n}{\widehat{\sigma}_1^2}.
\end{aligned}$$

Esto implica que

$$(\det \widehat{I}_1)^{-\frac{1}{2}} = \frac{\widehat{\sigma}_1}{\sqrt{2n}}.$$

Por otro lado, bajo $M_2, X_i \sim \mathcal{N}(\mu, \sigma_2^2)$ entonces se tiene que $k_2 = 2$, además,

$$\log f(x_1, x_2, \dots, x_n | \mu, \sigma) = -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma_2^2} - \frac{n}{2} \log(2\pi) - n \log \sigma_2,$$

y la matriz de información observada de Fisher es

$$\begin{aligned}
\widehat{I}_2(x) &= - \left[\begin{array}{cc} \frac{\partial^2 \log f(x|\mu, \sigma)}{\partial \mu^2} & \frac{\partial^2 \log f(x|\mu, \sigma)}{\partial \mu \partial \sigma_2} \\ \frac{\partial^2 \log f(x|\mu, \sigma)}{\partial \sigma_2 \partial \mu} & \frac{\partial^2 \log f(x|\mu, \sigma)}{\partial \sigma_2^2} \end{array} \right]_{(\mu, \sigma) = (\widehat{\mu}, \widehat{\sigma})} \\
&= - \left[\begin{array}{cc} -\frac{n}{\sigma_2^2} & -\frac{2 \sum_{i=1}^n (x_i - \mu)}{\sigma_2^3} \\ -\frac{2 \sum_{i=1}^n (x_i - \mu)}{\sigma_2^3} & -\frac{3 \sum_{i=1}^n (x_i - \mu)^2}{\sigma_2^4} + \frac{n}{\sigma_2^2} \end{array} \right]_{(\mu, \sigma) = \left(\bar{x}, \sqrt{\frac{s^2}{n}} \right)} \\
&= \left[\begin{array}{cc} \frac{n}{\widehat{\sigma}_2^2} & \frac{2 \sum_{i=1}^n (x_i - \bar{x})}{\widehat{\sigma}_2^3} \\ \frac{2 \sum_{i=1}^n (x_i - \bar{x})}{\widehat{\sigma}_2^3} & \frac{2n}{\widehat{\sigma}_2^2} \end{array} \right],
\end{aligned}$$

de aquí tenemos que

$$\left(\det \hat{I}_2\right)^{-\frac{1}{2}} = \frac{\hat{\sigma}_2^2}{\sqrt{2n}},$$

por lo tanto,

$$B_{21}^L = \frac{\left(1/(2\pi)^{n/2} \hat{\sigma}_2^n\right) \cdot \exp\left[-\sum_{i=1}^n (x_i - \hat{\mu})^2 / 2\hat{\sigma}_2^2\right] \cdot \frac{\hat{\sigma}_2^2}{\sqrt{2n}} \cdot (2\pi) \cdot \pi_2(\hat{\theta}_2)}{\left(1/(2\pi)^{n/2} \hat{\sigma}_1^n\right) \cdot \exp\left[-\sum_{i=1}^n x_i^2 / 2\hat{\sigma}_1^2\right] \cdot \frac{\hat{\sigma}_1}{\sqrt{2n}} \cdot (2\pi)^{1/2} \cdot \pi_1(\hat{\theta}_1)}.$$

Sustituyendo los estimadores máximo verosímiles correspondientes, en esta expresión se tiene que, tomando en cuenta que $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ y $\sum_{i=1}^n x_i^2 = s^2 + n\bar{x}^2$,

$$\begin{aligned} B_{21}^L &= \frac{\hat{\sigma}_1^n \cdot \hat{\sigma}_2^2 \exp(-n/2) \cdot (2\pi)^{1/2} \cdot \pi_2(\hat{\theta}_2)}{\hat{\sigma}_1 \cdot \hat{\sigma}_2^n \exp(-n/2) \sqrt{n} \cdot \pi_1(\hat{\theta}_1)} \\ &= \sqrt{\frac{2\pi}{n}} \cdot \frac{\hat{\sigma}_1^n}{\hat{\sigma}_2^n} \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\theta}_2)}{\pi_1(\hat{\theta}_1)} \\ &= \sqrt{\frac{2\pi}{n}} \cdot \frac{(\bar{x}^2 + s^2/n)^{n/2}}{(s^2/n)^{n/2}} \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\theta}_2)}{\pi_1(\hat{\theta}_1)} \\ &= \sqrt{\frac{2\pi}{n}} \cdot \left(1 + \frac{n\bar{x}^2}{s^2}\right)^{n/2} \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\theta}_2)}{\pi_1(\hat{\theta}_1)} \\ &= B_{21}^N \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\theta}_2)}{\pi_1(\hat{\theta}_1)}, \end{aligned}$$

que demuestra la afirmación anterior.

Aproximación de Schwarz.- Está dada por

$$B_{21}^S = \frac{B_{21}^N}{\sqrt{2\pi}}.$$

Aproximación de Jeffreys. Para las distribuciones a priori en (3.7), Jeffreys (1961) aproximó B_{21} por

$$B_{21}^J = B_{21}^N \cdot \hat{\sigma}_2 \cdot \frac{1}{\hat{\sigma}_2 \pi [1 + \hat{\mu}^2 / \hat{\sigma}_2^2]}. \quad (3.26)$$

Aproximación de Smith y Spiegelhalter (1980).- El factor de Bayes de estos autores es de la forma

$$B_{21}^{SS} = \frac{B_{21}^N}{\sqrt{\pi}}.$$

Es útil volver a escribir $\pi_2(\mu, \sigma_2)$ como

$$\pi_2(\mu, \sigma_2) = \pi_2(\mu | \sigma_2) \cdot \pi_2(\sigma_2).$$

Si ahora elegimos las distribuciones a priori no informativas $\pi_1(\sigma_1) = 1/\sigma_1$ y $\pi_2(\sigma_2) = 1/\sigma_2$ entonces B_{21}^L se convierte en

$$\begin{aligned} B_{21}^L &= B_{21}^N \cdot \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1} \cdot \frac{\pi_2(\hat{\mu} | \hat{\sigma}_2) \frac{1}{\hat{\sigma}_2}}{\frac{1}{\hat{\sigma}_1}} \\ &= B_{21}^N \cdot \hat{\sigma}_2 \cdot \pi_2(\hat{\mu} | \hat{\sigma}_2). \end{aligned} \quad (3.27)$$

Nótese que B_{21}^J es de esta forma, con $\pi_2(\mu | \sigma_2)$ una distribución *Cauchy*(0, σ_2); por lo que se tiene que B_{21}^J es consistente con (3.7).

Asimismo, (3.23) y (3.24) pueden reescribirse como

$$B_{21}^{IAE} = B_{21}^N \cdot \left(\frac{1 - \exp\{-n\bar{x}^2/s^2\}}{2\sqrt{\pi} [\hat{\mu}^2/\hat{\sigma}_2^2]} \right) = B_{21}^N \cdot \hat{\sigma}_2 \cdot \left(\frac{1 - \exp\{-\hat{\mu}^2/\hat{\sigma}_2^2\}}{2\sqrt{\pi} [\hat{\mu}^2/\hat{\sigma}_2^2]} \right), \quad (3.28)$$

y,

$$B_{21}^{IGE} = B_{21}^N \cdot \hat{\sigma}_2 \cdot \left(\frac{1}{\hat{\sigma}_2} E_{(\hat{\mu}, \hat{\sigma}_2)} \left[\log \left(\frac{\{X_i - X_j\}^2}{2\sqrt{\pi} \{X_i^2 + X_j^2\}} \right) \right] \right).$$

Recordemos que una de las metas de Berger y Pericchi (1996) era desarrollar un método automático que pudiera “reproducir” un factor de Bayes verdadero. Pues bien, resulta que B_{21}^{IAE} se comporta increíblemente bien ya que se demuestra que

$$\pi_2^I(\mu | \sigma_2) = \frac{1 - \exp\{-\mu^2/\sigma_2^2\}}{2\sqrt{\pi} [\mu^2/\sigma_2^2]}, \quad (3.29)$$

es una distribución a priori propia (integra uno sobre μ) y, además, es prácticamente equivalente a la elección de Jeffreys de $\pi_2(\mu | \sigma_2)$ como *Cauchy*(0, σ_2); de hecho, las densidades de las dos distribuciones a priori nunca difieren en más del 15%.

Esta extraordinaria propiedad de B_{21}^{IAE} no es compartida, desafortunadamente, por B_{21}^{IGE} . Éste no corresponde a un factor de Bayes para una distribución a priori propia.

Comparando B_{21}^J o B_{21}^{IAE} con B_{21}^{SS} , vemos que el último es más grande por un factor de $\sqrt{\pi} (1 + \hat{\mu}^2/\hat{\sigma}_2^2)$, es decir,

$$B_{21}^J = B_{21}^{SS} \cdot \frac{1}{\sqrt{\pi} (1 + \hat{\mu}^2/\hat{\sigma}_2^2)},$$

que siempre es más grande que uno (y puede ser arbitrariamente grande). Este fenómeno es generalmente verdadero y propociona el apoyo para la predisposición en la subsección 3.2.3. de que B_{21}^{SS} está sesgado hacia el modelo más complejo.

Similarmente, B_{21}^S es más grande por un factor de $\sqrt{\pi/2} (1 + \hat{\mu}^2/\hat{\sigma}_2^2)$, es decir,

$$B_{21}^J = B_{21}^S \cdot \frac{1}{\sqrt{\pi/2} (1 + \hat{\mu}^2/\hat{\sigma}_2^2)},$$

y está, en sí mismo, sesgado hacia el modelo más complejo.

Si se hubiera escogido $\pi_2(\mu|\sigma_2)$ como una $\mathcal{N}(0, \sigma_2^2)$, entonces

$$B_{21}^L = B_{21}^S \cdot \exp\{-\bar{x}^2/2\hat{\sigma}_2^2\}.$$

Si M_1 fuera el verdadero modelo y n fuera grande, entonces $\bar{x} \cong 0$ y $B_{21}^L \cong B_{21}^S$. Esta es la base para el argumento de Kass y Wasserman (1994) que dicen que B_{21}^S es aproximadamente un factor de Bayes cuando el modelo más simple es el verdadero, en una situación de modelos anidados.

La historia para B_{21}^{IA} y B_{21}^{IG} es similar. Notablemente, el promedio en B_{21}^{IA} otra vez corresponde a una distribución a priori propia. El término correspondiente de (3.13) es cualitativamente similar, pero otra vez, no corresponde a una distribución a priori propia.

3.4. Distribuciones a priori intrínsecas

3.4.1. Definición y motivación

Desde el punto de vista teórico, la principal razón referente a los factores de Bayes intrínsecos y factores de Bayes fraccionales es determinar si efectivamente se corresponden con un factor de Bayes verdadero para distribuciones a priori razonables. De ser así, se tiene una justificación bayesiana para ambos métodos y la consistencia de estos factores de Bayes se cumpliría automáticamente. Berger y Pericchi (1996) resolvieron asintóticamente esta cuestión obteniendo las denominadas **distribuciones a priori intrínsecas**.

En el ejemplo 5 (continuación), en la subsección 3.3.5, vimos que B_{21}^{IA} y B_{21}^{IAE} eran aproximadamente iguales a los factores de Bayes verdaderos para la distribución a priori (condicional) propia (3.29). Tal distribución a priori si existe es llamada **distribución a**

priori intrínseca. Esto eliminaría la necesidad de calcular muestras de entrenamiento y eliminaría lo concerniente acerca de la estabilidad de los factores de Bayes intrínsecos.

Alternativamente, los factores de Bayes intrínsecos pueden verse como un método para aplicar a “muestras de entrenamiento imaginarias”, para determinar distribuciones a priori convencionales (razonables) para usar en problemas de selección de modelos y contraste de hipótesis. Además, esto podría verse como el complemento a la teoría de las distribuciones a priori de referencia (Bernardo (1979); Berger y Bernardo (1992)), en la cual se desarrollan distribuciones a priori convencionales para problemas de estimación.

Como sucede con las distribuciones a priori de referencia, las definiciones de las distribuciones a priori intrínseca, de Berger y Pericchi (1996), se centran alrededor del uso de una muestra de entrenamiento imaginaria (asintótica). Con frecuencia, se pueden utilizar argumentos asintóticos para verificar la siguiente aproximación, que Berger y Pericchi (1996) establecieron como condición.

Condición A. Cuando el tamaño de muestra tiende a infinito, para distribuciones a priori en una clase apropiada, Υ , (3.2) puede aproximarse por

$$B_{ji} = B_{ji}^N \cdot \frac{\pi_j(\hat{\theta}_j) \pi_i^N(\hat{\theta}_i)}{\pi_j^N(\hat{\theta}_j) \pi_i(\hat{\theta}_i)} \cdot (1 + o(1)), \quad (3.30)$$

donde $\hat{\theta}_j$ y $\hat{\theta}_i$ son los estimadores máximo verosímiles bajo M_j y M_i , respectivamente, y $o(1) \rightarrow 0$ en probabilidad bajo M_j y M_i .

Esta condición puede verse fácilmente para comprender la situación asintótica estándar de B_{ji}^L , pero también para comprender las situaciones no estándares tales como la del ejemplo 6; ver (3.17) y (3.20).

Para definir las distribuciones a priori intrínsecas, Berger y Pericchi (1996) comenzaron igualando (3.30), con (3.12) o (3.13), produciendo

$$\frac{\pi_j(\hat{\theta}_j) \pi_i^N(\hat{\theta}_i)}{\pi_j^N(\hat{\theta}_j) \pi_i(\hat{\theta}_i)} \cdot (1 + o(1)) = \tilde{B}_{ij}^N, \quad (3.31)$$

definiendo a \tilde{B}_{ji}^N como la media aritmética o la media geométrica de $B_{ji}^N(x(l))$. Posteriormente hicieron algunos supuestos sobre el comportamiento límite de las cantidades en (3.31). La siguiente condición típicamente se satisface.

Condición B. Cuando el tamaño de muestra tiende a infinito, se tiene lo siguiente:

- i) Bajo M_j , $\hat{\theta}_j \rightarrow \theta_j$, $\hat{\theta}_i \rightarrow \psi_i(\theta_j)$, y $\tilde{B}_{ij}^N \rightarrow B_j^*(\theta_j)$.
- ii) Bajo M_i , $\hat{\theta}_i \rightarrow \theta_i$, $\hat{\theta}_j \rightarrow \psi_j(\theta_i)$, y $\tilde{B}_{ij}^N \rightarrow B_i^*(\theta_i)$.

Donde, $\psi_i(\theta_j)$ denota el límite del estimador máximo verosímil $\hat{\theta}_i$ bajo el modelo M_j en el punto θ_j .

Normalmente, para $k = i$ o $k = j$,

$$B_k^*(\theta_k) = \lim_{L \rightarrow \infty} E_{\theta_k}^{M_k} \left[\frac{1}{L} \sum_{l=1}^L B_{ij}^N(X(l)) \right] \text{ caso aritmético,}$$

$$B_k^*(\theta_k) = \lim_{L \rightarrow \infty} \exp \left\{ E_{\theta_k}^{M_k} \left[\frac{1}{L} \sum_{l=1}^L \log B_{ij}^N(X(l)) \right] \right\} \text{ caso geométrico,} \quad (3.32)$$

donde si las $X(l)$ son intercambiables, entonces los límites y los promedios sobre L pueden eliminarse.

Usando la condición B y tomando límites en (3.31), primero bajo M_j y después bajo M_i ; Berger y Pericchi (1996) obtuvieron dos ecuaciones funcionales, que son las que definen las distribuciones a priori intrínsecas, (π_j^I, π_i^I) .

Las dos ecuaciones funcionales son la siguientes:

$$\frac{\pi_j^I(\theta_j) \pi_i^N(\psi_i(\theta_j))}{\pi_j^N(\theta_j) \pi_i^I(\psi_i(\theta_j))} = B_j^*(\theta_j), \quad (3.33)$$

y

$$\frac{\pi_j^I(\psi_j(\theta_i)) \pi_i^N(\theta_i)}{\pi_j^N(\psi_j(\theta_i)) \pi_i^I(\theta_j)} = B_i^*(\theta_i). \quad (3.34)$$

Otra vez, la motivación, se centra en que las distribuciones a priori que satisfagan las ecuaciones (3.33) y (3.34), tendrán respuestas que serán asintóticamente equivalentes a aquéllas obtenidas usando los factores de Bayes intrínsecos.

También puede notarse que las soluciones no son necesariamente únicas ni necesariamente propias. Por ejemplo, esto se tiene en las situaciones en las cuales

$$B_j^*(\theta_j) = B_i^*(\theta_i) = 1,$$

ya que en estos casos se ve que

$$\pi_k^I(\theta_k) = \pi_k^N(\theta_k), \text{ para } k = i, j,$$

son soluciones de las ecuaciones anteriores, resultando las distribuciones a priori intrínsecas, de nuevo las distribuciones a priori no informativas.

3.4.2. Distribuciones a priori intrínsecas para modelos anidados.

En el escenario de los modelos anidados de la subsección 3.3.1 y bajo el supuesto 1, las soluciones a (3.33) y (3.34), están dadas por

$$\pi_1^I(\theta_1) = \pi_1^N(\theta_1), \quad \pi_2^I(\theta_2) = \pi_2^N(\theta_2) B_2^*(\theta_2). \quad (3.35)$$

Demostración.

Sustituyendo (3.35), en las ecuaciones (3.33) y (3.34) tenemos que

$$\frac{\pi_2^N(\theta_2) B_2^*(\theta_2) \pi_1^N(\psi_1(\theta_2))}{\pi_2^N(\theta_2) \pi_1^N(\psi_1(\theta_2))} = B_2^*(\theta_2),$$

y

$$\frac{\pi_2^N(\psi_2(\theta_1)) B_2^*(\psi_2(\theta_1)) \pi_1^N(\theta_1)}{\pi_2^N(\psi_2(\theta_1)) \pi_1^N(\theta_1)} = B_2^*(\psi_2(\theta_1)).$$

Existen también otras soluciones, quizás incluso soluciones que son distribuciones propias, pero las soluciones en (3.35) son las más simples. (Ver Dmochowski (1994) para caracterizaciones interesantes de distribuciones a priori intrínsecas para modelos anidados).

Ejemplo 5.- (Continuación). Para B_{21}^{IA} , se sigue de (3.23) y (3.32) que

$$B_2^*(\theta_2) = \sigma_2 \cdot \pi_2^I(\mu|\sigma_2),$$

donde $\pi_2^I(\mu|\sigma_2)$ está definido en (3.29).

Por lo tanto las distribuciones a priori intrínsecas son

$$\pi_1^I(\sigma_1) = \pi_1^N(\sigma_1) = 1/\sigma_1,$$

$$\pi_2^I(\mu, \sigma_2) = \pi_2^N(\sigma_2) B_2^*(\mu, \sigma_2) = \frac{1}{\sigma_2} \cdot \pi_2^I(\mu|\sigma_2). \quad (3.36)$$

Por tanto, B_{21}^{IA} se comporta (asintóticamente) como un factor de Bayes verdadero que usa las distribuciones a priori no informativas de referencia para σ_1 y σ_2 , y la distribución propia $\pi_2^I(\mu|\sigma_2)$, para la distribución a priori condicional de μ dada σ_2 . Además de la propiedad de $\pi_2^I(\mu|\sigma_2)$, es también notable que la distribución a priori intrínseca para σ_2 es la distribución a priori de referencia $1/\sigma_2$, y no la distribución a priori de Jeffreys (formal) $1/\sigma_2^2$ que se utilizó para calcular B_{21}^{IA} . Berger y Pericchi (1996) también observaron este último comportamiento en otros ejemplos; los factores de Bayes intrínsecos parecen intentar convertir la original π_i^N en distribuciones a priori de referencia para parámetros comunes o modelos similares.

Ejemplo 6.- (Continuación). Para B_{21}^{IA} , vemos de (3.18), (3.24), y (3.32) que $B_2^*(\theta_2) = \Phi(-\theta_2/\sqrt{2})$. Por lo tanto, las distribuciones a priori intrínsecas (3.35) son

$$\pi_1^I(\theta_1) = \pi_1^N(\theta_1) = 1,$$

$$\pi_2^I(\theta_2) = 1 \cdot \Phi(-\theta_2/\sqrt{2}). \quad (3.37)$$

Dos características de las distribuciones a priori intrínsecas son de particular interés. Primera, en $(-\infty, 0)$, las distribuciones a priori π_1^I, π_2^I no son proporcionales. Recordemos que los modelos se escribieron como $M_1 : \theta_1 < 0, M_2 : \theta_2 \in \mathfrak{R}$, a diferencia de $M_1 : \theta < 0, M_2 : \theta \in \mathfrak{R}$, para enfatizar que $\theta_i, i = 1, 2$, podrían tener diferentes interpretaciones bajo cada modelo y diferentes distribuciones a priori intrínsecas, incluso en su dominio común. Esta posibilidad parece haberse realizado. Nótese, sin embargo, que sobre $(-\infty, 0)$, π_1^I, π_2^I difieren substancialmente cerca del cero.

La segunda característica interesante de las distribuciones a priori intrínsecas es que

$$\int_0^\infty \pi_2^I(\theta_2) d\theta_2 = \int_0^\infty \Phi(-\theta_2/\sqrt{2}) d\theta_2 = \frac{1}{\sqrt{\pi}},$$

por lo tanto $\pi_2^I(\theta_2 | \{\theta_2 > 0\})$ es propia.

El comportamiento de las distribuciones a priori intrínsecas, observado en los ejemplos anteriores es típico para modelos anidados. Parámetros “comunes” (o al menos, parámetros que pueden ser identificados en el sentido de la definición (3.9)) normalmente tienen distribuciones a priori intrínsecas que son distribuciones a priori estándares no informativas o ligeras variantes, mientras que parámetros que están sólo en los modelos más complejos (o que tienen dominios extendidos en los modelos más complejos) tienen distribuciones a priori intrínsecas propias (condicionales). (Ver Berger y Pericchi (1995) para una aplicación en modelos lineales normales anidados).

En general, dados $M_1 : \{f_1(x|\theta_1), \pi_1(\theta_1)\}$ y $M_2 : \{f_2(x|\theta_2), \pi_2^N(\theta_2)\}$ dos modelos para los cuales se tiene que:

(i) $f_1(x|\theta_1)$ está anidado en $f_2(x|\theta_2)$,

(ii) $\pi_1(\theta_1)$ es una distribución a priori propia,

(iii) $\pi_2^N(\theta_2)$ es una distribución a priori impropia, y suponiendo que para una muestra de tamaño n la función de verosimilitud $f_2(x_1, x_2, \dots, x_n|\theta_2)$ es integrable con respecto a $\pi_2^N(\theta_2)$, se demuestra que asintóticamente la metodología intrínseca genera una única función de densidad $\pi_2^I(\theta_2)$ llamada distribución a priori intrínseca (Berger y Pericchi (1996)). Esta distribución es de la forma

$$\pi_2^I(\theta_2) = \pi_2^N(\theta_2) E_{x(l)|\theta_2}^{M_2} B_{12}^N(x(l)), \quad (3.38)$$

donde la esperanza se calcula con respecto a la densidad de la muestra de entrenamiento, $x(l)$, y considerada bajo el modelo M_2 . Además, se cumple que (3.38) es una distribución a priori propia.

Teorema 1 *Para el factor de Bayes intrínseco aritmético (FBIA), supóngase que (3.32) se cumple y que $\pi_1^N(\theta_1)$, es propia. Entonces $\pi_2^I(\theta_2)$, definida en (3.35), es también propia.*

Demostración.-

Puesto que se asume que existe el límite en (3.32), entonces es cierto que

$$\begin{aligned}
\int \pi_2^I(\theta_2) &= \int \pi_2^N(\theta_2) \left(\lim_{L \rightarrow \infty} E_{\theta_2}^{M_2} \left[\frac{1}{L} \sum_{l=1}^L B_{12}^N(X(l)) \right] \right) d\theta_2 & (3.39) \\
&= \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{l=1}^L \int \int \pi_2^N(\theta_2) f_2(x(l) | \theta_2) B_{12}^N(x(l)) dx(l) d\theta_2 \\
&= \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{l=1}^L \int m_2^N(x(l)) \cdot \frac{m_1^N(x(l))}{m_2^N(x(l))} dx(l) \\
&= \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{l=1}^L \int \int m_1^N(x(l)) dx(l) \\
&= \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{l=1}^L (1) \\
&= \lim_{L \rightarrow \infty} \frac{L}{L} = 1,
\end{aligned}$$

la penúltima línea se sigue del hecho de que si π_1^N es propia entonces m_1^N también lo es.

La gran generalidad de este teorema es atenuada por la condición de que $\pi_1^N(\theta)$ sea propia. Por otra parte, Berger y Pericchi esperan mostrar que la distribución impropia π_1^N pueda ser manejada por un argumento límite apropiado.

Finalmente, nótese que no existe un teorema análogo al teorema 1 para B_{21}^{IG} , ya que, como vimos en la subsección 3.3.5, B_{21}^{IG} (equivalente a B_{21}^{IGE} en términos de distribuciones a priori intrínsecas, por (3.32)) no corresponde a un factor de Bayes con distribuciones a priori propias.

Capítulo 4

Pruebas de igualdad de coeficientes de regresión

Tanto para el enfoque frecuentista como para el enfoque bayesiano, comprobar si los vectores de coeficientes de regresión son iguales, para dos modelos de regresión normales independientes, cuando no se conocen las varianzas de los errores, presenta varias dificultades.

En el enfoque frecuentista, no se aplica la teoría para el contraste de hipótesis normales debido a que las varianzas no están relacionadas, y en el enfoque bayesiano, las distribuciones a priori generalmente usadas son impropias y, por lo tanto, la solución basada en el factor de Bayes no puede usarse. Moreno, Torres y Casella (2002), proponen una solución interesante para este problema, en el cual no se considera ninguna información a priori subjetiva.

En este capítulo se hará un análisis exhaustivo de la propuesta de estos autores, en donde primero se genera una distribución a priori propia “objetiva” (distribución a priori intrínseca) para la cual el factor de Bayes y la probabilidad a posteriori del modelo están bien definidos, y en donde la probabilidad a posteriori de cada modelo se usa como una herramienta de selección.

Este procedimiento de contraste de hipótesis se compara con algunas aproximaciones frecuentistas propuestas en la literatura, tales como el estadístico modificado de Chow, propuesto por Conerly y Mansfiel (1988).

A continuación se define el estadístico de Chow.

4.1. El estadístico de Chow

Existen diversas técnicas para contrastar las situaciones de inestabilidad o sensibilidad de una ecuación estimada para una muestra dada. Todas están en torno a la discusión de que, sí es factible suponer cambios en cada uno de los puntos muestrales, o que puede identificarse una observación de cambio llamada punto de cambio o quiebre, un contraste muy conocido para la estabilidad de las relaciones entre dos ecuaciones

(modelos) estimadas es el de Chow.

La filosofía principal del estadístico de Chow, consiste en evaluar estadísticamente si el vector de parámetros estimado para toda la muestra difiere de aquel que se calcula considerando dos o más submuestras.

Consideremos el modelo lineal $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, donde \mathbf{X} es una matriz de dimensión $n \times k$, si hay dos submuestras, (1) y (2), en las cuales los parámetros no son necesariamente los mismos, entonces podemos escribir:

$$\mathbf{y}_i = \begin{cases} \mathbf{X}_i\boldsymbol{\theta}_1 + \boldsymbol{\varepsilon}_i & i \in (1) \\ \mathbf{X}_i\boldsymbol{\theta}_2 + \boldsymbol{\varepsilon}_i & i \in (2) \end{cases}$$

y la hipótesis que deseamos probar es:

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$$

el número total de observaciones es $n = n_1 + n_2$ y el número de parámetros es k .

Es decir, el estadístico de Chow está basado en una subdivisión de la muestra, a continuación se describe el proceso paso a paso.

1.- Dividimos la muestra de n observaciones en dos grupos. El primer grupo tendrá n_1 observaciones, y el segundo grupo tendrá n_2 observaciones, donde $n_2 = n - n_1$.

2.- Se ejecuta una regresión separadamente para cada uno de los dos grupos de la muestra y calculamos la suma de los cuadrados de los errores para cada grupo de la muestra. SCE_1 y SCE_2 . La suma de los cuadrados del error es la suma de los cuadrados de los errores de cada muestra: $SCE_1 + SCE_2$.

3.- Asumimos que el coeficiente de regresión es el mismo antes y después del periodo n_1 . Estimamos el modelo otra vez pero sobre las n observaciones. Y obtenemos la suma de cuadrados del error: SCE .

4.- El estadístico de prueba es:

$$F = \frac{(SCE - SCE_1 - SCE_2) / k}{(SCE_1 + SCE_2) / (n - 2k)}.$$

5.- Si $F > F_{k, n-2k}$, rechazamos la hipótesis nula, y entonces se tendría un cambio estructural en la muestra.

Conerly y Mansfiel (1988), proponen un estadístico de Chow modificado, el cuál se basa en la aproximación de Satterthwaite (1946), es decir, en las estimaciones usuales de mínimos cuadrados de los dos vectores paramétricos. (para más detalles ver Conerly y Mansfiel, (1988)).

4.2. Definición y motivación

Supóngase que se tiene disponible un conjunto de n_1 observaciones sobre (y, x_2, \dots, x_k) que satisface al modelo de regresión normal

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\theta}_1 + \boldsymbol{\varepsilon}_1, \quad \boldsymbol{\varepsilon}_1 \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}), \quad (4.1)$$

donde $\mathbf{y}_1 = (y_1, y_2, \dots, y_{n_1})^t$, $\boldsymbol{\theta}_1 = (\theta_{11}, \theta_{21}, \dots, \theta_{k1})^t$ y \mathbf{X}_1 es una matriz de dimensión $n_1 \times k$ de valores observados de los $(k-1)$ regresores. Como es usual, la primera columna de \mathbf{X}_1 se toma como un vector de unos.

Además, se dispone de un conjunto de n_2 observaciones para el cual

$$\mathbf{y}_2 = \mathbf{X}_2 \boldsymbol{\theta}_2 + \varepsilon_2, \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2 \mathbf{I}). \quad (4.2)$$

El problema de contraste $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ vs $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ es de interés práctico y además teóricamente es desafiante; lo han estudiado muchos autores entre los que destacan Zellner (1962), Theil (1971, pp.131), Smith (1975), (1980), Smith y Cook (1980), y Williford, Smith, Mansfiel, Conerly y Bishop (1986).

Por tanto para las observaciones $(n_1, \mathbf{y}_1, \mathbf{X}_1)$ se tiene que

$$y_T = \mathbf{x}_1^t(T) \boldsymbol{\theta}_1 + \varepsilon_1(T), \quad \varepsilon_1(T) \sim \mathcal{N}(0, \sigma_1^2),$$

para $T = 1, 2, \dots, n_1$, y para las observaciones $(n_2, \mathbf{y}_2, \mathbf{X}_2)$ se tiene que

$$y_T = \mathbf{x}_2^t(T) \boldsymbol{\theta}_2 + \varepsilon_2(T), \quad \varepsilon_2(T) \sim \mathcal{N}(0, \sigma_2^2),$$

para $T = n_1 + 1, \dots, n_1 + n_2$. El contraste H_0 vs H_1 , es equivalente al problema de demostrar si ocurre un cambio puntual sobre $T = n_1 + 1$.

Bajo el enfoque frecuentista el problema tiene una solución exacta solamente cuando se acepta la condición de homoscedasticidad, es decir, cuando se tiene que $\sigma_1 = \sigma_2$, se obtiene usando el procedimiento del cociente de verosimilitudes y se basa en el estadístico

$$F = \left(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \right)^t \left[\left(\mathbf{X}_1^t \mathbf{X}_1 \right)^{-1} - \left(\mathbf{X}_2^t \mathbf{X}_2 \right)^{-1} \right]^{-1} \left(\hat{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_1 \right) / ks^2, \quad (4.3)$$

donde

$$\hat{\boldsymbol{\theta}}_i = \left(\mathbf{X}_i^t \mathbf{X}_i \right)^{-1} \mathbf{X}_i^t \mathbf{y}_i,$$

$$(n_1 + n_2 - 2k) s^2 = (n_1 - k) s_1^2 + (n_2 - k) s_2^2,$$

y

$$(n_i - k) s_i^2 = \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\theta}}_i \right)^t \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\theta}}_i \right).$$

Cuando H_0 es verdadero, el estadístico (4.3) tiene una distribución F con k y $(n_1 + n_2 - 2k)$ grados de libertad (Chow 1960), tal y como se expone en la primera sección de este capítulo.

Cuando no se acepta la condición de homoscedasticidad, el estadístico F tiene una distribución que depende del parámetro de ruido σ_1/σ_2 . Algunas referencias de aproximaciones a la distribución del estadístico F , modificaciones de él, y pruebas para el contraste H_0 vs H_1 son, entre otras, Toyoda (1974); Schmidt y Sickles (1977); Ali y

Silver (1985) y Conerly y Mansfiel (1988). Sin embargo, no se tiene claro cual de éstas aproximaciones puede usarse para una aplicación específica.

En el enfoque bayesiano para el contraste de hipótesis, el factor de Bayes es la herramienta principal, (ver, por ejemplo, Jeffreys (1961), pp. 245-249 y Berger (1985), sección 4.3.3.). Esta es una alternativa a la utilización de regiones de máxima densidad, como veíamos en la subsección 2.2.5, los cuales sitúan el problema de contraste de hipótesis en un marco de estimación (Box y Tiao (1973), sección 2.9).

Moreno, Torres y Casella (2002), notan que el contraste H_0 vs H_1 es un problema de dos muestras, que puede plantearse como un problema de comparación de dos modelos. De hecho, si tomamos $\phi_p(y|\mu, \Sigma)$, que denota la densidad de la distribución normal p -variada con media μ y matriz de covarianzas Σ , entonces, bajo H_0 , la muestra $(n_1, \mathbf{y}_1, \mathbf{X}_1, n_2, \mathbf{y}_2, \mathbf{X}_2)$ tiene una densidad

$$f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) = \phi_{n_1}(\mathbf{y}_1 | \mathbf{X}_1 \boldsymbol{\theta}, \tau_1^2 \mathbf{I}) \phi_{n_2}(\mathbf{y}_2 | \mathbf{X}_2 \boldsymbol{\theta}, \tau_2^2 \mathbf{I}), \quad (4.4)$$

y bajo H_1

$$f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = \phi_{n_1}(\mathbf{y}_1 | \mathbf{X}_1 \boldsymbol{\theta}_1, \sigma_1^2 \mathbf{I}) \phi_{n_2}(\mathbf{y}_2 | \mathbf{X}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}). \quad (4.5)$$

Para comparar los modelos (4.4) y (4.5), calculamos el factor de Bayes

$$B_{12}(\mathbf{y}_1, \mathbf{y}_2) = m_1(\mathbf{y}_1, \mathbf{y}_2) / m_2(\mathbf{y}_1, \mathbf{y}_2),$$

donde

$$m_1(\mathbf{y}_1, \mathbf{y}_2) = \int f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \pi_1(\boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta} d\tau_1 d\tau_2,$$

y

$$m_2(\mathbf{y}_1, \mathbf{y}_2) = \int f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) \pi_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2,$$

son las distribuciones marginales de los datos para las distribuciones a priori

$$\pi_1(\boldsymbol{\theta}, \tau_1, \tau_2)$$

y

$$\pi_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2)$$

respectivamente.

Con información a priori no subjetiva, la distribución a priori comúnmente usada es una densidad uniforme para las $\boldsymbol{\theta}$'s y una densidad uniforme para $\log \sigma^2$, como vimos en la subsección 2.2.3. Por lo tanto, una formulación bayesiana por defecto para el

contraste de hipótesis de H_0 vs H_1 es la de comparación del modelo M_1 contra el modelo M_2 , (ver subsección 2.2.5) donde

$$M_1 : \left\{ f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2), \pi_1(\boldsymbol{\theta}, \tau_1, \tau_2) = \frac{c_1}{\tau_1^2 \tau_2^2} \right\}, \quad (4.6)$$

y

$$M_2 : \left\{ f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2), \pi_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = \frac{c_2}{\sigma_1^2 \sigma_2^2} \right\}. \quad (4.7)$$

Aquí, las c_i 's son constantes positivas arbitrarias que no pueden determinarse (ver Berger y Pericchi, (1996)), porque π_i^N no es una función integrable. Para estas distribuciones a priori impropias el factor de Bayes resulta

$$B_{12}^N(\mathbf{y}_1, \mathbf{y}_2) = \frac{c_1 \int f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \tau_1^{-2} \tau_2^{-2} d\boldsymbol{\theta} d\tau_1 d\tau_2}{c_2 \int f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2},$$

por lo que puede apreciarse que depende de la constante arbitraria c_1/c_2 , y en consecuencia no está definido de manera única. (Moreno, E. Torres, F. and Casella, G. (2002)). De nuevo, la inclusión de estas c_i es debatible, y si consideramos valores constantes 1, también pueden realizarse los cálculos correspondientes.

Sin embargo, las distribuciones a priori impropias para parámetros en modelos anidados pueden convertirse, a través del método intrínseco (Berger y Pericchi (1996); Moreno (1997); y Moreno, Bertolino y Racugno (1988)) en distribuciones a priori intrínsecas, para las cuales el factor de Bayes está bien definido. Algunas justificaciones para el uso de distribuciones a priori intrínsecas para selección de modelos y contraste de hipótesis han sido proporcionadas por Berger y Pericchi (1996, 1997a, 1997b, 1988), Moreno (1997), Moreno, Bertolino y Racugno (1998) y De Santis y Spezzaferri (1999). El método ha sido previamente usado para proporcionar distribuciones a priori intrínsecas razonables para una variedad de problemas de contraste de hipótesis (Berger y Pericchi (1996, 1998); Moreno, Bertolino y Racugno (1998, 1999); Bertolino, Moreno y Racugno (2000); Moreno y Liseo (2001); entre otros).

Como resumen, podemos decir que por definición, el factor de Bayes para distribuciones a priori intrínsecas y el factor de Bayes intrínseco aritmético (FBIA), definido en el capítulo 3, coinciden asintóticamente. Para tamaños de muestra pequeños o moderados, el FBIA puede ser inestable, por ejemplo, para muestras que tienen un número pequeño de muestras de entrenamiento minimal. Además, el FBIA no existe cuando el tamaño de muestra es más pequeño que el tamaño de muestra de entrenamiento minimal, una situación que se encuentra en el problema de cambio puntual. Finalmente, el FBIA no satisface la relación de coherencia $P(M_1 | \mathbf{x}) = 1 - P(M_2 | \mathbf{x})$. El factor de Bayes para distribuciones a priori intrínsecas es un factor de Bayes verdadero asintóticamente y por lo tanto, su comportamiento es coherente.

Puesto que el modelo M_1 está anidado en el modelo M_2 , las distribuciones a priori intrínsecas pueden ser derivadas usando la teoría intrínseca estándar.

En un escenario por defecto, es usual asumir que cada modelo tiene una probabilidad a priori de $1/2$, entonces la probabilidad a posteriori de M_1 está dada por

$$P(M_1|\mathbf{y}_1, \mathbf{y}_2) = \frac{1}{1 + B_{21}(\mathbf{y}_1, \mathbf{y}_2)}, \quad (4.8)$$

ya que por definición

$$\begin{aligned} P(M_1|\mathbf{y}_1, \mathbf{y}_2) &= \frac{P(\mathbf{y}_1, \mathbf{y}_2|M_1) P(M_1)}{\sum_{i=1}^2 P(\mathbf{y}_1, \mathbf{y}_2|M_i) P(M_i)} = \frac{P(\mathbf{y}_1, \mathbf{y}_2|M_1)^{\frac{1}{2}}}{P(\mathbf{y}_1, \mathbf{y}_2|M_1)^{\frac{1}{2}} + P(\mathbf{y}_1, \mathbf{y}_2|M_2)^{\frac{1}{2}}} \\ &= \frac{P(\mathbf{y}_1, \mathbf{y}_2|M_1)}{P(\mathbf{y}_1, \mathbf{y}_2|M_2) + P(\mathbf{y}_1, \mathbf{y}_2|M_2)} = \frac{m_1(\mathbf{y}_1, \mathbf{y}_2)}{m_1(\mathbf{y}_1, \mathbf{y}_2) + m_2(\mathbf{y}_1, \mathbf{y}_2)} \\ &= \frac{1}{\frac{m_1(\mathbf{y}_1, \mathbf{y}_2) + m_2(\mathbf{y}_1, \mathbf{y}_2)}{m_1(\mathbf{y}_1, \mathbf{y}_2)}} = \frac{1}{1 + \frac{m_2(\mathbf{y}_1, \mathbf{y}_2)}{m_1(\mathbf{y}_1, \mathbf{y}_2)}} = \frac{1}{1 + B_{21}(\mathbf{y}_1, \mathbf{y}_2)}, \end{aligned}$$

lo que demuestra el resultado (4.8).

La probabilidad a posteriori en (4.8) es una medida del soporte que los datos dan a la hipótesis nula (ver, por ejemplo, Lavine y Schervish (1999)). En efecto, la decisión óptima bajo pérdida 0-1, es elegir M_1 si la desigualdad $P(M_1|\mathbf{y}_1, \mathbf{y}_2) \geq P(M_2|\mathbf{y}_1, \mathbf{y}_2)$ se cumple. Además, ésta es una regla consistente: se elige el verdadero modelo con probabilidad que tiende a uno cuando el tamaño de la muestra tiende a infinito (O'Hagan (1994), sección 7.52).

El supuesto de homoscedasticidad reduce la complejidad de los modelos subyacentes. De aquí en adelante, estaremos interesados en comprobar si realmente tal hipótesis es apoyada por los datos disponibles.

4.3. Distribuciones a priori intrínsecas y factor de Bayes

En esta sección se hará uso de los factores de Bayes intrínsecos definidos en el capítulo anterior. Así como la metodología intrínseca desarrollada en el mismo capítulo, para calcular las distribuciones a priori intrínsecas para modelos de regresión heteroscedásticos.

4.3.1. Distribuciones a priori intrínsecas.

Como vimos en el capítulo 3, si se asume que $f_1(\mathbf{x}|\boldsymbol{\theta}_1)$ está anidado en $f_2(\mathbf{x}|\boldsymbol{\theta}_2)$ y se tiene que para algún tamaño de muestra n la función de verosimilitud

$$f_2(x_1, x_2, \dots, x_n|\boldsymbol{\theta}_2)$$

es una función integrable con respecto a $\pi_2^N(\boldsymbol{\theta}_2)$, entonces, la distribución a priori intrínseca (Berger y Pericchi, (1996)) es la distribución a priori propia $\pi_2^I(\boldsymbol{\theta}_2)$ definida por

$$\pi_2^I(\boldsymbol{\theta}_2) = \pi_2^N(\boldsymbol{\theta}_2) E_{x(l)|\boldsymbol{\theta}_2}^{M_2} [B_{12}^N(x(l))], \quad (4.9)$$

donde la esperanza se toma con respecto a la densidad de la muestra de entrenamiento minimal $x(l)$ bajo el modelo M_2 .

La construcción de la distribución a priori intrínseca (4.9) requiere de una distribución a priori propia $\pi_1(\boldsymbol{\theta}_1)$ para el modelo más simple. Sin embargo, en el análisis por defecto, típicamente esta distribución a priori es impropia. En esta situación podemos proceder de la siguiente manera, colocamos una distribución a priori degenerada sobre un punto $\boldsymbol{\theta}_1$, y se halla la correspondiente distribución a priori intrínseca (4.9) condicional sobre este punto, $\pi_2^I(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$. Entonces, la distribución a priori intrínseca $\pi_2^I(\boldsymbol{\theta}_2)$ se obtiene integrando sobre el parámetro $\boldsymbol{\theta}_1$ en $\pi_2^I(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ usando $\pi_1^N(\boldsymbol{\theta}_1)$.

Este procedimiento es equivalente al procedimiento límite intrínseco propuesto por Moreno, Bertolino y Racugno (1998). La ventaja es que la distribución a priori propia $\pi_2^I(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ tiene una interpretación interesante, como veremos en secciones posteriores.

4.3.2. Distribuciones a priori intrínsecas para modelos de regresión heteroscedásticos.

Para los modelos (4.6) y (4.7) es fácil ver que una muestra de entrenamiento es un vector $(\mathbf{z}_1, \mathbf{z}_2)$ de $2(k+1)$ réplicas independientes de (y, x_2, \dots, x_k) .

Para demostrar que el tamaño de muestra minimal para los modelos (4.6) y (4.7) es un vector $(\mathbf{z}_1, \mathbf{z}_2)$ hay que verificar que

$$0 < m_i(\mathbf{z}_1, \mathbf{z}_2) < \infty, \quad i = 1, 2.$$

Demostración. -

$$\begin{aligned}
m_1(\mathbf{z}_1, \mathbf{z}_2) &= \int \phi_{k+1}(\mathbf{z}_1 | \mathbf{Z}_1 \boldsymbol{\theta}, \tau_1^2 \mathbf{I}) \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}, \tau_2^2 \mathbf{I}) \frac{1}{\tau_1^2 \tau_2^2} d\boldsymbol{\theta} d\tau_1 d\tau_2 \\
&= \iiint |\tau_1^2 \mathbf{I}|^{-1/2} (2\pi)^{-(k+1)/2} \exp \left\{ -\frac{1}{2\tau_1^2} (\mathbf{z}_1 - \mathbf{Z}_1 \boldsymbol{\theta})^t (\mathbf{z}_1 - \mathbf{Z}_1 \boldsymbol{\theta}) \right\} \\
&\quad \times |\tau_2^2 \mathbf{I}|^{-1/2} (2\pi)^{-(k+1)/2} \exp \left\{ -\frac{1}{2\tau_2^2} (\mathbf{z}_2 - \mathbf{Z}_2 \boldsymbol{\theta})^t (\mathbf{z}_2 - \mathbf{Z}_2 \boldsymbol{\theta}) \right\} \\
&\quad \times \frac{1}{\tau_1^2 \tau_2^2} d\boldsymbol{\theta} d\tau_1 d\tau_2 \\
&= \iiint (\tau_1 \tau_2)^{-(k+1)} (2\pi)^{-(k+1)} \\
&\quad \times \exp \left\{ -\frac{1}{2\tau_1^2} (\mathbf{z}_1^t \mathbf{z}_1 - \mathbf{z}_1^t \mathbf{Z}_1 \boldsymbol{\theta} - \boldsymbol{\theta}^t \mathbf{Z}_1^t \mathbf{z}_1 - \boldsymbol{\theta}^t \mathbf{Z}_1^t \mathbf{Z}_1 \boldsymbol{\theta}) \right\} \\
&\quad \times \exp \left\{ -\frac{1}{2\tau_2^2} (\mathbf{z}_2^t \mathbf{z}_2 - \mathbf{z}_2^t \mathbf{Z}_2 \boldsymbol{\theta} - \boldsymbol{\theta}^t \mathbf{Z}_2^t \mathbf{z}_2 - \boldsymbol{\theta}^t \mathbf{Z}_2^t \mathbf{Z}_2 \boldsymbol{\theta}) \right\} \frac{1}{\tau_1^2 \tau_2^2} d\boldsymbol{\theta} d\tau_1 d\tau_2 \\
&= \iiint (\tau_1 \tau_2)^{-(k+1)} (2\pi)^{-(k+1)} \exp \left\{ -\frac{\mathbf{z}_1^t \mathbf{z}_1}{2\tau_1^2} - \frac{\mathbf{z}_2^t \mathbf{z}_2}{2\tau_2^2} \right\} \\
&\quad \times \exp \left\{ \frac{1}{2} \boldsymbol{\theta}' \left(\frac{\mathbf{Z}_1^t \mathbf{Z}_1}{\tau_1^2} + \frac{\mathbf{Z}_2^t \mathbf{Z}_2}{\tau_2^2} \right) \boldsymbol{\theta} + \boldsymbol{\theta}' \left(\frac{\mathbf{Z}_1^t \mathbf{z}_1}{\tau_1^2} + \frac{\mathbf{Z}_2^t \mathbf{z}_2}{\tau_2^2} \right) \right. \\
&\quad \left. + \frac{1}{2} \boldsymbol{\theta}' \left(\frac{\mathbf{Z}_1^t \mathbf{z}_1}{\tau_1^2} + \frac{\mathbf{Z}_2^t \mathbf{z}_2}{\tau_2^2} \right) \boldsymbol{\theta} \right\} \\
&\quad \times \frac{1}{\tau_1^2 \tau_2^2} d\boldsymbol{\theta} d\tau_1 d\tau_2,
\end{aligned}$$

completando cuadrados y tomando

$$\begin{aligned}
-\frac{\mathbf{z}_1^t \mathbf{z}_1}{2\tau_1^2} - \frac{\mathbf{z}_2^t \mathbf{z}_2}{2\tau_2^2} &= -\frac{1}{2} \sum_{i=1}^2 \mathbf{z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{z}_i, \\
\frac{\mathbf{Z}_1^t \mathbf{z}_1}{\tau_1^2} + \frac{\mathbf{Z}_2^t \mathbf{z}_2}{\tau_2^2} &= \sum_{i=1}^2 \mathbf{Z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{z}_i, \\
\frac{\mathbf{Z}_1^t \mathbf{Z}_1}{\tau_1^2} + \frac{\mathbf{Z}_2^t \mathbf{Z}_2}{\tau_2^2} &= \sum_{i=1}^2 \mathbf{Z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{Z}_i,
\end{aligned}$$

tenemos que

$$\begin{aligned}
 m_1(\mathbf{z}_1, \mathbf{z}_2) &= \iint \exp \left\{ -\frac{1}{2} \sum_{i=1}^2 \mathbf{z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{z}_i + \frac{1}{2} \left(\sum_{i=1}^2 \mathbf{Z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{z}_i \right)^t \right. \\
 &\quad \left. \times \left(\sum_{i=1}^2 \mathbf{Z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{Z}_i \right)^{-1} \left(\sum_{i=1}^2 \mathbf{Z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{z}_i \right) \right\} \\
 &\quad \frac{1}{(2\pi)^{(k+2)/2} (\tau_1)^{k+1} (\tau_2)^{k+1} \left| \sum_{i=1}^2 \mathbf{Z}_i^t (\tau_i^2 \mathbf{I})^{-1} \mathbf{Z}_i \right|^{1/2}} \\
 &\quad \times \frac{1}{\tau_1^2 \tau_2^2} d\tau_1 d\tau_2,
 \end{aligned}$$

haciendo un cambio de variables a coordenadas polares, tomando $\tau_1 = r \operatorname{sen} \varphi$ y $\tau_2 = r \cos \varphi$; donde $0 < r < \infty$ y $0 < \varphi < \frac{\pi}{2}$; nos queda que

$$m_1(\mathbf{z}_1, \mathbf{z}_2) = \frac{\Gamma\left(\frac{k+4}{2}\right)}{2(2\pi)^{(k+2)/2}} J < \infty,$$

donde

$$J = \int_0^{\pi/2} \frac{(\operatorname{sen} \varphi)^{(k+1)} (\cos \varphi)^{(k+1)}}{Q_1^{(k+4)/2} |\mathbf{P}_1|^{1/2}} d\varphi,$$

$$\mathbf{P}_1 = \mathbf{Z}_1^t \mathbf{Z}_1 (\operatorname{sen} \varphi)^2 + \mathbf{Z}_2^t \mathbf{Z}_2 (\cos \varphi)^2,$$

y

$$Q_1 = \mathbf{z}_1^t \mathbf{z}_1 (\operatorname{sen} \varphi)^2 + \mathbf{z}_2^t \mathbf{z}_2 (\cos \varphi)^2 - \mathbf{C}^t \mathbf{P}_1^{-1} \mathbf{C},$$

donde

$$\mathbf{C} = \mathbf{Z}_1^t \mathbf{z}_1 (\operatorname{sen} \varphi)^2 + \mathbf{Z}_2^t \mathbf{z}_2 (\cos \varphi)^2,$$

la integral J no puede calcularse analíticamente, pero la integración numérica es muy práctica ya que su dominio es un conjunto acotado.

Por otro lado

$$\begin{aligned}
 &m_2(\mathbf{z}_1, \mathbf{z}_2) \\
 &= \iiint \phi_{k+1}(\mathbf{z}_1 | \mathbf{Z}_1 \boldsymbol{\theta}_1, \sigma_1^2 \mathbf{I}) \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}) \frac{1}{\sigma_1^2 \sigma_2^2} d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2 \\
 &= \iiint \left| \sigma_1^2 \mathbf{I} \right|^{-1/2} (2\pi)^{-(k+1)/2} \exp \left\{ -\frac{1}{2\sigma_1^2} (\mathbf{z}_1 - \mathbf{Z}_1 \boldsymbol{\theta}_1)^t (\mathbf{z}_1 - \mathbf{Z}_1 \boldsymbol{\theta}_1) \right\} d\boldsymbol{\theta}_1 \\
 &\quad \times \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}) \frac{1}{\sigma_1^2 \sigma_2^2} d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2,
 \end{aligned}$$

pero

$$\begin{aligned}
& (\mathbf{z}_1 - \mathbf{Z}_1 \boldsymbol{\theta}_1)^t (\mathbf{z}_1 - \mathbf{Z}_1 \boldsymbol{\theta}_1) \\
&= \left(\mathbf{z}_1 - \mathbf{Z}_1 \hat{\boldsymbol{\theta}}_1 + \mathbf{Z}_1 \hat{\boldsymbol{\theta}}_1 - \mathbf{Z}_1 \boldsymbol{\theta}_1 \right)^t \left(\mathbf{z}_1 - \mathbf{Z}_1 \hat{\boldsymbol{\theta}}_1 + \mathbf{Z}_1 \hat{\boldsymbol{\theta}}_1 - \mathbf{Z}_1 \boldsymbol{\theta}_1 \right) \\
&= \mathbf{z}_1^t \mathbf{H}_1 \mathbf{z}_1 + \left(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \right)^t \mathbf{Z}_1^t \mathbf{Z}_1 \left(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \right),
\end{aligned}$$

donde $\mathbf{H}_1 = \mathbf{I} - \mathbf{Z}_1 (\mathbf{Z}_1^t \mathbf{Z}_1)^{-1} \mathbf{Z}_1^t$, entonces sustituyendo en $m_2(\mathbf{z}_1, \mathbf{z}_2)$ tenemos que

$$\begin{aligned}
m_2(\mathbf{z}_1, \mathbf{z}_2) &= \iiint (\sigma_1)^{-(k+1)} (2\pi)^{-(k+1)/2} \\
&\quad \times \exp \left\{ -\frac{1}{2\sigma_1^2} \left(\mathbf{z}_1^t \mathbf{H}_1 \mathbf{z}_1 + \left(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \right)^t \mathbf{Z}_1^t \mathbf{Z}_1 \left(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \right) \right) \right\} d\boldsymbol{\theta}_1 \\
&\quad \times \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}) \frac{1}{\sigma_1^2 \sigma_2^2} d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2 \\
&= \iiint (\sigma_1)^{-(k+1)} (2\pi)^{-(k+1)/2} \exp \left\{ -\frac{\mathbf{z}_1^t \mathbf{H}_1 \mathbf{z}_1}{2\sigma_1^2} \right\} \\
&\quad \times \exp \left\{ -\frac{1}{2\sigma_1^2} \left(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \right)^t \mathbf{Z}_1^t \mathbf{Z}_1 \left(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \right) \right\} d\boldsymbol{\theta}_1 \\
&\quad \times \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}) \frac{1}{\sigma_1^2 \sigma_2^2} d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2 \\
&= \iiint (\sigma_1)^{-(k+1)} (2\pi)^{-(k+1)/2} \exp \left\{ -\frac{\mathbf{z}_1^t \mathbf{H}_1 \mathbf{z}_1}{2\sigma_1^2} \right\} \frac{(2\pi)^{k/2}}{\left| \left(\frac{\mathbf{Z}_1^t \mathbf{Z}_1}{\sigma_1^2} \right)^{-1} \right|^{-1/2}} \\
&\quad \times \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}) \frac{1}{\sigma_1^2 \sigma_2^2} d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2 \\
&= \iiint \frac{(\sigma_1)^{-1} (2\pi)^{-1/2}}{|\mathbf{Z}_1^t \mathbf{Z}_1|^{1/2}} \exp \left\{ -\frac{\mathbf{z}_1^t \mathbf{H}_1 \mathbf{z}_1}{2\sigma_1^2} \right\} \\
&\quad \times \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}) \frac{1}{\sigma_1^2 \sigma_2^2} d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2.
\end{aligned}$$

Análogamente

$$\int \phi_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}) d\boldsymbol{\theta}_2 = \frac{(\sigma_2)^{-1} (2\pi)^{-1/2}}{|\mathbf{Z}_2^t \mathbf{Z}_2|^{1/2}} \exp \left\{ -\frac{\mathbf{z}_2^t \mathbf{H}_2 \mathbf{z}_2}{2\sigma_2^2} \right\}.$$

Sustituyendo en $m_2(\mathbf{z}_1, \mathbf{z}_2)$ tenemos que, por el apéndice A3.3 tomando $p = 2$, $a = (\mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i) / 2$,

$$\begin{aligned}
 & m_2(\mathbf{z}_1, \mathbf{z}_2) \\
 = & \iint \frac{(\sigma_1)^{-1} (2\pi)^{-1/2}}{|\mathbf{Z}_1^t \mathbf{Z}_1|^{1/2}} \exp \left\{ -\frac{\mathbf{z}_1^t \mathbf{H}_1 \mathbf{z}_1}{2\sigma_1^2} \right\} \frac{(\sigma_2)^{-1} (2\pi)^{-1/2}}{|\mathbf{Z}_2^t \mathbf{Z}_2|^{1/2}} \\
 & \times \exp \left\{ -\frac{\mathbf{z}_2^t \mathbf{H}_2 \mathbf{z}_2}{2\sigma_2^2} \right\} \frac{d\sigma_1 d\sigma_2}{\sigma_1^2 \sigma_2^2} \\
 = & (2\pi)^{-1} \prod_{i=1}^2 \int \frac{\sigma_i^3}{|\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2}} \exp \left\{ -\frac{\mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i}{2\sigma_i^2} \right\} d\sigma_i \\
 = & (2\pi)^{-1} \prod_{i=1}^2 \frac{1}{2} \left(\frac{\mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i}{2} \right)^{-1} \frac{1}{|\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2}} = \frac{1}{(2\pi) \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i} < \infty,
 \end{aligned}$$

lo que demuestra el resultado anterior.

Las densidades de la muestra de entrenamiento para los modelos son

$$(\mathbf{z}_1, \mathbf{z}_2) \sim \mathcal{N}_{k+1}(\mathbf{Z}_1 \boldsymbol{\theta}, \tau_1^2 \mathbf{I}) \times \mathcal{N}_{k+1}(\mathbf{Z}_2 \boldsymbol{\theta}, \tau_2^2 \mathbf{I}) \text{ bajo el modelo } M_1,$$

y

$$(\mathbf{z}_1, \mathbf{z}_2) \sim \mathcal{N}_{k+1}(\mathbf{Z}_1 \boldsymbol{\theta}_1, \sigma_1^2 \mathbf{I}) \times \mathcal{N}_{k+1}(\mathbf{Z}_2 \boldsymbol{\theta}_1, \sigma_2^2 \mathbf{I}) \text{ bajo el modelo } M_2.$$

Se asume que $\text{rango}(\mathbf{Z}_i) = k$, para $i = 1, 2$.

Lema 2 La distribución a priori intrínseca para la comparación de los modelos (4.6) y (4.7) es

$$\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = \int \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta} d\tau_1 d\tau_2,$$

donde

$$\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\theta}, \tau_1, \tau_2) = \prod_{i=1}^2 \phi_k(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) \times \frac{1}{\tau_i} \left(1 + \frac{\sigma_i^2}{\tau_i^2} \right)^{-3/2}, \quad (4.10)$$

donde ϕ_k es la densidad de una distribución $\mathcal{N}_k(\boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1})$.

Demostración.-

Puesto que se quiere la distribución a priori condicional sobre $(\boldsymbol{\theta}, \tau_1, \tau_2)$, entonces aplicamos la ecuación (4.9) a los modelos

$$M_1^* : \mathcal{N}_{k+1}(\mathbf{z}_1 | \mathbf{Z}_1 \boldsymbol{\theta}, \tau_1^2 \mathbf{I}) \cdot \mathcal{N}_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}, \tau_2^2 \mathbf{I}), \quad \pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) = \delta(\boldsymbol{\theta}, \tau_1, \tau_2),$$

y

$$M_2^* : \mathcal{N}_{k+1}(\mathbf{z}_1 | \mathbf{Z}_1 \boldsymbol{\theta}_1, \tau_1^2 \mathbf{I}) \cdot \mathcal{N}_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \tau_2^2 \mathbf{I}), \quad \pi_2^N(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = \frac{c}{\sigma_1^2 \sigma_2^2},$$

donde $\delta(\boldsymbol{\theta}, \tau_1, \tau_2)$ es la delta de Dirac en el punto $(\boldsymbol{\theta}, \tau_1, \tau_2)$.

Haciendo los cálculos respectivos se puede ver que

$$B_{12}^N(\mathbf{z}_1, \mathbf{z}_2) = \frac{2\pi}{c} \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \cdot \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}, \tau_i^2 \mathbf{I}),$$

ya que, por definición tenemos

$$\begin{aligned} B_{12}^N(\mathbf{z}_1, \mathbf{z}_2) &= \frac{m_1^*(z_1, \mathbf{z}_2)}{m_2^*(z_1, \mathbf{z}_2)} = \frac{\prod_{i=1}^2 \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}, \tau_i^2 \mathbf{I})}{m_2(z_1, \mathbf{z}_2)} \\ &= \frac{\prod_{i=1}^2 \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}, \tau_i^2 \mathbf{I})}{\frac{c}{(2\pi)^2 \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i}} \\ &= \frac{2\pi}{c} \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}, \tau_i^2 \mathbf{I}), \end{aligned}$$

donde

$$\mathbf{H}_i = \mathbf{I} - \mathbf{Z}_i (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \mathbf{Z}_i^t.$$

Por (3.38), teníamos que

$$\begin{aligned} &\pi^I(\boldsymbol{\theta}_2) \\ &= \pi_2^N(\boldsymbol{\theta}_2) E_{x(l)|\boldsymbol{\theta}_2}^{M_2} [B_{12}^N(\mathbf{z}_1, \mathbf{z}_2)] \\ &= \frac{c}{\sigma_1^2 \sigma_2^2} \iint \frac{2\pi}{c} \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}, \tau_i^2 \mathbf{I}) \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) dz_1 dz_2 \\ &= \frac{2\pi}{\sigma_1^2 \sigma_2^2} \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \prod_{i=1}^2 \int \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}, \tau_i^2 \mathbf{I}) \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) dz_i. \end{aligned}$$

Ahora, desarrollando y completando cuadrados tenemos que :

$$\begin{aligned} &\int \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \cdot \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}, \tau_i^2 \mathbf{I}) \cdot \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) dz_i \\ &= \frac{1}{(2\pi)^{1/2}} \mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) \frac{\sigma_i^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{-1/2}}{\tau_i (1 + \sigma_i^2 / \tau_i^2)^{3/2}}, \end{aligned}$$

y sustituyendo en $\pi^I(\boldsymbol{\theta}_2)$ tenemos que

$$\begin{aligned}
 & \pi^I(\boldsymbol{\theta}_2) \\
 &= \frac{2\pi}{\sigma_1^2 \sigma_2^2} \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \prod_{i=1}^2 \frac{1}{(2\pi)^{1/2}} \mathcal{N}_k \left(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right) \frac{\sigma_i^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{-1/2}}{\tau_i (1 + \sigma_i^2 / \tau_i^2)^{3/2}} \\
 &= \frac{2\pi}{\sigma_1^2 \sigma_2^2} |\mathbf{Z}_1^t \mathbf{Z}_1|^{1/2} |\mathbf{Z}_2^t \mathbf{Z}_2|^{1/2} \frac{1}{(2\pi)^{1/2}} \frac{1}{(2\pi)^{1/2}} \frac{\sigma_1^2 |\mathbf{Z}_1^t \mathbf{Z}_1|^{-1/2}}{\tau_1 (1 + \sigma_1^2 / \tau_1^2)^{3/2}} \frac{\sigma_2^2 |\mathbf{Z}_2^t \mathbf{Z}_2|^{-1/2}}{\tau_2 (1 + \sigma_2^2 / \tau_2^2)^{3/2}} \\
 &\quad \times \prod_{i=1}^2 \mathcal{N}_k \left(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right) \\
 &= \prod_{i=1}^2 \mathcal{N}_k \left(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right) \times \frac{1}{\tau_i (1 + \sigma_i^2 / \tau_i^2)^{3/2}} \\
 &= \prod_{i=1}^2 \mathcal{N}_k \left(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right) \times \frac{1}{\tau_i} (1 + \sigma_i^2 / \tau_i^2)^{-3/2},
 \end{aligned}$$

por lo tanto,

$$\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\theta}, \tau_1, \tau_2) = \prod_{i=1}^2 \mathcal{N}_k \left(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right) \times \frac{1}{\tau_i} (1 + \sigma_i^2 / \tau_i^2)^{-3/2},$$

por lo que el lema queda demostrado.

Por construcción, la distribución a priori intrínseca $\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\theta}, \tau_1, \tau_2)$ integra uno para cualquier punto $(\boldsymbol{\theta}, \tau_1, \tau_2)$. Sin embargo, la distribución a priori intrínseca incondicional es impropia; aunque sea límite único de distribuciones a priori intrínsecas (Moreno, Bertolino y Racugno, (1998)). De aquí en adelante, nos referiremos al par $(\pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2), \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2))$ como distribuciones a priori intrínsecas para los modelos M_1 y M_2 respectivamente.

Nota 4 En la distribución a priori intrínseca (4.10), $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ son independientes, condicional sobre $(\boldsymbol{\theta}, \tau_1, \tau_2, \sigma_1, \sigma_2)$, también σ_1 y σ_2 son independientes, condicional sobre (τ_1, τ_2) .

4.3.3. Consideraciones de diseño.

La distribución a priori intrínseca (4.10) depende de la matriz de diseño \mathbf{Z}_i , $i = 1, 2$, que corresponde a la muestra de entrenamiento minimal $(\mathbf{z}_1, \mathbf{z}_2)$. Para determinar $(\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}$ podemos proceder usando la idea fundamental en el factor de Bayes intrínseco aritmético (FBIA); promediando sobre todas las submatrices observadas de orden $(k+1) \times k$. Esto es, dada la muestra de variables regresoras de tamaño n_i con

$$\mathbf{X}_i = \begin{pmatrix} 1 & x_{21}^i & \cdots & x_{k1}^i \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_{2n_i}^i & \cdots & x_{kn_i}^i \end{pmatrix},$$

se consideran todas las submatrices de $(k+1) \times k$. Se denotará esta clase de matrices como $\{\mathbf{Z}_i(l), l = 1, \dots, L_i\}$. Entonces, $(\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}$ se define como

$$(\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} = \frac{1}{L_i} \sum_{l=1}^{L_i} (\mathbf{Z}_i^t(l) \mathbf{Z}_i(l))^{-1}, \quad i = 1, 2.$$

4.3.4. El factor de Bayes para modelos de regresión heteroscedásticos.

El factor de Bayes para comparar el modelo M_2 vs M_1 con las distribuciones a priori intrínsecas $(\pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2), \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2))$ está dado por

$$B_{21}(\mathbf{y}_1, \mathbf{y}_2) = \frac{\int m_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta} d\tau_1 d\tau_2}{\int f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta} d\tau_1 d\tau_2}, \quad (4.11)$$

donde,

$$\begin{aligned} & m_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \\ &= \int f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 d\sigma_1 d\sigma_2, \end{aligned}$$

y

$$\pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) = \frac{c_1}{\tau_1^2 \tau_2^2}.$$

Nótese que $B_{21}(\mathbf{y}_1, \mathbf{y}_2)$ está bien definido, puesto que la constante arbitraria c_1 que aparece en $\pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2)$ se cancela.

Lema 3 *El factor de Bayes para probar la igualdad de los coeficientes de regresión, bajo heteroscedasticidad, es*

$$B_{21}(\mathbf{y}_1, \mathbf{y}_2) = \frac{I}{J}, \quad (4.12)$$

donde

$$I = \int_0^{\pi/2} \int_0^{\pi/2} \int_0^{\pi/2} \frac{N}{Q^{(n-k+2)}} d\varphi_1 d\varphi_2 d\varphi_3,$$

$n = n_1 + n_2$, y

$$J = \int_0^{\pi/2} \frac{\text{sen}^{n_1} \varphi \cdot \cos^{n_2} \varphi}{Q_1^{(n-k+2)} |\mathbf{P}_1|^{1/2}} d\varphi,$$

donde

$$N = \frac{\text{sen}\varphi_2 (\cos^2 \varphi_2 + \text{sen}^2 \varphi_2 \text{sen}^2 \varphi_3)^{-3/2}}{\text{sen}\varphi_1 (\cos^2 \varphi_1 + \text{sen}^2 \varphi_1 \text{sen}^2 \varphi_2 \cos^2 \varphi_3)^{3/2}},$$

$$P = |\mathbf{A}_1|^{1/2} |\mathbf{A}_2|^{1/2} |\mathbf{A}|^{1/2},$$

$$Q = \sum_{i=1}^2 \mathbf{y}_i^t \mathbf{A}_i^{-1} \mathbf{y}_i - \left(\sum_{i=1}^2 \mathbf{X}_i^t \mathbf{A}_i^{-1} \mathbf{y}_i \right)^t \mathbf{A}^{-1} \left(\sum_{i=1}^2 \mathbf{X}_i^t \mathbf{A}_i^{-1} \mathbf{y}_i \right),$$

$$\mathbf{A}_1 = \left(\mathbf{I}_{n_1} + \mathbf{X}_1 (\mathbf{Z}_1^t \mathbf{Z}_1)^{-1} \mathbf{X}_1^t \right) \text{sen}^2 \varphi_1 \text{sen}^2 \varphi_2 \cos^2 \varphi_3 + \mathbf{X}_1 (\mathbf{Z}_1^t \mathbf{Z}_1)^{-1} \mathbf{X}_1^t \cos^2 \varphi_1,$$

$$\mathbf{A}_2 = \text{sen}^2 \varphi_1 \left[\left(\mathbf{I}_{n_2} + \mathbf{X}_2 (\mathbf{Z}_2^t \mathbf{Z}_2)^{-1} \mathbf{X}_2^t \right) \text{sen}^2 \varphi_2 \text{sen}^2 \varphi_3 + \mathbf{X}_2 (\mathbf{Z}_2^t \mathbf{Z}_2)^{-1} \mathbf{X}_2^t \cos^2 \varphi_2 \right],$$

$$\mathbf{A} = \sum_{i=1}^2 \mathbf{X}_i^t \mathbf{A}_i^{-1} \mathbf{X}_i,$$

$$\mathbf{P}_1 = \mathbf{X}_1^t \mathbf{X}_1 \text{sen}^2 \varphi + \mathbf{X}_2^t \mathbf{X}_2 \cos^2 \varphi,$$

$$\mathbf{Q}_1 = \mathbf{y}_1^t \mathbf{y}_1 \text{sen}^2 \varphi + \mathbf{y}_2^t \mathbf{y}_2 \cos^2 \varphi - \mathbf{C}^t \mathbf{P}_1^{-1} \mathbf{C},$$

y

$$\mathbf{C} = \mathbf{X}_1^t \mathbf{y}_1 \text{sen}^2 \varphi + \mathbf{X}_2^t \mathbf{y}_2 \cos^2 \varphi.$$

Demostración.- El numerador de la ecuación (4.11) se reduce a una simple integración numérica después de implementar los dos pasos siguientes. Primero, se hace una integración con respecto a $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ y $\boldsymbol{\theta}$.

(i) Por un lado

$$\begin{aligned} & \int \mathcal{N}_{n_i}(\mathbf{y}_i | \mathbf{X}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) \mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) d\boldsymbol{\theta}_i \\ &= \int |\sigma_i^2 \mathbf{I}|^{-1/2} (2\pi)^{-n_i/2} \exp \left\{ -\frac{1}{2\sigma_i^2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta}_i)^t (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta}_i) \right\} \\ & \quad \times \left| (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right|^{-1/2} (2\pi)^{-k/2} \\ & \quad \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\theta})^t \left((\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right)^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta}) \right\} d\boldsymbol{\theta}_i \\ &= \frac{\sigma_i^{-n_i} (2\pi)^{-n_i/2} |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2}}{(\sigma_i^2 + \tau_i^2)^{k/2} (2\pi)^{k/2}} \\ & \quad \times \int \exp \left\{ -\frac{1}{2\sigma_i^2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta}_i)^t (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\theta}_i) \right\} \\ & \quad \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\theta})^t \left((\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \right)^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta}) \right\} d\boldsymbol{\theta}_i. \end{aligned}$$

Análogamente a cálculo anteriores, tenemos que desarrollando y completando cuadrados se llega a que

$$\begin{aligned} & \int \mathcal{N}_{n_i}(\mathbf{y}_i | \mathbf{X}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) \mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma_i^2 + \tau_i^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) d\boldsymbol{\theta}_i \\ &= \mathcal{N}_{n_i}(\mathbf{y}_i | \mathbf{X}_i \boldsymbol{\theta}, \sigma_i^2 \mathbf{I} + (\sigma_i^2 + \tau_i^2) \mathbf{X}_i (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \mathbf{X}_i^t). \end{aligned}$$

(ii) Por otro lado

$$\begin{aligned} & \int \left\{ \prod_{i=1}^2 \mathcal{N}_{n_i}(\mathbf{y}_i | \mathbf{X}_i \boldsymbol{\theta}, \mathbf{A}_i) \right\} d\boldsymbol{\theta} \\ &= |\mathbf{A}_1|^{-1/2} |\mathbf{A}_2|^{-1/2} (2\pi)^{-(n_1+n_2)/2} \\ & \quad \times \int \exp \left\{ -\frac{1}{2} (\mathbf{y}_1 - \mathbf{X}_1 \boldsymbol{\theta})^t (\mathbf{A}_1)^{-1} (\mathbf{y}_1 - \mathbf{X}_1 \boldsymbol{\theta}) - \frac{1}{2} (\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\theta})^t (\mathbf{A}_2)^{-1} (\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\theta}) \right\} d\boldsymbol{\theta}, \end{aligned}$$

este cálculo es análogo a la integración con respecto a $\boldsymbol{\theta}$, en el cálculo de la distribución marginal $m_1(\mathbf{z}_1, \mathbf{z}_2)$, por lo tanto,

$$\begin{aligned} & \int \left\{ \prod_{i=1}^2 \mathcal{N}_{n_i}(\mathbf{y}_i | \mathbf{X}_i \boldsymbol{\theta}, \mathbf{A}_i) \right\} d\boldsymbol{\theta} \\ &= \frac{\exp \left\{ -\frac{1}{2} \sum_{i=1}^2 \mathbf{y}_i^t \mathbf{A}_i^{-1} \mathbf{y}_i + \frac{1}{2} \left(\sum_{i=1}^2 \mathbf{X}_i^t \mathbf{A}_i^{-1} \mathbf{y}_i \right)^t \mathbf{A}^{-1} \left(\sum_{i=1}^2 \mathbf{X}_i^t \mathbf{A}_i^{-1} \mathbf{y}_i \right) \right\}}{(2\pi)^{(n-k)/2} |\mathbf{A}_1|^{1/2} |\mathbf{A}_2|^{1/2} |\mathbf{A}|^{1/2}}, \end{aligned}$$

donde

$$A = \sum_{i=1}^2 \mathbf{X}_i^t \mathbf{A}_i^{-1} \mathbf{y}_i.$$

Segundo, transformando las variables $\tau_1, \tau_2, \sigma_1, \sigma_2$ a coordenadas polares se obtiene

$$\begin{aligned} & \int \int \int m_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta} d\tau_1 d\tau_2 \\ &= \int \int \int f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \\ & \quad \times \pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta} d\tau_1 d\tau_2 \\ &= \frac{c_1 \Gamma\left(\frac{n-k+2}{2}\right)}{2(2\pi)^{(n-k)/2}} \mathbf{I}, \end{aligned}$$

donde $n = n_1 + n_2$

El denominador de (4.11) se reduce a una simple integración numérica. En efecto, una integración analítica con respecto a $\boldsymbol{\theta}$ usando (ii) transformando las variables τ_1, τ_2 a coordenadas polares, y usando un procedimiento análogo al cálculo de $m_1(\mathbf{z}_1, \mathbf{z}_2)$, se obtiene

$$\int f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau_1, \tau_2) \pi_1^N(\boldsymbol{\theta}, \tau_1, \tau_2) d\boldsymbol{\theta} d\tau_1 d\tau_2 = \frac{c_1 \Gamma\left(\frac{n-k+2}{2}\right)}{2(2\pi)^{(n-k)/2}} \mathbf{J}.$$

Las integrales \mathbf{I} y \mathbf{J} no pueden calcularse analíticamente y por lo tanto se necesita de integración numérica.

4.4. Comparaciones

El factor de Bayes de la sección 4.2 se dedujo sin imponer alguna condición sobre las varianzas de los modelos. Cuando la homoscedasticidad puede aceptarse, la dimensión de los espacios paramétricos se reduce. Por otra parte, como el contraste de la sección 4.2 puede verse como un contraste para la igualdad de regresiones bajo el supuesto de heteroscedasticidad, se pueden combinar los resultados de las dos secciones para obtener una prueba incondicional para la igualdad de regresiones.

4.4.1. Prueba de homoscedasticidad.

Probar la condición de homoscedasticidad involucra la comparación entre los modelos

$$(\mathbf{y}_1, \mathbf{y}_2) \sim f_1(\mathbf{y}_1, \mathbf{y}_2 | \gamma_1, \gamma_2, \tau) = \mathcal{N}_{n_1}(\mathbf{X}_1 \gamma_1, \tau^2 \mathbf{I}) \times \mathcal{N}_{n_2}(\mathbf{X}_2 \gamma_2, \tau^2 \mathbf{I}), \quad (4.13)$$

y

$$(\mathbf{y}_1, \mathbf{y}_2) \sim f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma) = \mathcal{N}_{n_1}(\mathbf{X}_1 \boldsymbol{\theta}_1, \sigma_1^2 \mathbf{I}) \times \mathcal{N}_{n_2}(\mathbf{X}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}).$$

Bajo el punto de vista bayesiano por defecto los modelos a comparar son

$$M^{HO} : \left\{ f_1(\mathbf{y}_1, \mathbf{y}_2 | \gamma_1, \gamma_2, \tau), \quad \pi_1^N(\gamma_1, \gamma_2, \tau) = \frac{c_1}{\tau^2} \right\}, \quad (4.14)$$

y

$$M^{HE} : \left\{ f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2), \quad \pi_2^N(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = \frac{c_2}{\sigma_1^2 \sigma_2^2} \right\}. \quad (4.15)$$

La distribución a priori intrínseca para $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2)$ está dada en el siguiente lema.

Lema 4 La distribución a priori intrínseca para comparar los modelos (4.14) y (4.15) es

$$\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = \int \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \gamma_1, \gamma_2, \tau) \pi_1^N(\gamma_1, \gamma_2, \tau) d\gamma_1 d\gamma_2 d\tau,$$

donde

$$\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \gamma_1, \gamma_2, \tau) = \prod_{i=1}^2 \phi_k\left(\boldsymbol{\theta}_i | \gamma_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}\right) \times \frac{1}{\tau (1 + \sigma_i^2 / \tau^2)^{3/2}},$$

donde ϕ_k es la densidad de una distribución $\mathcal{N}\left(\gamma_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}\right)$.

Demostración.-

Puesto que se quiere la distribución a priori condicional sobre $(\gamma_1, \gamma_2, \tau)$, entonces aplicamos la ecuación (4.9) a los modelos

$$M_1^{**} : \mathcal{N}_{k+1}(\mathbf{z}_1 | \mathbf{Z}_1 \gamma_1, \tau^2 \mathbf{I}) \cdot \mathcal{N}_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \gamma_2, \tau^2 \mathbf{I}), \quad \pi_1^N(\gamma_1, \gamma_2, \tau) = \delta(\gamma_1, \gamma_2, \tau),$$

y

$$M_2^{**} : \mathcal{N}_{k+1}(\mathbf{z}_1 | \mathbf{Z}_1 \boldsymbol{\theta}_1, \sigma_1^2 \mathbf{I}) \cdot \mathcal{N}_{k+1}(\mathbf{z}_2 | \mathbf{Z}_2 \boldsymbol{\theta}_2, \sigma_2^2 \mathbf{I}), \quad \pi_2^N(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = \frac{c}{\sigma_1^2 \sigma_2^2},$$

donde $\delta(\gamma_1, \gamma_2, \tau)$ es la delta de Dirac en el punto $(\gamma_1, \gamma_2, \tau)$.

Por definición

$$\begin{aligned} & B_{12}^N(\mathbf{z}_1, \mathbf{z}_2) \\ &= \frac{m_1^{**}(\mathbf{z}_1, \mathbf{z}_2)}{m_2^{**}(\mathbf{z}_1, \mathbf{z}_2)} = \frac{\prod_{i=1}^2 \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \gamma_i, \tau^2 \mathbf{I})}{m_2(\mathbf{z}_1, \mathbf{z}_2)} \\ &= \frac{\prod_{i=1}^2 \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \gamma_i, \tau^2 \mathbf{I})}{\frac{c}{(2\pi)^2 \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i}} = \frac{2\pi}{c} \prod_{i=1}^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \gamma_i, \tau^2 \mathbf{I}), \end{aligned}$$

donde

$$\mathbf{H}_i = \mathbf{I} - \mathbf{Z}_i (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1} \mathbf{Z}_i^t,$$

por (3.38), teníamos que

$$\begin{aligned}
& \pi^I(\boldsymbol{\theta}_2) \\
&= \pi_2^N(\boldsymbol{\theta}_2) E_{x^{(l)}|\boldsymbol{\theta}_2}^{M_2} [B_{12}^N(\mathbf{z}_1, \mathbf{z}_2)] \\
&= \frac{c}{\sigma_1^2 \sigma_2^2} \iint \frac{2\pi}{c} \prod_{i=1}^2 |\mathbf{z}_i^t \mathbf{z}_i|^{1/2} \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\gamma}_i, \tau^2 \mathbf{I}) \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) d\mathbf{z}_1 d\mathbf{z}_2 \\
&= \frac{2\pi |\mathbf{Z}_1^t \mathbf{Z}_1|^{1/2} |\mathbf{Z}_2^t \mathbf{Z}_2|^{1/2}}{\sigma_1^2 \sigma_2^2} \prod_{i=1}^2 \int \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\gamma}_i, \tau^2 \mathbf{I}) \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) d\mathbf{z}_i.
\end{aligned}$$

Análogamente al resultado del lema 2 tenemos que, desarrollando exponenciales y completando cuadrados

$$\begin{aligned}
& \int \mathbf{z}_i^t \mathbf{H}_i \mathbf{z}_i \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\gamma}_i, \tau^2 \mathbf{I}) \mathcal{N}_{k+1}(\mathbf{z}_i | \mathbf{Z}_i \boldsymbol{\theta}_i, \sigma_i^2 \mathbf{I}) d\mathbf{z}_i \\
&= \frac{1}{(2\pi)^{1/2}} \mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\gamma}_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) \frac{\sigma_i^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{-1/2}}{\tau (1 + \sigma_i^2/\tau^2)^{3/2}}.
\end{aligned}$$

Sustituyendo,

$$\begin{aligned}
& \pi^I(\boldsymbol{\theta}_2) \\
&= \frac{2\pi |\mathbf{Z}_1^t \mathbf{Z}_1|^{1/2} |\mathbf{Z}_2^t \mathbf{Z}_2|^{1/2}}{\sigma_1^2 \sigma_2^2} \prod_{i=1}^2 \frac{1}{(2\pi)^{1/2}} \mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\gamma}_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) \\
& \quad \times \frac{\sigma_i^2 |\mathbf{Z}_i^t \mathbf{Z}_i|^{-1/2}}{\tau (1 + \sigma_i^2/\tau^2)^{3/2}} \\
&= \prod_{i=1}^2 \mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\gamma}_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) \times \frac{1}{\tau} (1 + \sigma_i^2/\tau^2)^{-3/2},
\end{aligned}$$

por lo tanto

$$\begin{aligned}
& \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \tau) \\
&= \prod_{i=1}^2 \left[\mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\gamma}_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) \times \frac{1}{\tau} (1 + \sigma_i^2/\tau^2)^{-3/2} \right],
\end{aligned}$$

lo que demuestra el lema anterior.

Nota 5 $\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \tau)$ es una densidad de probabilidad condicional sobre un punto arbitrario $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \tau)$. En esta distribución a priori, $\boldsymbol{\theta}_1$ es independiente de $\boldsymbol{\theta}_2$ condicional sobre $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \sigma_1, \sigma_2, \tau)$, y σ_1 es independiente de σ_2 condicional sobre τ .

La distribución a priori intrínseca incondicional para los parámetros $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2$ es

$$\begin{aligned}
& \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) \tag{4.16} \\
&= \iiint \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2 | \gamma_1, \gamma_2, \tau) \frac{c_1}{\tau^2} d\gamma_1 d\gamma_2 d\tau \\
&= \iiint \prod_{i=1}^2 \mathcal{N}_k(\boldsymbol{\theta}_i | \gamma_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) \times \frac{1}{\tau} (1 + \sigma_i^2/\tau^2)^{-3/2} \frac{c_1}{\tau^2} d\gamma_1 d\gamma_2 d\tau \\
&= c_1 \int \prod_{i=1}^2 \int \mathcal{N}_k(\boldsymbol{\theta}_i | \gamma_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) d\gamma_i \frac{1}{\tau^3} (1 + \sigma_i^2/\tau^2)^{-3/2} d\tau,
\end{aligned}$$

pero

$$\int \mathcal{N}_k(\boldsymbol{\theta}_i | \gamma_i, (\sigma_i^2 + \tau^2) (\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}) d\gamma_i = 1.$$

Por lo tanto, se tiene que

$$\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2) = c_1 \int \left[\prod_{i=1}^2 \frac{1}{\tau^3} (1 + \sigma_i^2/\tau^2)^{-3/2} \right] d\tau,$$

que es una densidad impropia pero,

- (i) Es límite único para la distribución a priori intrínseca propia.
- (ii) No depende de la matriz de diseño \mathbf{Z}_i .
- (iii) El factor de Bayes para $(\pi_1^N(\gamma_1, \gamma_2, \tau), \pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2))$ está bien definido, ya que la constante arbitraria c_1 que aparece en $\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2)$ es exactamente la misma que la que aparece en $\pi_1^N(\gamma_1, \gamma_2, \tau)$.

Lema 5 *El factor de Bayes para comparar los modelos de prueba (4.13) con las distribuciones a priori intrínsecas $\pi_1^N(\gamma_1, \gamma_2, \tau)$ y $\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1, \sigma_2)$ está dado por*

$$\tilde{B}_{21}(\mathbf{y}_1, \mathbf{y}_2) = D^{(n-2k+1)/2} \int_0^{\pi/2} \int_0^{\pi/2} \frac{N_2}{Q_2^{(n-2k+1)/2}} d\varphi_1 d\varphi_2, \tag{4.17}$$

donde $\mathbf{D} = \sum_{i=1}^2 \mathbf{y}_i^t \mathbf{H}_i \mathbf{y}_i$, $\mathbf{H}_i = \mathbf{I}_{n_i} - \mathbf{X}_i (\mathbf{X}_i^t \mathbf{X}_i)^{-1} \mathbf{X}_i^t$,

$$N_2 = \frac{\text{sen}^2 \varphi_1 \cos^2 \varphi_1 \text{sen}^{n_1-k+1} \varphi_2 \cos^{n_2-k+1} \varphi_2}{(\cos^2 \varphi_1 + \text{sen}^2 \varphi_1 \text{sen}^2 \varphi_2)^{3/2} (\cos^2 \varphi_1 + \text{sen}^2 \varphi_1 \cos^2 \varphi_2)^{3/2}},$$

y

$$\mathbf{Q}_2 = \mathbf{y}_i^t \mathbf{H}_i \mathbf{y}_i \text{sen}^2 \varphi_2 + \mathbf{y}_i^t \mathbf{H}_i \mathbf{y}_i \cos^2 \varphi_2.$$

Asumiendo a priori que $P(M^{HO}) = P(M^{HE}) = \frac{1}{2}$, la probabilidad a posteriori, de que la condición de homoscedasticidad se cumpla, está dado por

$$P(M^{HO}|\text{datos}) = \frac{1}{1 + \tilde{B}_{21}(\mathbf{y}_1, \mathbf{y}_2)}.$$

4.4.2. Prueba de igualdad de regresiones bajo homoscedasticidad.

Cuando $P(M^{HO}|\text{datos})$ no es pequeño, puede ser de interés probar la igualdad de los coeficientes de regresión, esto es, probar H_0 vs H_1 , bajo la condición de homoscedasticidad. En este caso el problema es comparar los modelos

$$M_1^{HO} : \left\{ f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau), \quad \pi_1^N(\boldsymbol{\theta}, \tau) = \frac{c_1}{\tau^2} \right\}, \quad (4.18)$$

y

$$M_2^{HO} : \left\{ f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma), \quad \pi_1^N(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma) = \frac{c_2}{\sigma^2} \right\}, \quad (4.19)$$

donde

$$f_1(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}, \tau) = \mathcal{N}_{n_1}(\mathbf{X}_1 \boldsymbol{\theta}, \tau^2 \mathbf{I}) \times \mathcal{N}_{n_2}(\mathbf{X}_2 \boldsymbol{\theta}, \tau^2 \mathbf{I}),$$

y

$$f_2(\mathbf{y}_1, \mathbf{y}_2 | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma) = \mathcal{N}_{n_1}(\mathbf{X}_1 \boldsymbol{\theta}_1, \sigma^2 \mathbf{I}) \times \mathcal{N}_{n_2}(\mathbf{X}_2 \boldsymbol{\theta}_2, \sigma^2 \mathbf{I}).$$

La distribución a priori intrínseca de $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma)$ condicional sobre $(\boldsymbol{\theta}, \tau)$ es

$$\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma | \boldsymbol{\theta}, \tau) = \frac{1}{\tau(1 + \sigma^2/\tau^2)^{3/2}} \prod_{i=1}^2 \mathcal{N}_k(\boldsymbol{\theta}_i | \boldsymbol{\theta}, (\sigma^2 + \tau^2)(\mathbf{Z}_i^t \mathbf{Z}_i)^{-1}), \quad (4.20)$$

donde la matriz de diseño \mathbf{Z}_1 es de orden $(k+1) \times k$ y \mathbf{Z}_2 es de orden $k \times k$. La diferencia del orden de estas matrices se debe al hecho de que bajo la condición de homoscedasticidad, la muestra de entrenamiento minimal \mathbf{z}_1 y \mathbf{z}_2 tienen tamaños $k+1$ y k respectivamente.

El factor de Bayes para las distribuciones a priori intrínsecas $\pi^I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma | \boldsymbol{\theta}, \tau)$ y $\pi_1^N(\boldsymbol{\theta}, \tau)$ es

$$B_{12}^*(\mathbf{y}_1, \mathbf{y}_2) = |\mathbf{X}^t \mathbf{X}|^{1/2} (\mathbf{y}^t \mathbf{H} \mathbf{y})^{(n-k+1)/2} \int_0^{\pi/2} \frac{M^{-(n-k+1)/2}}{|\mathbf{A}|^{1/2} \prod_{i=1}^2 |\mathbf{A}_i|^{1/2}} d\varphi, \quad (4.21)$$

donde

$$\mathbf{H} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix},$$

$$\mathbf{A} = \sum_{i=1}^2 \mathbf{X}_i' \mathbf{A}_i^{-1} \mathbf{X}_i = \mathbf{I}_{n_i} \cos^2 \varphi + \mathbf{X}_i (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{X}_i,$$

y

$$M = \sum_{i=1}^2 \mathbf{y}_i' \mathbf{A}_i^{-1} \mathbf{y}_i - \left(\sum_{i=1}^2 \mathbf{X}_i' \mathbf{A}_i^{-1} \mathbf{y}_i \right)' \mathbf{A}^{-1} \left(\sum_{i=1}^2 \mathbf{X}_i' \mathbf{A}_i^{-1} \mathbf{y}_i \right).$$

4.4.3. Una prueba incondicional

En una situación general, probar la hipótesis nula de que los coeficientes de regresión en dos modelos lineales son iguales se basa en la probabilidad a posteriori de la hipótesis nula. Esta probabilidad se obtiene condicionando sobre la homoscedasticidad y sobre la heteroscedasticidad, esto es

$$P(H_0|\text{datos}) = P(H_0|\text{datos}, HE) P(HE|\text{datos}) + P(H_0|\text{datos}, HO) P(HO|\text{datos}).$$

De los resultados de las secciones anteriores tenemos lo siguiente

$$P(H_0|\text{datos}, HE) = \frac{1}{1 + B_{21}},$$

donde B_{21} está dado en la expresión (4.12),

$$P(HO|\text{datos}) = 1 - P(HE|\text{datos}) = \frac{1}{1 + \tilde{B}_{21}},$$

donde \tilde{B}_{21} está dado en la expresión (4.17),

$$P(H_0|\text{datos}, HO) = \frac{1}{1 + B_{21}^*},$$

donde B_{21}^* está dado en la expresión (4.21).

4.5. Ejemplos.

Comparación con el estadístico de Chow modificado.

Como un primer ejemplo, se hace una pequeña simulación y se compara la probabilidad a posteriori de la hipótesis nula con el p -value frecuentista. Específicamente, se contrastan los modelos (4.1) y (4.2) usando la probabilidad a posteriori de H_0 (4.8), el factor de Bayes (4.12) y también el estadístico de Chow modificado propuesto por Conerly y Mansfield (1988).

Considérense los modelos

$$M_1 : y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \tau_i^2), i = 1, 2,$$

y

$$M_2 : y_i = a_i + b_i x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), i = 1, 2,$$

bajo cuatro casos:

Tabla 1 : Configuraciones paramétricas para las simulaciones

	Homoscedasticidad	Heteroscedasticidad
M_1 verdadero	$a = 1, b = 2$ $\tau_1 = \tau_2 = 1$	$a = 1, b = 2$ $\tau_1 = 1, \tau_2 = 2$
M_2 verdadero	(a) $a_1 = 1, b_1 = 2;$ $a_2 = 2, b_2 = 2$ $\sigma_1 = \sigma_2 = 1$	(b) $a_1 = 1, b_1 = 2;$ $a_2 = 2, b_2 = 2$ $\sigma_1 = 1, \sigma_2 = 3$
	(c) $a_1 = 1, b_1 = 2;$ $a_2 = 1, b_2 = 3$ $\sigma_1 = \sigma_2 = 1$	(d) $a_1 = 1, b_1 = 2;$ $a_2 = 1, b_2 = 3$ $\sigma_1 = 1, \sigma_2 = 3$

Para cada situación se generan valores de x_i uniformes en el rango $(0, 5)$, y se generan las y_i de la distribución normal apropiada, con $n_1 = 9$ observaciones tomadas de M_1 y $n_2 = 11$ observaciones tomadas de M_2 . Este proceso se generó 100 veces.

Generamos 100 datos de una distribución normal con media $1 + 2x$ (con x ya prefijado) y varianza 1, obtenemos datos sobre la vertical de $1 + 2x$, centrados en $1 + 2x$ y con varianza 1.

Si movemos x entre 0 y 5, obtenemos que la nube de puntos se mueve, pero manteniendo la varianza. Esto es lo que se muestra en el primer caso. Así para todos y cada uno de los cuatro casos.

Los resultados se presentan en la siguiente tabla (los errores estándares están entre paréntesis).

Tabla2 : Probabilidad a posteriori de H_0 basada en el factor de Bayes (4.12), y

los p – values del estadístico de Chow modificado (los errores estándares están entre paréntesis)

	<i>Homoscedasticidad</i>		<i>Heteroscedasticidad</i>	
	Prob. a pos.	Chow Mod.	Prob. a pos.	Chow Mod.
M_1 verdadero	.971 (,004)	.470 (.042)	.962 (,005)	.426 (.042)
M_2 verdadero	(a) .849 (,042)	.176 (.029)	(b) .849 (,062)	.148 (.026)
	(c) .066 (,026)	.004 (.001)	(d) .592 (,059)	.049 (.009)

Las gráficas dadas en la figura 1 (apéndice A4.1) demuestran una realización de los datos de cada uno de las cuatro configuraciones de los parámetros (a)-(d), los cuales corresponden a cuatro casos, en los cuales el modelo nulo es falso.

La posición (1,1) de la tabla 2, significa que se ha simulado de un modelo homoscedástico con varianza 1, y que es una sola línea de regresión con $a = 1$, y $b = 2$. Para cada una de las 100 simulaciones se ha calculado el valor de probabilidad a posteriori del modelo verdadero (es decir, del modelo del que provienen las simulaciones). Ahora, el valor medio de esas probabilidades a posteriori es 0,971 y el error estándar 0,004. Análogamente se hizo el cálculo para las otras tres configuraciones de los parámetros. El resultado es excelente pues con esa media habrá pocos valores menores a 0,5 (en cuyo caso se rechazaría el modelos del que provienen las simulaciones). Sin embargo en el estadístico de Chow los p – values son altos pero no tan altos como cabría esperar (uno espera p – values cercanos a unos).

En cada caso la probabilidad a posteriori del modelo nulo es más grande que el p-value del estadístico de Chow modificado, demostrando que el estadístico de Chow modificado tiene una gran tendencia a rechazar el modelo de la hipótesis nula (una tendencia que se observa en muchos procedimientos de selección de modelos). Esto es particularmente evidente en los casos (a) y (b), donde la probabilidad a posteriori de la hipótesis nula es bastante grande, y el p-value, aunque es “no-significativo” tiende a favorecer a la hipótesis alternativa.

Ejemplo 7 Abundancia de especie animal seleccionada

MacPherson (1990) describe un estudio que compara dos especies de alga marina con diferentes características morfológicas. Para cada especie de alga marina, se han medido la relación entre su biomasa (el peso seco) y la abundancia de especies de

animales que usaron la planta como un huésped. Los datos se muestran en la siguiente tabla

Tabla 3 Biomasa (peso seco) y abundancia de especies de animales, sobre dos plantas de algas marinas (MacPherson, (1990)).

<i>Caulocystis</i>	<i>cephalornothos (C)</i>	<i>Sargassum</i>	<i>verruculosum (S)</i>
Abundancia	Peso seco	Abundancia	Peso seco
681	28.8	67	2.5
320	12.1	222	12.5
244	7.5	131	8.6
219	6.5	225	10.8
422	9.9	71	3.7
441	12.7	391	12.0
347	7.6	142	4.5
176	8.5	123	4.9
229	7.3	186	8.0
363	14.1	239	9.7
310	11.0	294	4.1
161	9.6	256	6.7
337	12.0	435	11.2
331	9.2	460	10.5
337	21.5	222	4.8
825	30.4	572	10.1
707	24.4	224	6.4
514	11.1	344	4.6
901	28.5	661	14.3
957	26.0	439	14.2
		277	8.7
		129	3.6
		430	8.6
		338	4.5
		530	11.5

para cada especie de alga marina, se hace una regresión del $\log(\text{abundancia})$ sobre la biomasa, y la pregunta de interés es si la relación es la misma en las dos especies.

Un análisis de covarianzas, ajustando una regresión común a ambos modelos, considerando una hipótesis con pendiente común y otra hipótesis con término independiente común y haciendo una regresión por separado para cada modelo, producen un p -value de 0,0477 para la hipótesis del término independiente común y 0,0153 para la hipótesis de pendientes común, lo cual nos lleva a concluir que la respuesta de especies de animales es diferente. Sin embargo, existe evidencia de heteroscedasticidades en la gráfica de la figura 2, (apéndice A4.2) y en los errores residuales estándar de las regresiones individuales (0,4593 para la especie S y 0,2895 para la especie C). Por lo tanto, se

procede con la prueba de igualdad de regresiones bajo heteroscedasticidad, y se obtiene un p-value del estadístico de Chow modificado de 0,634 y una probabilidad a posteriori de H_0 de 0,997 usando el factor de Bayes intrínseco. Por lo tanto, una vez considerada la heterocedasticidad, la evidencia para la respuesta de especies de animales similares es bastante fuerte.

Conclusiones

Una de las conclusiones es que podemos apreciar que los factores de Bayes intrínsecos son completamente automáticos en el sentido de que sólo se basan en los datos y en distribuciones a priori no informativas estándares. Sin embargo, aun existe el problema de la elección óptima de la muestra de entrenamiento minimal.

Otra de las ventajas que se tiene es que los factores de Bayes intrínsecos pueden usarse para contrastes de hipótesis bayesianas estándar.

Puede ser usado para modelos anidados o no anidados y para comparación múltiple de modelos y predicción. Son invariantes para transformaciones univariadas de los datos. Su comportamiento asintótico como un verdadero factor de Bayes para distribuciones a priori intrínsecas razonables, lo hace preferible a otras aproximaciones, tal como el método de Jeffrey.

Ahora, sobre las desventajas de los factores de Bayes intrínsecos es que computacionalmente es muy costoso, ya que su cálculo puede ser bastante intensivo, pueden no ser estable para un tamaño de muestra pequeño y también que no son invariantes para transformaciones multivariadas de los datos.

Otra de las cosas, es que podemos apreciar que el problema de igualdad de coeficientes en dos modelos de regresión lineal normal puede ser tratado como un problema de dos muestras para modelos anidados. Esto nos permite resolver el problema de calibrado que surge cuando se usan las distribuciones a priori convencionales estándares para los parámetros. En un ambiente homoscedástico y heteroscedástico se obtienen distribuciones a priori calibradas con las cuales se pueden calcular los factores de Bayes y las probabilidades a posteriori de los modelos. El procedimiento de comparación de modelos presentado se basa en estas probabilidades a posteriori.

También se ha visto que no hay ningún problema numérico especial en el cálculo de las probabilidades a posteriori de los modelos. Las simulaciones muestran que el seleccionador de modelos basado en las distribuciones a priori intrínsecas tiende a modelos más parsimoniosos que el estadístico de Chow modificado, requiriendo una fuerte evidencia antes de rechazar la hipótesis nula. Además, se ha visto que al considerar la

heteroscedasticidad pueden derrumbarse fácilmente los resultados de un análisis en los que no se considera.

Notación

μ	Media poblacional
σ^2	Varianza poblacional
α	Nivel de significancia
$\pi(\theta)$	Distribución a priori del parámetro θ
$f(x \theta)$	Función de densidad de los datos
Θ	Espacio paramétrico
$I(\theta)$	Matriz de información esperada de Fisher
$\pi(\theta x)$	Distribución a posteriori del parámetro θ
$h(x, \theta)$	Distribución conjunta de x y θ
$\mathcal{N}(\theta, \sigma^2)$	Distribución normal con media θ y varianza σ^2
$\mathcal{N}(\mu, \tau^2)$	Distribución normal con media μ y varianza τ^2
$\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$	Distribución normal multivariada con media $\mathbf{0}$ y varianza $\sigma^2\mathbf{I}$
s^2	Varianza muestral
t	Distribución t de Student
F	Estadístico de prueba
$F_{k, n-2k}$	Distribución F con k , y $n - 2k$ grados de libertad
Γ	Función gamma
Υ	Clase de todas las distribuciones de probabilidad sobre $(0, \infty)$.
$\pi_i^N(\theta)$	Distribución a priori impropia de θ
$\pi_i^I(\theta_i)$	Distribución a priori intrínseca para θ_i
$x(l)$	Muestra de entrenamiento minimal
$m_i(\mathbf{y}_1, \mathbf{y}_2)$	Distribución marginal i de $(\mathbf{y}_1, \mathbf{y}_2)$
\mathbf{X}_i	Matriz de muestras regresoras de tamaño n_i
$\det \hat{I}$	Determinante de la matriz observada de Fisher
$\Phi_p(\mathbf{y} \mu, \Sigma)$	Densidad de la distribución normal p-variada con media μ y matriz de covarianza Σ

$P(M_i y_1, y_2)$	Probabilidad a posteriori del modelo M_i
$E_{x^{(l)} \theta_2}^{M_2}$	Esperanza con respecto a la densidad de la muestra de entrenamiento minimal $x^{(l)}$ bajo el modelo M_2
$\Phi_{k+1}(\mathbf{z}_i \mathbf{Z}_i\theta, \tau_i^2 I)$	Densidad de la distribución normal multivariada con media $\mathbf{Z}_i\theta$ y varianza $\tau_i^2 I$

Apéndice

A2.1

Lema 6 Sea \mathbf{t} un estadístico suficiente para θ , con distribución conjunta $g(\mathbf{t}|\theta)$. Entonces,

$$l(\theta|\mathbf{x}) \propto l_1(\theta|\mathbf{t}) \quad \text{donde} \quad l_1(\theta|\mathbf{t}) \propto g(\mathbf{t}|\theta).$$

En otras palabras, la función de verosimilitud obtenida de la distribución de \mathbf{t} es la misma que la obtenida de la distribución de \mathbf{x} .

Demostración. Por el teorema de factorización (estadístico suficiente) sabemos que

$$f(\mathbf{x}|\theta) = g(\mathbf{t}|\theta)m(\mathbf{x}).$$

La distribución de probabilidad $f(x|\theta)$ puede ser considerada como una función de θ y no de \mathbf{x} , Fisher (1922) le llamó función de verosimilitud de θ dado \mathbf{x} , y la designó por $l(\theta|\mathbf{x})$, donde

$$l(\theta|\mathbf{x}) = g(\mathbf{t}|\theta)m(\mathbf{x}).$$

Esto significa que la función de verosimilitud es proporcional a $g(\mathbf{t}|\theta)$, ya que $m(\mathbf{x})$ puede verse solamente como una constante de proporcionalidad, es decir

$$l(\theta|\mathbf{x}) \propto g(\mathbf{t}|\theta),$$

que prueba el lema.

Ejemplos del lema anterior son los siguientes.

Ejemplo 8 Distribución normal, varianza conocida. Supóngase que \mathbf{y} es un vector aleatorio de una distribución normal $\mathcal{N}(\theta, \sigma^2)$, donde σ se supone conocida. Entonces, la función de verosimilitud de θ es

$$l(\theta|\sigma, \mathbf{y}) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right]. \quad (1)$$

Por otro lado, la media muestral \bar{y} se distribuye como

$$f(\bar{y}|\theta, \sigma^2) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n}{2\sigma^2}(\bar{y} - \theta)^2\right], \quad -\infty < \bar{y} < \infty,$$

así que, dado \bar{y} ,

$$l_1(\theta|\sigma, \bar{y}) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right], \quad (2)$$

de aquí se tiene que

$$l(\theta|\sigma, \mathbf{y}) \propto l_1(\theta|\sigma, \bar{y}).$$

Ejemplo 9 *Distribución normal, media y varianza desconocidas. En este caso, la función de verosimilitud $l(\theta, \sigma|\mathbf{y})$ basada en las n observaciones independientes, \mathbf{y} , está dada por*

$$l(\theta, \sigma|\mathbf{y}) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \theta)^2]\right\}. \quad (3)$$

Ahora, es bien conocido que: a) $\bar{y} \sim \mathcal{N}(\theta, \sigma^2/n)$, b) $(n-1)s^2 = \sum_{u=1}^n (y_u - \bar{y})^2 \sim \sigma^2 \chi_{n-1}^2$, y c) \bar{y} y s^2 son estadísticamente independientes. Así,

$$\pi(\bar{y}, s^2|\theta, \sigma) = \pi(\bar{y}|\theta, \sigma^2) \pi(s^2|\sigma^2),$$

donde

$$\pi(\bar{y}|\theta, \sigma^2) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right], \quad -\infty < \bar{y} < \infty,$$

y

$$\pi(s^2|\sigma^2) = \frac{1}{\Gamma\left[\frac{1}{2}(n-1)\right]} \left(\frac{n-1}{2\sigma^2}\right)^{\frac{1}{2}(n-1)} (s^2)^{\frac{1}{2}(n-1)-1} \times \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right],$$

se sigue que, dado (\bar{y}, s^2) ,

$$l_1(\theta, \sigma|\bar{y}, s^2) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\theta - \bar{y})^2]\right\}. \quad (4)$$

De aquí se tiene que

$$l(\theta, \sigma|\mathbf{y}) \propto l_1(\theta, \sigma|\bar{y}, s^2).$$

A3.1

Para $a > 0, p > 0$

$$\int_0^\infty x^{p-1} e^{-ax} dx = a^{-p} \Gamma(p).$$

A3.2

Para $a > 0, p > 0$

$$\int_0^\infty x^{-(p+1)} e^{-ax^{-1}} dx = a^{-p} \Gamma(p).$$

A3.3

Para $a > 0, p > 0$

$$\int_0^{\infty} x^{p-1} e^{-ax^2} dx = \frac{1}{2} a^{-p/2} \Gamma(p/2).$$

A3.4

Para $a > 0, p > 0$

$$\int_0^{\infty} x^{-(p+1)} e^{-ax^{-2}} dx = \frac{1}{2} a^{-p/2} \Gamma(p/2).$$

A4.1

Código en R para generar la gráfica del ejemplo 1.

```
#Inciso a
for (j in 1:50) {
e1_rnorm(1,0,1)
for (i in 1:9) {
x1_runif(1,min=0,max=5)
y1_1+2*x1+e1
ab_cbind(x1,y1) # array bidimensional combinando las columnas
write.table(ab, file="c:\\proba.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
}
}
e2_rnorm(1,0,1)
for (i in 1:11) {
x2_runif(1,min=0,max=5)
y2_2+2*x2+e2
ab_cbind(x2,y2) # array bidimensional combinando las columnas
write.table(ab, file="c:\\proba1.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
}
}
# Inciso b
for (j in 1:50) {
e1_rnorm(1,0,1)
for (i in 1:9) {
x1_runif(1,min=0,max=5)
y1_1+2*x1+e1
ab_cbind(x1,y1) # array bidimensional combinando las columnas
write.table(ab, file="c:\\probb.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
```

```

}
e2_rnorm(1,0,3)
for (i in 1:11) {
x2_runif(1,min=0,max=5)
y2_2+2*x2+e2
ab_cbind(x2,y2) # array bidimensional combinando las columnas
write.table(ab, file="c:\\probb1.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
}
}
# Inciso c
for (j in 1:50) {
e1_rnorm(1,0,1)
for (i in 1:9) {
x1_runif(1,min=0,max=5)
y1_1+2*x1+e1
ab_cbind(x1,y1) # array bidimensional combinando las columnas
write.table(ab, file="c:\\probc.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
}
}
e2_rnorm(1,0,1)
for (i in 1:11) {
x2_runif(1,min=0,max=5)
y2_1+3*x2+e2
ab_cbind(x2,y2) # array bidimensional combinando las columnas
write.table(ab, file="c:\\probc1.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
}
}
# Inciso d
for (j in 1:50) {
e1_rnorm(1,0,1)
for (i in 1:9) {
x1_runif(1,min=0,max=5)
y1_1+2*x1+e1
ab_cbind(x1,y1) # array bidimensional combinando las columnas
write.table(ab, file="c:\\probd.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
}
}
e2_rnorm(1,0,3)
for (i in 1:11) {
x2_runif(1,min=0,max=5)
y2_1+3*x2+e2

```

```

ab_ cbind(x2,y2) # array bidimensional combinando las columnas
write.table(ab, file="c:\\probd1.txt", append = TRUE , col.names = FALSE,
row.names=FALSE)
}
}
datos_a <- read.table("c:\\proba.txt", header=F)
datos_a1 <- read.table("c:\\proba1.txt", header=F)
datos_b <- read.table("c:\\probb.txt", header=F)
datos_b1 <- read.table("c:\\probb1.txt", header=F)
datos_c <- read.table("c:\\probc.txt", header=F)
datos_c1 <- read.table("c:\\probc1.txt", header=F)
datos_d <- read.table("c:\\probd.txt", header=F)
datos_d1 <- read.table("c:\\probd1.txt", header=F)
par(mfrow=c(2,2))
plot(datos_a, xlab="(a)",ylab="", ylim=c(0,15))
points(datos_a1, col=2)
plot(datos_b, xlab="(b)",ylab="", ylim=c(0,15))
points(datos_b1, col=2)
plot(datos_c, xlab="(c)",ylab="", ylim=c(0,15))
points(datos_c1, col=2)
plot(datos_d, xlab="(d)",ylab="", ylim=c(0,15))
points(datos_d1, col=2)

```

A4.2

Código en R para hacer la gráfica del ejemplo 2.

```

abundancia.c_c(681,320,244,219,422,441,347,176,229,363,310,161,337,331,337,
825,707,514,901,957)
biomasa.c_c(28.8,12.1,7.5,6.5,9.9,12.7,7.6,8.5,7.3,14.1,11.0,9.6,12.0,9.2,21.5,30.4,
24.4,11.1,28.5,26.0)
log.abundancia.c_log(abundancia.c)
m1_lm(log.abundancia.c ~biomasa.c)
abundancia.s_c(67,222,131,225,71,391,142,123,186,239,294,256,435,460,222,572,
224,344,661,439,277,129,430,338,530)
biomasa.s_c(2.5,12.5,8.6,10.8,3.7,12.0,4.5,4.9,8.0,9.7,4.1,6.7,11.2,10.5,4.8,10.1,6.4,
4.6,14.3,14.2,8.7,3.6,8.6,4.5,11.5)
log.abundancia.s_log(abundancia.s)
m2_lm(log.abundancia.s ~biomasa.s)
summary(m1)
Call:
glm(formula = labundancia1 ~biomasa1)
Deviance Residuals:
Min 1Q Median 3Q Max

```

```

-0.58618 -0.11019 -0.01120 0.18065 0.49374
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.149821 0.136718 37.668 < 2e-16 ***
biomasa1 0.053933 0.008063 6.689 2.83e-06 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.0838382)
Null deviance: 5.2606 on 19 degrees of freedom
Residual deviance: 1.5091 on 18 degrees of freedom
AIC: 11.073
Number of Fisher Scoring iterations: 2
summary(m2)
Call:
glm(formula = labundancia2 ~biomasa2)
Deviance Residuals:
Min 1Q Median 3Q Max
-0.75682 -0.30000 0.01585 0.31114 0.71518
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.58389 0.23121 19.825 5.85e-16 ***
biomasa2 0.11773 0.02639 4.461 0.000178 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.2109639)
Null deviance: 9.0507 on 24 degrees of freedom
Residual deviance: 4.8522 on 23 degrees of freedom
AIC: 35.961
Number of Fisher Scoring iterations: 2
# Valores ajustados
yg.c_5.14982+0.05393*biomasa.c
yg.s_4.5839+0.1177*biomasa.s
# Gráficas
plot(biomasa.s,log.abundancia.s, xlab="Biomasa",ylab="log(abundancia)",
ylim=c(4,7))
points(biomasa.c, log.abundancia.c, col=2)
lines(biomasa.s, yg.s, col=3)
lines(biomasa.c, yg.c, col=4)
legend(3,7,c("Especie S", "Especie C", "Ajuste S", "Ajuste C"), lty=rep(1,4),
col=(1:4)).

```

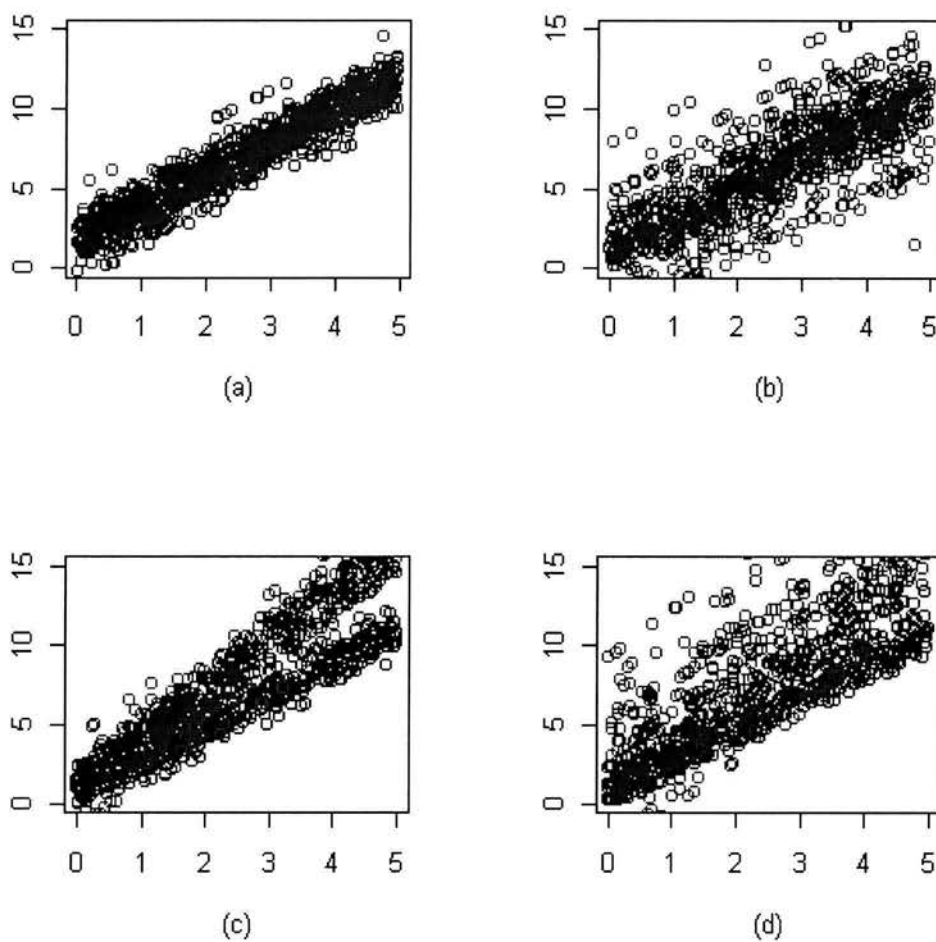


Figura 1: Las cuatro gráficas de dispersión corresponden a las cuatro configuraciones paramétricas dadas en la tabla 1.

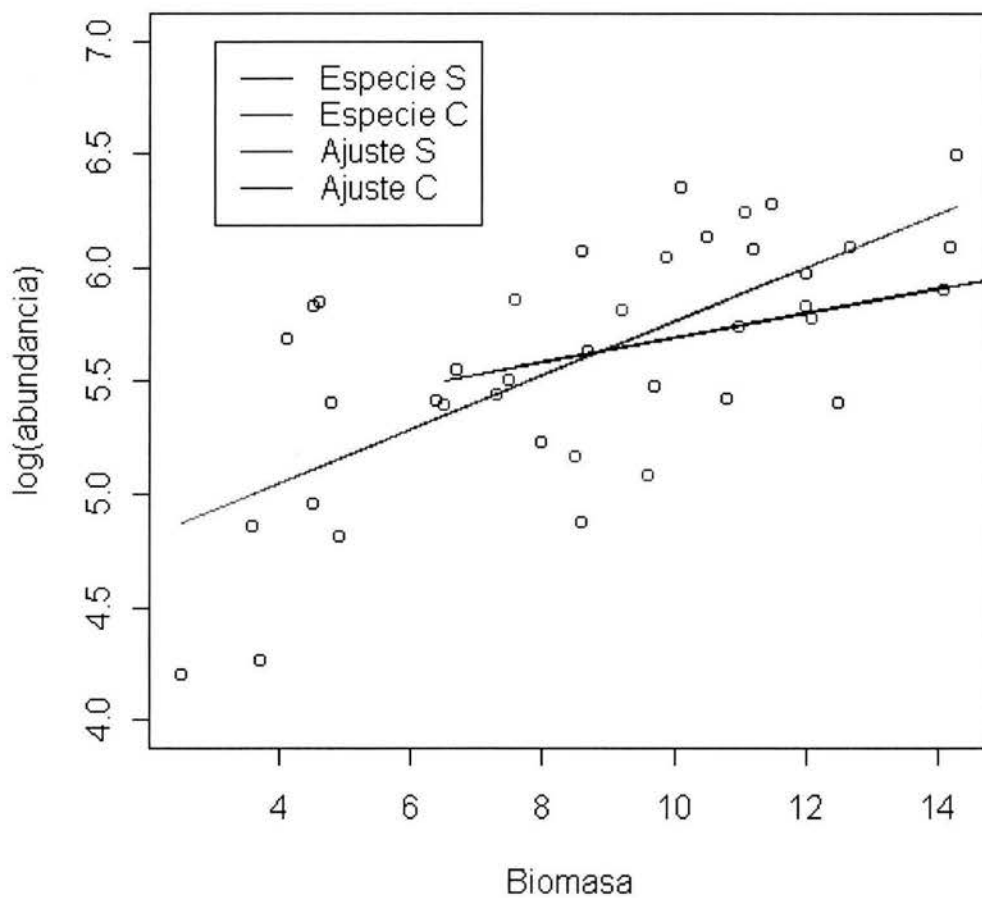


Figura 2: Gráfica de dispersión de los datos dados en la tabla 3. Las líneas son los ajustes por mínimos cuadrados.

Bibliografía

- [1] Aitkin, M. (1991). Posterior Bayes Factors (with discussion). *J. Roy. Statist. Soc. B*, **53**, 111-142.
- [2] Albert, J. H. (1990). A Bayesian Test for a Two-Way Contingency Table Using Independence Priors. *Canadian Journal of Statistics*, **14**, p. 1583-1590.
- [3] Berger, J. O. (1985) Statistical Decision Theory and Bayesian Analysis. *Second Edition*. Springer-Verlag.
- [4] Berger, J. O. and Bernardo, J. M. (1992). On the development of the reference prior method. In *Bayesian Statistics IV*, (eds., J. M. Bernardo, et, al.), London: Oxford University Press.
- [5] Berger, J. O. and Pericchi, L. R. (1993). Intrinsic Bayes factors for model selection and prediction in the general linear model. Technical Report.
- [6] Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for model selection and Prediction. *Journal of the American Statistical Association*, **91**, 109-122.
- [7] Berger, J.O. and Pericchi, L. R. (1997). On the justification of default and intrinsic Bayes Factor. In *Modeling and Prediction*, (Eds. J.C. Lee et al), New York: Springer-Verlag, 276-293.
- [8] Berger, J. O. and Pericchi, L. R: (1998). On Criticism and Comparison of Default Bayes Factors for Model Selection and Hypothesis Testing. *Proceedings of the Workshop on Model Selection*, Ed. W. Racugno, Pitágora Ed., Bologna, 1-50.
- [9] Berger, J. O. and Pericchi, L. R: (1998). On the justification of Default and Intrinsic Bayes Factor. In *Modelling and Prediction*, eds. J. C. Lee et. al., Springer- Verlag, 276-293.
- [10] Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, **41**, 113-147.
- [11] Bernardo, J. M. and Smith, A. F. M. (1993). *Bayesian Theory*. John Wiley and Sons, Inc. New York. NY.

- [12] Box, G. E. P., and Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis. Reading, M. A: Addison-Wesley.
- [13] Casella, G. and Moreno, E. (2002). Objective Bayesian variable selection. Technical Report. University of Granada (submitted).
- [14] Conerly, M. D. and Mansfield, E. R. (1988). an approximate test for comparing heteroscedastic regression models. *Journal of the American Statistical Association*, **83**, 811-817.
- [15] Dmochowski, J. (1994). Intrinsic Priors Via Kullback-Liebler Geometry. Technical Report , Purdue University, Department of Statistics.
- [16] George E. I., and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, *88*, 881-889.
- [17] Hodges, J. L. and Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B*. 16.
- [18] Jeffreys, H. (1961), *Theory of Probability*, London: Oxford University Press.
- [19] Kass, R. E. and Raftery, Adrián E. (1993). Bayes Factor. Technical Report no. 254 Department of Statistics, University of Washington. Technical Report no. 571 Department of Statistics Carnegie-Mellon University, Revision 3: July, 6.1994.
- [20] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343-1370.
- [21] Lehmann E.L.(1959). *Testing Statistical Hypotheses*. John Wiley 85 Sons, Inc., New York, NY.
- [22] Lindley, D. V.(1970). *An Introduction to Probability and Statistics from a Bayesian Viewpoint* (Vol. 2). Cambridge University Press, Cambridge.
- [23] McCulloch, R. E, and Rossi, P. E. (1993). Bayes Factors for Nonlinear Hypotheses and Likelihood Distributions, *Biometrika*, *79*, 663-676.
- [24] MacPherson, G. (1990). *Statistics in Scientific Investigation*. New York: Springer-Verlag.
- [25] Mitchell, T. J. and Beuachamp, J. J. (1988). Bayesian Variable Selection in Regression. *Journal of the American Statistical Association*, *83*, p. 1023-1032.
- [26] Moreno, E., Bertolino, F. and Racugno, W. (1999). Default Bayesian Analysis of the Behrens-Fisher Problem. *Journal of Statistical Planning and Inference*, **81**, 323-333.

- [27] Moreno, E., Bertolino, F. and Racugno, W. (2001). An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing. *Journal of the American Statistical Association*, **93**, 1451-1460.
- [28] Moreno, E., Bertolino F. and Racugno W. Bayesian Model Selection Approach to Analysis of Variance under Heteroscedasticity. Technical Report. University of Granada.
- [29] Moreno, E. and Liseo, B. (2001). Default priors for testing the number of components of a mixture. *Journal of Statistical Planning and Inference* (to appear).
- [30] Moreno, E. Torres, F. and Casella, G. (2002). Testing Equality of Regression Coefficients in Heteroscedastic Normal Regression Models. Technical Report. University of Granada.
- [31] O'Hagan, A. (1995)-Fractional Bayes Factors for Model Comparison. *Journal Royal Statistical Society B*, **57**, No. 1, 99-138
- [32] O'Hagan, A. (1997). Properties of Intrinsic and Fractional Bayes Factors. *Test*, **1**, 101-118.
- [33] Peña, D. A. (2002). Análisis de datos multivariantes. McGraw-Hill/Interamericana de España, S.A.U.
- [34] Poirier, D. J.(1985). Bayesian Hypothesis Testing in Linear Models with Continuously Induced conjugate Prior Across Hypotheses. Bayesian Statistics 2, eds. J. M. Bernardo, et, al., New York: Elsevier, p. 711-722.
- [35] Raftery, A. E. (1993). Aproximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models. Technical Report No. 255. University of Washington, Department of Statistics.
- [36] San Martini, A. and Spezzaferrri, F. (1984). A predictive model selection criterion. *Journal Royal Statistical Society, B*, **46** 296-303.
- [37] Schwarz, G. (1978). Estimating the dimension of a model *Annals of Statistics*, Volumen 6, Issue 2, 461-464.
- [38] Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factor and choice criteria for linear models, *Journal of the Royal Statistical Society, Serie B*, **42**, 213-220.
- [39] Stewart, L. (1987). Hierarchical Bayesian Analysis using Monte Carlo Integration: Computing Posterior Distribution When There are Many Possible Models. *The Statistician*, **36**, p. 211-219.
- [40] Zellner, A. and Siow (1980). Posterior Odds for Selected Regression Hypotheses. Bayesian Statistics 1, eds. J. M. Bernardo, et, al., Valencia: Valencia University Press.