Introducción a los Modelos Bayesianos con aplicaciones en R y JAGS

Estadística Bayesiana

Introducción a los Modelos Bayesianos con aplicaciones en R y JAGS

Estadística Bayesiana

Andrés Gutiérrez Rojas, M.Sc., Ph.D.
Hanwen Zhang, M.Sc., Ph.D.

Bogotá D.C., Colombia.
2009

Andrés Gutiérrez Rojas
Hanwen Zhang
Centro de Investigaciones y Estudios Estadísticos
Facultad de Estadística
Universidad Santo Tomás

Email: hugogutierrez@usantotomas.edu.co

ISBN-10: 1-387-30814-8

Impreso en Bogotá, Colombia.

Para Yong Gutiérrez y Ming Gutiérrez

Prefacio

≪La estadística Bayesiana cumple un paradigma que es muy bonito≫.

No podría dejar de agradecer a mis mejores maestros que, con su estilo en el tablero y su forma particular forma de ver el desarrollo de la ciencia estadística, me introdujeron al mundo bayesiano. El primero de ellos, el profesor Luis Guillermo Díaz, quien en sus cátedras de modelos lineales e inferencia estadística multivariada en posgrado sembró la curiosidad en sus alumnos acerca de distintas maneras ajuste de relaciones causales.

Este libro es el resultado de un estudio juicioso y aplicado de los métodos bayesianos que llevé a cabo como preparación para mi examen de calificación doctoral.

Agradecer a Jairo Fúquene por la ayuda en algunos programas computacionales.

Contenido

1 Tópicos básicos

1.1 Teoría de la decisión

El problema estadístico de estimar un parámetro se puede ver dentro del contexto de la teoría de decisión: la estimación que proveemos, sea en el ámbito de la estadística clásica o la estadística bayesiana, depende de los datos muestrales, $\mathbf{X}$, de tal forma que si éstos cambian, nuestra estimación también cambia. De esta manera, el proceso de estimación se puede ver como una función que toma un conjunto de datos muestrales y los convierte en una estimación de nuestro parámetro de interés, $A(\mathbf{X})$ o simplemente A. En la teoría de decisión, la anterior función se conoce como una regla de decisión.

Así como en la vida cotidiana, por la incertidumbre del futuro, en el ámbito estadístico, por la incertidumbre acerca del parámetro, toda acción que uno toma (toda estimación que uno provea) puede traer consigo un grado de falla o riesgo. Y es necesario tomar la acción óptima que de alguna forma minimice ese riesgo. Formalizando esta idea intuitiva, tenemos la función de pérdida L que asocia cada dupla de la acción tomada y el parámetro de interés θ, $(A,\ \theta)$ con un número no negativo que cuantifica la pérdida que ocasiona la acción (o la estimación) A con respecto al parámetro θ.

Es claro que se desea escoger aquella acción que minimice de alguna forma la pérdida que ésta ocasiona, pero la función L no se puede minimizar directamente, puesto que

- En el ámbito de la estadística clásica, el parámetro θ se considera fijo, y los datos muestrales $\mathbf{X}$ aleatorios, así como la función de pérdida L depende de $\mathbf{X}$, entonces ésta también será una variable aleatoria, y no se puede minimizar directamente. Y se define el riesgo o la pérdida promedio como la esperanza matemática de L; denotando el riesgo como R, éste está definido como $R = E(L)$. La esperanza se toma con respecto a la distribución probabilística de $\mathbf{X}$.
- En el ámbito de la estadística bayesiana, θ es una cantidad aleatoria, y la herramienta fundamental para conocer características de θ es su función de densidad posterior $p(\theta|\mathbf{X})$. En este caso, el riesgo R se define como

$$R = E(L) = \int L(A, \theta) p(\theta|\mathbf{X}) d\theta$$

En cualquier de los dos casos anteriores, buscaremos la estimación que minimice el riesgo R. Ilustramos los anteriores conceptos en los siguientes ejemplos tanto en la estadística clásica como en la estadística bayesiana.

Ejemplo 1.1.1. *Sea X_i con $i = 1, \cdots, n$ una muestra aleatoria con media θ y varianza σ^2, ambas fijas, y suponga que se desea encontrar el mejor estimador de θ bajo la función de pérdida cuadrática dada por*

$$L(A, \theta) = (A - \theta)^2$$

cuyo riesgo asociado está dado por $R = E(A-\theta)^2$. En primer lugar buscaremos dicho estimador dentro de todas las formas lineales de X_i, es decir, los estimadores de la forma $A = \sum_{i=1}^{n} c_i X_i$, de esta forma, el riesgo se puede expresar como

$$\begin{aligned} R = E(A-\theta)^2 &= Var(A) + (E(A) - \theta)^2 \\ &= \sum_{i=1}^{n} c_i^2 \sigma^2 + \theta^2 (\sum_{i=1}^{n} c_i - 1)^2 \end{aligned}$$

Y al buscar los coeficientes c_i que minimizan la anterior expresión, encontramos que $c_i = \theta^2/(\sigma^2 + n\theta^2)$ para todo i. Como estos coeficientes conducen a un estimador que depende del parámetro desconocido, concluimos que no hay ningún estimador que minimiza el riesgo.

Para encontrar una solución, es necesario restringir aún más el rango de estimadores, para eso, se restringe que $\sum_{i=1}^{n} c_i = 1$, de esta forma el riesgo está dado por $R = \sum c_i^2 \sigma^2$, y al minimizar $\sum c_i^2$ sujeto a la restricción de $\sum c_i = 1$. La solución está dada por $c_i = 1/n$ para todo i, y así encontramos que el mejor estimador dentro de todas formas lineales con $\sum c_i = 1$ es la media muestral $\bar{X}$.

Ejemplo 1.1.2. *Suponga que se desea estimar un parámetro de interés θ en el contexto de la estadística bayesiana y denotamos la función de densidad posterior de θ como $\pi(\theta)$, entonces si utilizamos la función de pérdida cuadrática, entonces el riesgo asociado será*

$$R = E(L(A, \theta)) = E(A - \theta)^2 = Var(\theta) + (E(\theta) - A)^2$$

que es minimizado si $A = E(\theta)$. Es decir la mejor acción para estimar θ es utilizar la esperanza de θ tomada con respecto a la distribución posterior $\pi(\theta)$.

Ejemplo 1.1.3. *En el mismo contexto del ejemplo anterior, si cambiamos la función de pérdida a la siguiente*

$$L(A, \theta) = |A - \theta| = (A - \theta) I_{(A \geq \theta)} + (\theta - A) I_{(\theta > A)}$$

Y el riesgo está dado por

$$\begin{aligned} R &= E(L(A, \theta)) \\ &= \int L(A, \theta) \pi(\theta) d\theta \\ &= \int_{(A \geq \theta)} (A - \theta) \pi(\theta) d\theta + \int_{(\theta > A)} (\theta - A) \pi(\theta) d\theta \end{aligned}$$

Derivando el riesgo con respecto a la acción A, se tiene que

$$\frac{\partial R}{\partial A} = \int_{(A \geq \theta)} \pi(\theta) d\theta - \int_{(\theta > A)} \pi(\theta) d\theta$$

Igualando a cero, tenemos que

$$\int_{(A \geq \theta)} \pi(\theta) d\theta = \int_{(\theta > A)} \pi(\theta) d\theta = 0.5$$

Y concluimos que la acción A que induce menor riesgo corresponde al percentil 50 % o la mediana de la distribución posterior de θ.

De los anteriores ejemplos vemos que bajo un mismo contexto, cuando se utilizan diferentes funciones de pérdidas, también obtenemos distintas estimaciones.

1.2 Algunos resultados de probabilidad

A continuación se presentan definiciones y resultados de probabilidad. en términos de notación se utilizará indistintamente la expresión de integral, $\int$, que implicará la integral, en el caso de las variables aleatorias continuas, o la sumatoria, en el caso de las variables aleatorias discretas.

DefiniciÃ³n1.2.1. *Sean $\mathbf{X} = (X_1, \ldots, X_p)'$, $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ dos vectores aleatorios definidos sobre los espacios de muestreo $\mathcal{X}$, $\mathcal{Y}$, respectivamente. Suponga que la distribución conjunta de estos vectores aleatorios está dada por $p(\mathbf{X}, \mathbf{Y})$. Se define la distribución marginal de $\mathbf{X}$ como*

$$p(\mathbf{X}) = \int p(\mathbf{X}, \mathbf{Y}) \; d\mathbf{Y} \tag{1.2.1}$$

y la distribución condicional de $\mathbf{X}$ dado $\mathbf{Y}$ como

$$p(\mathbf{X} \mid \mathbf{Y}) = \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{Y})} \tag{1.2.2}$$

Resultado 1.2.1. *Suponga los vectores $\mathbf{X}$, $\mathbf{Y}$ y un tercer vector $\mathbf{Z} = (Z_1, \ldots, Z_r)'$ definido sobre el espacio de muestreo $\mathcal{Z}$. Entonces se tiene que*

$$p(\mathbf{X} \mid \mathbf{Z}) = \int p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \; d\mathbf{Y} \tag{1.2.3}$$

y

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} \tag{1.2.4}$$

Prueba. En primer lugar, nótese que

$$\begin{aligned}\int p(\mathbf{X},\mathbf{Y}\mid\mathbf{Z})\ d\mathbf{Y} &= \int \frac{p(\mathbf{X},\mathbf{Y},\mathbf{Z})}{p(\mathbf{Z})}\ d\mathbf{Y}\\ &= \frac{1}{p(\mathbf{Z})}\int p(\mathbf{X},\mathbf{Y},\mathbf{Z})\ d\mathbf{Y}\\ &= \frac{1}{p(\mathbf{Z})}p(\mathbf{X},\mathbf{Z}) = p(\mathbf{X}\mid\mathbf{Z})\end{aligned}$$

Por otro lado,

$$\frac{p(\mathbf{X},\mathbf{Y}\mid\mathbf{Z})}{p(\mathbf{Y}\mid\mathbf{Z})} = \frac{p(\mathbf{X},\mathbf{Y},\mathbf{Z})}{p(\mathbf{Z})}\Big/\frac{p(\mathbf{Y},\mathbf{Z})}{p(\mathbf{Z})} = \frac{p(\mathbf{X},\mathbf{Y},\mathbf{Z})}{p(\mathbf{Y},\mathbf{Z})} = p(\mathbf{X}\mid\mathbf{Y},\mathbf{Z})$$

■

DefiniciÃ³n1.2.2. *Sean* $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ *vectores aleatorios. Se dice que* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *con respecto a* $\mathbf{Z}$ *si satisfacen la siguiente expresión*

$$p(\mathbf{X},\mathbf{Y}\mid\mathbf{Z}) = p(\mathbf{X}\mid\mathbf{Z})p(\mathbf{Y}\mid\mathbf{Z}) \tag{1.2.5}$$

Resultado 1.2.2. *Si* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *con respecto a* $\mathbf{Z}$*, entonces se tiene que*

$$p(\mathbf{X}\mid\mathbf{Y},\mathbf{Z}) = p(\mathbf{X}\mid\mathbf{Z}) \tag{1.2.6}$$

Prueba. Como $p(\mathbf{X},\mathbf{Y}\mid\mathbf{Z}) = \dfrac{p(\mathbf{X},\mathbf{Y},\mathbf{Z})}{p(\mathbf{Z})}$, entonces

$$p(\mathbf{X}\mid\mathbf{Y},\mathbf{Z}) = \frac{p(\mathbf{X},\mathbf{Y},\mathbf{Z})}{p(\mathbf{Y},\mathbf{Z})} = \frac{p(\mathbf{X},\mathbf{Y}\mid\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{Y},\mathbf{Z})} = \frac{p(\mathbf{X}\mid\mathbf{Z})p(\mathbf{Y}\mid\mathbf{Z})}{p(\mathbf{Y}\mid\mathbf{Z})} = p(\mathbf{X}\mid\mathbf{Z})$$

■

Resultado 1.2.3. *Si* $\mathbf{X}$ *es independiente de* $\mathbf{Y}$*, entonces* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *dada cualquier otro vector, digamos* $\mathbf{Z}$*.*

Prueba. Nótese que

$$p(X,Y\mid Z) = p(X\mid Y,Z)p(Y\mid Z) = p(X\mid Z)p(Y\mid Z)$$

puesto que, utilizando la hipótesis de independencia, se tiene que

$$p(X\mid Y) = p(X)$$

■

1.3 Teorema de Bayes

Desde la revolución estadística de Pearson y Fisher, la inferencia estadística busca encontrar los valores que parametrizan a la distribución desconocida de los datos. El primer enfoque, propuesto por Pearson, afirmaba que si era posible observar a la variable de interés en todos y cada uno de los individuos de una población, entonces era posible calcular los parámetros de la distribución de la variable de interés; por otro lado, si sólo se tenía acceso a una muestra representativa, entonces era posible calcular una estimación de tales parámetros. Sin embargo, Fisher discrepó de tales argumentos, asumiendo que las observaciones están sujetas a un error de medición y por lo tanto, así se tuviese acceso a toda la población, es imposible calcular los parámetros de la distribución de la variable de interés.

Del planteamiento de Fisher resultaron una multitud de métodos estadísticos para la estimación de los parámetros poblacionales. Es decir, si la distribución de $\mathbf{Y}$ está parametrizada por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, $\boldsymbol{\theta} \in \Theta$ con Θ el espacio paramétrico inducido por el comportamiento de la variable de interés, el objetivo de la teoría estadística inferencial es calcular una estimación $\hat{\boldsymbol{\theta}}$ del parámetro $\boldsymbol{\theta}$ por medio de los datos observados. En este enfoque, los parámetros se consideran cantidades fijas y constantes. Sin embargo, en la última mitad del siglo XX, algunos investigadores estadísticos comenzaron a reflexionar acerca de la naturaleza de $\boldsymbol{\theta}$ y enfocaron la inferencia estadística de una manera distinta: asumiendo que la distribución de la variable de interés está condicionada a valores específicos de los parámetros. Es decir, en términos de notación, si la variable de interés es $\mathbf{Y}$, su distribución condicionada a los parámetros toma la siguiente forma $p(\mathbf{Y} \mid \boldsymbol{\theta})$. Esto implica claramente que en este nuevo enfoque la naturaleza de los parámetros no es constante sino estocástica.

En términos de inferencia para $\boldsymbol{\theta}$, es necesario encontrar la distribución de los parámetros condicionada a la observación de los datos. Para este fin, es necesario definir la distribución conjunta de la variable de interés con el vector de parámetros.

$$p(\boldsymbol{\theta}, \mathbf{Y}) = p(\boldsymbol{\theta}) p(\mathbf{Y} \mid \boldsymbol{\theta})$$

A la distribución $p(\boldsymbol{\theta})$ se le conoce con el nombre de distribución *previa* y en ella se enmarcan todas y cada una de las creencias que se tienen acerca del comportamiento estocástico del vector de parámetros antes de que ocurra la recolección de los datos y $p(\mathbf{Y} \mid \boldsymbol{\theta})$ es la distribución de muestreo o verosimilitud o distribución de los datos.Por otro lado, la distribución del vector de parámetros condicionada a los datos observados está dada por

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{p(\boldsymbol{\theta}, \mathbf{Y})}{p(\mathbf{Y})} = \frac{p(\boldsymbol{\theta}) p(\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y})} \tag{1.3.1}$$

A la distribución $p(\boldsymbol{\theta} \mid \mathbf{Y})$ se le conoce con el nombre de distribución *posterior* y en ella se enmarcan las creencias actualizadas acerca del comportamiento estocástico del vector de parámetros teniendo en cuenta los datos observados. Nótese que la expresión (**??**), se compone de una fracción cuyo denominador no depende del

vector de parámetros y considerando a los datos observados como fijos, corresponde a una constante y puede ser obviada. Por lo tanto, otra representación de la regla de Bayes está dada por

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.3.2}$$

? menciona que esta expresión se conoce como la distribución *a posterior no-normalizada* y encierra el núcleo técnico de la inferencia bayesiana.

Resultado 1.3.1. *La expresión $p(\mathbf{Y})$ corresponde a una constante k tal que*

$$k = p(\mathbf{Y}) = E_{\boldsymbol{\theta}}[p(Y \mid \boldsymbol{\theta})]$$

Prueba. Nótese que

$$k = p(\mathbf{Y}) = \int p(\mathbf{Y}, \boldsymbol{\theta})\ d\boldsymbol{\theta} = \int p(\boldsymbol{\theta})p(\mathbf{Y} \mid \boldsymbol{\theta})\ d\boldsymbol{\theta}.$$

entonces

$$\begin{aligned} k &= \int p(\mathbf{Y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &= E_{\boldsymbol{\theta}}[p(Y \mid \boldsymbol{\theta})] \end{aligned}$$

■

El reverendo Thomas Bayes murió sin publicar este resultado.....

Ejemplo 1.3.1. *Uno de los primeros acercamientos de cualquier profesional a la estadística bayesiana se da en un curso básico de probabilidades en donde el docente presenta con cierta rigurosidad los conceptos básicos e introductorios de la teoría de probabilidad. En un sobrevuelo de tales conceptos es posible recordar términos como experimento, espacio muestral, función de probabilidad y sigma álgebra. Justo después del repaso de rigor acerca de los axiomas de probabilidades y sus teoremas más significativos, el curso da una curva cerrada y el alumno es introducido en conceptos más profundos como la probabilidad condicional.*

En estos tópicos, tanto el maestro como el alumno asumen que los temas básicos ya están entendidos y que no existe necesidad de volver atrás. A manera de introducción, los autores desean hacer notar a los lectores que requieren de herramientas de modelamiento más sofisticadas, que es necesario volver atrás - al menos en esta primera página - para sentar las bases de la autopista de alta velocidad como lo es el análisis bayesiano. No tiene sentido que el investigador utilice las poderosas herramientas bayesianas si no entiende que sus bases probabilísticas están bien sustentadas.

Para entrar en detalle, vamos a utilizar un ejemplo en donde el lector se sentirá identificado con aquellas épocas universitarias de un curso de probabilidades: suponga que una fábrica del sector industrial produce bolígrafos y que la producción está a cargo de tres máquinas. La primera máquina produce el 50 % del total

de bolígrafos en el año, la segunda máquina produce el 30 % y la última maquina produce el restante 20 %. Por supuesto, esta producción esta sujeta al error y por tanto, basados en la experiencia, es posible reconocer que, de los artículos producidos por la primera máquina, el 5 % resultan defectuosos; de los artículos producidos por la segunda máquina, el 2 % resultan defectuosos y , de los artículos producidos por la última máquina, el 6 % resultan defectuosos.

Una pregunta natural que surge es acerca de la probabilidad de selección de un artículo defectuoso y para responder a esta pregunta con «rigurosidad de probabilísta» es necesario enfocar nuestra atención en los tópicos básicos que dejamos atrás. En primer lugar el experimento en cuestión es la selección de un bolígrafo. Para este experimento, una terna $(\Omega, \mathfrak{F}, P)$ [1]*, llamada comúnmente espacio de medida o espacio de probabilidad, está dada por*

1. *El espacio muestral:* $\Omega = \{defectuoso, Nodefectouso\}$
2. *La* σ*-álgebra:* $\mathfrak{F} = \{\Omega, \phi, \{Defectuoso\}, \{NoDefectuoso\}\}$
3. *La función de probabilidad:*

$$\begin{aligned} p : \mathfrak{F} &\longrightarrow [0,1] \\ \Omega &\longrightarrow 1 \\ \phi &\longrightarrow 0 \\ \{Defectuoso\} &\longrightarrow P(D) \\ \{Defectuoso\} &\longrightarrow 1 - P(D) \end{aligned}$$

en donde, acudiendo al teorema de probabilidad total, se define

$$p(D) = p(D \mid M1)P(M1) + p(D \mid M2)P(M2) + p(D \mid M3)P(M3)$$

Sin embargo, también es posible plantearse otro tipo de preguntas que sirven para calibrar el proceso de producción de artículos defectuosos. Por ejemplo, cabe preguntarse acerca de la probabilidad de que habiendo seleccionado un artículo defectuoso, éste provenga de la primera máquina[2]*. En esta ocasión, el experimento ha cambiado y ahora se trata de seleccionar un artículo defectuoso y para responder a tal cuestionamiento, se debe establecer rigurosamente el espacio de probabilidad que puede estar dado por*

1. *El espacio muestral:* $\Omega = \{M1, M2, M3\}$
2. *La* σ*-álgebra:* $\mathfrak{F}^{+} = \{\Omega, \phi, \{M1\}, \{M2, M3\}\}$

[1] Ω denota el conjunto de todos lo posibles resultados del experimento, $\mathfrak{F}$ denota una σ-álgebra y P hace referencia ana medida de probabilidad propiamente definida.

[2] Por supuesto que la pregunta también es válida al indagar por la probabilidad de que habiendo seleccionado un artículo defectuoso, éste provenga de la segunda o tercera máquina.

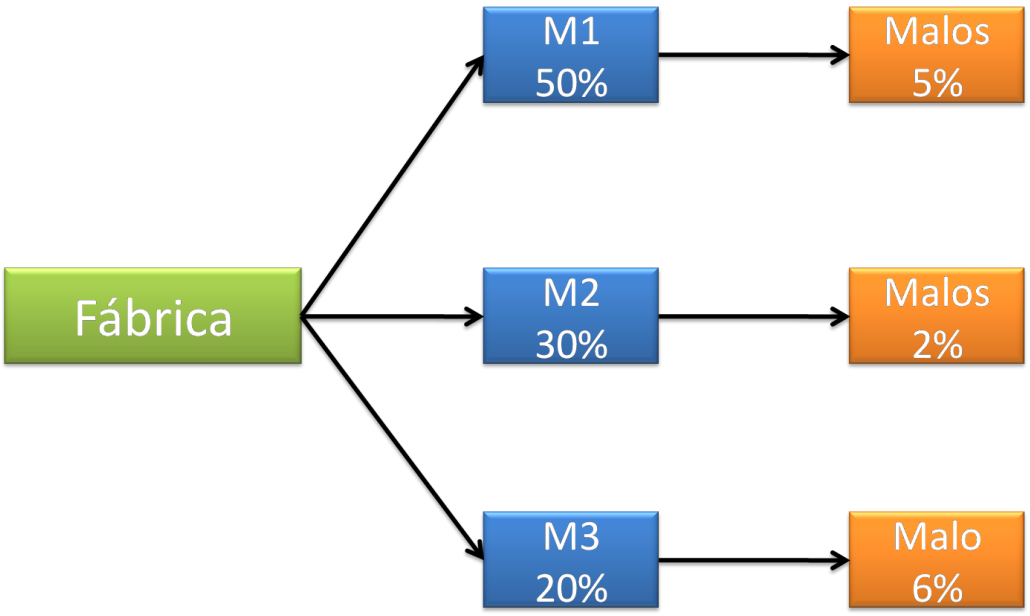

Figura 1.1: *Plano del proceso industrial en la fábrica de bolígrafos*

3. La función de probabilidad:

$$\begin{aligned} p : \mathfrak{F}^{+} &\longrightarrow [0,1] \\ \Omega &\longrightarrow 1 \\ \phi &\longrightarrow 0 \\ \{M1\} &\longrightarrow p(M1 \mid D) \\ \{M2, M3\} &\longrightarrow 1 - p(M1 \mid D) \end{aligned}$$

en donde, acudiendo a la definición de probabilidad condicional, se define

$$p(M1 \mid D) = \frac{p(D \mid M1)P(M1)}{p(D \mid M1)P(M1) + p(D \mid M2)P(M2) + p(D \mid M3)P(M3)}$$

La anterior función de probabilidad se conoce con el nombre de regla de probabilidad de Bayes y, aparte de ser el baluarte de la mayoría de investigaciones estadísticas que se plantean hoy en día, ha sido la piedra de tropiezo de muchos investigadores radicales que trataron de estigmatizar este enfoque tildando a sus seguidores de mediocres matemáticos y pobres probabilistas afirmando que la regla de probabilidad de Bayes es sólo un artilugio diseñado para divertirse en el tablero.

Pues bien, la interpretación de la regla de bayes se puede realizar en el sentido de actualización de la estructura probabilística que gobierna el experimento. Y esta actualización tiene mucho sentido práctico cuando se cae en la cuenta de que la vida real está llena de calibradores y que las situaciones generadas son consecuencia de algún cambio estructural. De esta forma, el conocimiento de la probabilidad de que el artículo sea producido por la primera máquina se actualiza al conocer que este artículo particular es defectuoso y de esta manera calibra la estructura aleatoria que existe detrás del contexto de la fábrica de bolígrafos. Aparte de servir para resolver problemas como el anteriormente mencionado, la regla de bayes ha marcado el comienzo de un nuevo enfoque de análisis de datos, no solamente porque hace explícitas las relaciones causales entre los procesos aleatorios, sino también porque facilita la inferencia estadística y la interpretación de los resultados.

En el blog de $http://www.iq.harvard.edu/blog/sss/archives/2009/11/breast_cancer_r.shtml$ XXXXXXXXXXXXXXXXXXXX, expone el siguiente ejemplo interesante de la aplicación del teorema de Bayes.

Ejemplo 1.3.2. *El Grupo de Trabajo de Servicios Preventivos de los Estados Unidos (USPSTF por sus siglas en inglés) hizo unas nuevas y controversiales recomendaciones sobre la detección del cáncer de mama (ver página $http://www.uspreventiveservicestaskforce.o$ dentro de los cuales, no recomienda el examen de la mamografía en mujeres entre 40 y 49 años de edad, afirmando que la práctica bienal de este examen debe ser una decisión individual según el contexto particular de la paciente, mientras que por muchos años, se han dicho a las mujeres que se debe realizar la mamografía una vez cumplidos los 40 años. Por otro lado, USPSTF sí recomienda tal práctica de forma bienal en grupos de mujeres de entre 50 y 74 años de*

edad, puesto que USPSTF no encontró suficiente evidencia de beneficio o daño adicional en realizar este examen en mujeres mayores que los 74 años. Otra recomendación que hizo USPSTF es no realizar auto exámanes de senos, contrario a las recomendaciones y consejos que da la mayoría de los profesionales y organizaciones de la salud, incluyendo la Amerian Cancer Society *(ver http : //www.cancer.org/acs/groups/cid/documents/webcontent/*003164 − *pdf.pdf).*

El autor del blog, después de algunas averiguaciones, encontró que

- *Los expertos estiman que un 12.3 % de las mujeres desarrollan formas invasivas del cáncer de mama durante la vida.*
- *La probabilidad de que una mujer desarrolle el cáncer de mama entre los 40 y los 479 años de edad es 1 en 69, y esta probabilidad aumenta a medida que envejezca, de tal forma que llega a ser de 1 en 38 en mujeres de entre 50 y 59 años.*
- *El cáncer de mama es más difícil de detectar en mujeres jóvenes puesto que el tejido mamario es más denso y fibroso. Los expertos estiman que la tasa de un falso positivo es de 97.8 por cada 1000 mujeres de 40 y 49 años, y esta tasa disminuye a 86.6 por cada 1000 mujeres entre 50 y 59 años.*
- *La tasa de un falso negativo es de 1 por cada 1000 mujeres de 40 y 49 años, y es de 1.1 por cada 1000 mujeres entre 50 y 59 años.*

Resumiendo las anteriores afirmaciones, tenemos las siguientes probabilidades

Probabilidad	Edad	
	40 - 49 años	50 - 59 años
p(Cáncer)	1/69	1/38
p(No cáncer)	68/69	37/38
p(Positivo| No cáncer)	0.0978	0.0866
p(Negativo| No cáncer)	0.9022	0.9134
p(Negativo| Cáncer)	0.001	0.0011
p(Positivo| Cáncer)	0.999	0.9989

Utilizando la regla de Bayes, se tiene que

$$j$$

En este capítulo se quiere regresar al camino de atrás exponiendo conceptos básicos que más adelante serán obviados pero que permanecerán latentes a los largo del desarrollo de los conceptos del libro.

1.4 Inferencia bayesiana

El enfoque bayesiano, además de especificar un modelo para los datos observados $\mathbf{Y} = (y_1, \ldots, y_n)$ dado un vector de parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, usualmente en forma de densidad condicional $p(\mathbf{Y} \mid \boldsymbol{\theta})$, supone que $\boldsymbol{\theta}$ es aleatorio

y que tiene un densidad *previa* $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$, donde $\boldsymbol{\eta}$ es un vector de hiper-parámetros. De esta forma, la inferencia concerniente a $\boldsymbol{\theta}$ se basa en una densidad *posterior* $p(\boldsymbol{\theta} \mid \mathbf{Y})$.

En términos de estimación, inferencia y predicción, el enfoque Bayesiano supone dos momentos o etapas:

1. Antes de la recolección de las datos, en donde el investigador propone, basado en su conocimiento, experiencia o fuentes externas, una distribución de probabilidad *previa* para el parámetro de interés. Con esta distribución es posible calcular estimaciones puntuales y por intervalo con el fin de confirmar que la distribución propuesta se ajusta al problema de estudio. En etapa, basados en la distribución *previa*, también es posible hacer predicciones de cantidades observables.
2. Después de la recolección de los datos. Siguiendo el teorema de Bayes, el investigador actualiza su conocimiento acerca del comportamiento probabilístico del parámetro de interés mediante la distribución *posterior* del parámetro de interés. Con esta distribución es posible calcular estimaciones puntuales y por intervalo justo como en el enfoque frecuentista. En etapa, basados en la distribución *posterior*, también es posible hacer predicciones de cantidades observables y pruebas de hipótesis acerca de la adecuación del mejor modelo a los datos observados.

1.4.1 Inferencia previa

Con las anteriores expresiones es posible calcular la probabilidad *previa* de que $\boldsymbol{\theta}$ esté en una determinada región G como

$$Pr(\boldsymbol{\theta} \in G) = \int_G p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.1}$$

En esta primera etapa es posible calcular, con fines confirmatorios **?**, la estimación puntual para el vector $\boldsymbol{\theta}$ dada por alguna medida de tendencia central para la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. En particular, si se escoge la media, entonces

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}) = \int \boldsymbol{\theta}\ p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.2}$$

También es posible calcular una región C de 100(1-α) % de credibilidad[3] para $\boldsymbol{\theta}$ que en esta primera etapa es tal que

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C) = \int_C p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.3}$$

[3]La interpretación de las regiones de credibilidad bayesianas difiere de la interpretación de las regiones de confianza frecuentistas. La primera se refiere a la probabilidad de que el verdadero valor de $\boldsymbol{\theta}$ este en la región. La segunda se refiere a la región de la distribución muestral para $\boldsymbol{\theta}$ tal que, dados los datos observados, se podría esperar que el 100(α) % de las futuras estimaciones de $\boldsymbol{\theta}$ no pertenecieran a dicha región.

1.4.2 Inferencia posterior

Es posible calcular la probabilidad *posterior* de que $\boldsymbol{\theta}$ esté en la región G dados los datos observados como

$$Pr(\boldsymbol{\theta} \in G \mid \mathbf{Y}) = \int_G p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{1.4.4}$$

De esta forma, es posible calcular la estimación puntual para el vector $\boldsymbol{\theta}$ dados los datos observados. Ésta está dada por alguna medida de tendencia central para la distribución $p(\boldsymbol{\theta} \mid \mathbf{Y})$. En particular, si se escoge la media, entonces

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \mathbf{Y}) = \int \boldsymbol{\theta}\ p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{1.4.5}$$

La región C de 100(1-α) % de credibilidad es tal que

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C \mid \mathbf{Y}) = \int_C p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{1.4.6}$$

Suponiendo que existen dos modelos $M1$ y $M2$ candidatos para $\mathbf{Y}$, se define el *Factor de Bayes* en favor del modelo $M1$ como la razón de las densidades marginales de los datos para los dos modelos y es posible demostrar que equivale a la siguiente expresión

$$FB = \frac{p(\mathbf{Y} \mid M1)}{p(\mathbf{Y} \mid M2)} = \frac{Pr(M1 \mid \mathbf{Y})/Pr(M2 \mid \mathbf{Y})}{Pr(M1)/Pr(M2)} \tag{1.4.7}$$

Para evaluar esta última expresión es necesario recurrir a las expresiones (3) y (4). El factor de bayes sólo está definido cuando la integral de la densidad marginal de $\mathbf{Y}$ bajo cada modelo converge.

1.4.3 Inferencia predictiva

En términos de inferencia predictiva existen dos etapas que cubren las «actuales» suposiciones acerca del vector de parámetros $\boldsymbol{\theta}$. En una primera etapa - antes de la observación de los datos - la suposición «actual» de $\boldsymbol{\theta}$ está dada por la densidad *previa* $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. En estos términos, utilizando el Resultado 1.2.1, la distribución predictiva *previa* de $\mathbf{Y}$ está dada por

$$p(\mathbf{y}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.8}$$

La segunda etapa - después de la recolección de los datos - actualiza las suposiciones acerca de $\boldsymbol{\theta}$ puesto que ahora éste sigue una distribución *posterior* dada

por (**??**). Por lo tanto, la distribución predictiva *posterior* de $\mathbf{Y}$ está dada por

$$\begin{aligned} p(\tilde{\mathbf{y}} \mid \mathbf{Y}) &= \int p(\tilde{\mathbf{y}}, \boldsymbol{\theta} \mid \mathbf{y}) \, d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \end{aligned} \tag{1.4.9}$$

donde $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ es la distribución de los datos evaluada en los nuevos valores $\tilde{\mathbf{y}}$. La segunda línea de anterior igualdad se obtiene utilizando el Resultado 1.1.1 y la última línea se obtiene del resultado 1.1.2 de la independencia condicional.

1.5 Información previa

blabla

1.5.1 Distribuciones conjugadas

Como se verá en los capítulos siguientes muchos problemas de inferencia bayesiana comparten la agradable cualidad de que la forma funcional de la distribución *previa* para el parámetro de interés resulta ser la misma de la distribución *posterior*. Por ejemplo:

- Cuando se tiene una muestra aleatoria de variables con distribución Bernoulli de parámetro θ, es factible pensar que una distribución *previa* apropiada para este parámetro es la distribución Beta; bajo este escenario, la distribución *posterior* también resulta ser Beta.
- En el caso en que se quiera modelar el parámetro θ concerniente a una variable aleatoria con distribución Poisson, es posible asignar como candidata para distribución *previa* a la distribución Gamma; en este caso la distribución *posterior* también resulta ser Gamma.

DefiniciÃ³n1.5.1. *Sea $\mathcal{F} = \{p(\mathbf{Y} \mid \boldsymbol{\theta})\}$ una familia de distribuciones de probabilidad. Una familia de distribuciones $\mathcal{P}$ se dice conjugada con respecto a $\mathcal{F}$ si para toda distribución* previa *$p(\boldsymbol{\theta}) \in \mathcal{P}$ y para toda distribución de muestreo o verosimilitud de las observaciones $p(\mathbf{Y} \mid \boldsymbol{\theta})$, entonces $p(\boldsymbol{\theta} \mid \mathbf{Y})$ pertenece a la familia $\mathcal{P}$.*

Esta definición es en la mayoría de los casos prácticos muy útil. Sin embargo, **?** describe los siguientes dos casos en donde esta definición es completamente inútil:

1. (Caso amplio) Sea $\mathcal{P} = \{$*Todas las distribuciones de probabilidad*$\}$ y $\mathcal{F}$ cualquier familia de distribuciones de probabilidad. Entonces $\mathcal{P}$ es conjugada con respecto a $\mathcal{F}$ puesto que toda posible distribución *posterior* será un miembro de $\mathcal{P}$.

2. (Caso restringido) Sea $\mathcal{P} = \{p \mid p(\theta = \theta_0) = 1\}$. La anterior definición de $\mathcal{P}$ corresponde a todas las distribuciones concentradas en un punto. Sea $\mathcal{F}$ cualquier familia de distribuciones de probabilidad. De esta manera, la distribución *posterior* estará dada por

$$p(\theta \mid Y) \propto p(Y \mid \theta)p(\theta) = \begin{cases} p(Y \mid \theta) \times 1 & \text{si } \theta = \theta_0 \\ p(Y \mid \theta) \times 0 & \text{si } \theta \neq \theta_0 \end{cases}$$
$$= \begin{cases} p(Y \mid \theta) & \text{si } \theta = \theta_0 \\ 0 & \text{si } \theta \neq \theta_0 \end{cases}$$

De lo anterior y dado que $\int p(\theta \mid Y)\, d\theta = 1$, entonces $p(Y \mid \theta) = 1$ si y sólo si $\theta = \theta_0$. Con el anterior razonamiento, se concluye que $\mathcal{P}$ es conjugada con respecto a $\mathcal{F}$.

Por tanto, se deben buscar distribuciones *previa* que sean conjugadas de una forma tan amplia que permita proponer una distribución *previa* adecuada, pero al mismo tiempo tan restringida para que la definición de conjugada tenga sentido práctico.

Familia exponencial

De esta manera, existe una familia que es conjugada de manera natural. Ésta se trata de la familia exponencial, tanto en su versión uniparamétrica como en la multiparamétrica. La forma funcional y las propiedades de esta familia de distribuciones se pueden consultar en el apéndice.

Una distribución de probabilidad pertenece a la familia exponencial uniparamétrica si se puede escribir de la forma

$$p(Y \mid \theta) = \exp\{d(\theta)T(y) - c(\theta)\}h(y) \tag{1.5.1}$$

donde $T(y)$ y $h(y)$ son funciones que dependen de y únicamente, y $d(\theta)$ y $c(\theta)$ son funciones que depende de θ únicamente. Análogamente, una distribución de probabilidad pertenece a la familia exponencial multi-paramétrica si se puede escribir de la forma

$$p(Y \mid \boldsymbol{\theta}) = \exp\{\mathbf{d}(\boldsymbol{\theta})'\mathbf{T}(y) - c(\boldsymbol{\theta})\}h(y) \tag{1.5.2}$$

donde $\mathbf{T}(y)$ y $\mathbf{d}(\boldsymbol{\theta})$ son funciones vectoriales, $h(y)$ y $c(\boldsymbol{\theta})$ son funciones reales.

Es bien sabido que una distribución perteneciente a la familia exponencial puede escribirse

Resultado 1.5.1. *Sea Y una variable aleatoria con función de densidad perteneciente a la familia exponencial uniparamétrica, entonces la familia exponencial uniparamétrica es conjugada con respecto a sí misma.*

Prueba. Observando la expresión (1.5.1), se debe encontrar una distribución *previa* en la familia exponencial uniparamétrica, tal que la distribución *posterior*,

resultante del producto de la distribución *previa* con la verosimilitud, sea también miembro de la familia exponencial uniparamétrica. Con base en lo anterior, la distribución *previa*, parametrizada por el hiperparámetro α, debe ser una función exponencial de los términos $d(\theta)$ y $c(\theta)$ como lo afirma **?**. Esto es,

$$p(\theta \mid \alpha) \propto \exp\{\alpha d(\theta) - \delta c(\theta)\}, \tag{1.5.3}$$

donde δ es una constante real. Por otro lado, para garantizar que $p(\theta \mid \alpha)$ sea una auténtica función de densidad se normaliza de la siguiente manera

$$\frac{1}{k(\alpha, \delta)} \exp\{\alpha d(\theta) - \delta c(\theta)\}, \tag{1.5.4}$$

con

$$k(\alpha, \delta) = \int \exp\{\alpha d(\theta) - \delta c(\theta)\}\ d\theta.$$

De esta manera, no es difícil comprobar que la definición de distribución *previa*, parametrizada por el hiper-parámetro α, pertenece a la familia exponencial, puesto que

$$p(\theta \mid \alpha) = \exp\{\underbrace{\alpha}_{d(\alpha)}\underbrace{d(\theta)}_{T(\theta)} - \underbrace{\ln k(\alpha, \delta)}_{c(\alpha)}\}\underbrace{\exp\{-\delta c(\theta)\}}_{h(\theta)}. \tag{1.5.5}$$

Por otro lado, del teorema de Bayes se tiene que

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta) p(\theta \mid \alpha) \\ &= \exp\{\alpha d(\theta) + d(\theta) T(y) - c(\theta) - \ln k(\alpha, \delta)\} \exp\{-\delta c(\theta)\} h(y) \\ &= \exp\{\underbrace{[\alpha + T(y)]}_{d(y)}\underbrace{d(\theta)}_{T(\theta)} - \underbrace{[\ln k(\alpha, \delta) - \ln h(y)]}_{c(y)}\} \underbrace{\exp\{-(\delta + 1) c(\theta)\}}_{h(\theta)} \\ &\propto \exp\{[\alpha + T(y)] d(\theta)\} \exp\{-(\delta + 1) c(\theta)\}. \end{aligned}$$

Por lo tanto, la distribución *posterior* resultante también pertenece a la familia exponencial uniparamétrica. ■

Resultado 1.5.2. *Sean* $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ *una muestra aleatoria de variables distribuidas con función de densidad común perteneciente a la familia exponencial uniparamétrica, cuya función de densidad conjunta* $p(\mathbf{Y} \mid \theta)$ *también pertenece a la familia exponencial uniparamétrica. Bajo las anteriores condiciones la familia exponencial uniparamétrica es conjugada con respecto a sí misma.*

Prueba. La demostración es inmediata utilizando el resultado anterior y notando que la forma funcional de la densidad conjunta para $\mathbf{Y}$ es

$$p(\mathbf{Y} \mid \theta) = \exp\left\{ d(\theta) \sum_{i=1}^{n} T(y_i) - n c(\theta) \right\} \prod_{i=1}^{n} h(y_i) \tag{1.5.6}$$

la cual hace parte de la familia exponencial. ■

Resultado 1.5.3. *Sean Y una variable aleatoria con función de densidad perteneciente a la familia exponencial multiparamétrica. Sea $\boldsymbol{\theta}$ el parámetro de interés con distribución* previa *parametrizada por un vector de hiperparámetros $\boldsymbol{\eta}$ y perteneciente a la familia exponencial multiparamétrica. Entonces la familia exponencial multiparamétrica es conjugada con respecto a sí misma.*

Prueba. En primer lugar, la distribución de probabilidad de Y perteneciente a la familia exponencial multiparamétrica está dada por (1.5.2). Siguiendo el mismo razonamiento de la demostración del Resultado 1.5.1, la distribución *previa* del parámetro de interés debe estar definida de la siguiente manera

$$p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = \exp\left\{ \underbrace{\boldsymbol{\eta}'}_{\mathbf{d}(\boldsymbol{\eta})} \underbrace{\mathbf{d}(\boldsymbol{\theta})}_{\mathbf{T}(\boldsymbol{\theta})} - \underbrace{\ln k(\boldsymbol{\eta}, \delta)}_{c(\boldsymbol{\eta})} \right\} \underbrace{\exp\{-\delta c(\boldsymbol{\theta})\}}_{h(\boldsymbol{\theta})}, \tag{1.5.7}$$

con

$$k(\boldsymbol{\eta}, \delta) = \int \exp\{\boldsymbol{\eta}' \mathbf{d}(\boldsymbol{\theta}) - \delta c(\boldsymbol{\theta})\}\, d\boldsymbol{\theta}.$$

Utilizando el teorema de Bayes, se tiene que, la distribución *posterior* del parámetro θ es

$$\begin{aligned} p(\boldsymbol{\theta} \mid Y) &\propto p(Y \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \\ &= \exp\{\mathbf{T}(y)' \mathbf{d}(\boldsymbol{\theta}) - c(\boldsymbol{\theta}) + \boldsymbol{\eta}' \mathbf{d}(\boldsymbol{\theta}) - \delta c(\boldsymbol{\theta}) - \ln k(\boldsymbol{\eta}, \delta) + \ln h(y)\} \\ &= \exp\left\{ \underbrace{(\boldsymbol{\eta} + \mathbf{T}(y))'}_{\mathbf{d}(y)} \underbrace{\mathbf{d}(\boldsymbol{\theta})}_{\mathbf{T}(\theta)} - \underbrace{[\ln k(\boldsymbol{\eta}, \delta) - \ln h(y)]}_{c(y)} \right\} \underbrace{\exp\{-(\delta + 1) c(\boldsymbol{\theta})\}}_{h(\boldsymbol{\theta})} \end{aligned}$$

La anterior expresión también hace parte de la familia exponencial biparamétrica y con esto se concluye la demostración ■

Nótese que el anterior resultado también cobija situaciones donde la verosimilitud sea perteneciente a la familia exponencial uniparamétrica. Más aún, a cualquier familia exponencial multiparamétrica de orden menor o igual al orden de la distribución *previa*.

Resultado 1.5.4. *Sean $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ una muestra aleatoria con función de densidad conjunta o verosimilitud dada (1.4.4). Bajo este escenario la familia exponencial multi-paramétrica es conjugada con respecto a sí misma.*

Prueba. La demostración sigue los mismos lineamentos que la demostración del

resultado anterior concluyendo que la distribución *posterior* de $\boldsymbol{\theta}$ está dada por

$$
\begin{aligned}
&p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \\
&= \exp\left\{ \sum_{i=1}^{n} \mathbf{T}(y_i)' \mathbf{d}(\boldsymbol{\theta}) - n c(\boldsymbol{\theta}) + \boldsymbol{\eta}' \mathbf{d}(\boldsymbol{\theta}) - \delta c(\boldsymbol{\theta}) - \ln k(\boldsymbol{\eta}, \delta) + \sum_{i=1}^{n} \ln h(y_i) \right\} \\
&= \exp\left\{ \underbrace{\left(\boldsymbol{\eta} + \sum_{i=1}^{n} \mathbf{T}(y_i) \right)'}_{\mathbf{d}(\mathbf{y})} \underbrace{\mathbf{d}(\boldsymbol{\theta})}_{\mathbf{T}(\theta)} - \underbrace{\left[\ln k(\boldsymbol{\eta}, \delta) - \sum_{i=1}^{n} \ln h(y_i) \right]}_{c(\mathbf{y})} \right\} \\
&\times \underbrace{\exp\left\{ -(\delta + n) c(\boldsymbol{\theta}) \right\}}_{h(\boldsymbol{\theta})}
\end{aligned}
$$

La anterior expresión también hace parte de la familia exponencial. ■

Resultado 1.5.5. *Sea Y una variable aleatoria con función de densidad perteneciente a la familia exponencial, dada por (1.5.1). Sea θ el parámetro de interés con distribución* previa *en la familia exponencial biparamétrica. La distribución predictiva* previa *de Y está dada por*

$$
p(Y) = \frac{k(\alpha + T(y), \delta + 1)}{k(\alpha, \delta)} h(y) \tag{1.5.8}
$$

Prueba.

$$
\begin{aligned}
p(Y) &= \int p(\theta) p(Y \mid \theta)\ d\theta \\
&= \int \exp\{\alpha d(\theta) - \ln k(\alpha, \delta) - \delta c(\theta)\} \exp\{d(\theta) T(y) - c(\theta)\} h(y) d\theta \\
&= \frac{h(y)}{k(\alpha, \delta)} \int \exp\{[\alpha + T(y)] d(\theta) - (\delta + 1) c(\theta)\} d\theta \\
&= \frac{k(\alpha + T(y), \delta + 1) h(y)}{k(\alpha, \delta)}
\end{aligned}
$$

donde

$$
k(\alpha, \delta) = \int \exp\{\alpha d(\theta) - \delta c(\theta)\}\ d\theta
$$

y

$$
k(\alpha + T(y), \delta + 1) = \int \exp\{[\alpha + T(y)] d(\theta) - (\delta + 1) c(\theta)\}\ d\theta.
$$

■

Resultado 1.5.6. *Sea $\mathbf{Y} = \{Y_1 \ldots, Y_n\}$ una muestra aleatoria con función de densidad conjunta perteneciente a la familia exponencial, dada por (1.4.4). Sea θ*

el parámetro de interés con distribución previa *dada por (1.4.5). La distribución predictiva* previa *de* $\mathbf{Y}$ *está dada por*

$$p(\mathbf{Y}) = \frac{k(\alpha + T(\mathbf{y}), \delta + n)}{k(\alpha, \beta)} h(\mathbf{y}) \tag{1.5.9}$$

Prueba. La prueba se tiene de inmediato siguiendo los lineamentos de la demostración del anterior resultado. ■

Resultado 1.5.7. *En términos de la distribución predictiva* posterior*, se tiene que para una sola observación* $\tilde{y}$*, ésta está dada por*

$$p(\tilde{y} \mid Y) = \frac{k(\alpha + T(y) + T(\tilde{y}), \delta + 2)}{k(\alpha + T(y), \delta + 1)} h(\tilde{y}) \tag{1.5.10}$$

y en el caso en donde se tiene una muestra aleatoria, entonces la distribución predictiva posterior *para una nueva muestra* $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{n^*}\}$ *de tamaño* n^* *está dada por*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{k(\alpha + T(\mathbf{y}) + T(\tilde{\mathbf{y}}), \delta + n + n^*)}{k(\alpha + T(\mathbf{y}), \delta + n)} h(\tilde{\mathbf{y}}) \tag{1.5.11}$$

Prueba. De la definición de distribución predictiva *posterior* dada por la expresión (1.4.9) se tiene que

$$\begin{aligned} p(\tilde{y} \mid Y) &= \int p(\tilde{y} \mid \theta) p(\theta \mid y) \, d\theta \\ &= \int \exp\{d(\theta)T(\tilde{y}) - c(\theta)\} h(\tilde{y}) \frac{\exp\{[\alpha + T(y)]d(\theta) - (\delta + 1)c(\theta)\}}{k(\alpha + T(y), \delta + 1)} \, d\theta \\ &= \frac{h(\tilde{y})}{k(\alpha + T(y), \delta + 1)} \int \exp\{[\alpha + T(y) + T(\tilde{y})]d(\theta) - (\delta + 2)c(\theta)\} \, d\theta \\ &= \frac{k(\alpha + T(y) + T(\tilde{y}), \delta + 2)}{k(\alpha + T(y), \delta + 1)} h(\tilde{y}), \end{aligned}$$

con

$$k(\alpha + T(y) + T(\tilde{y}), \delta + 2) = \int \exp\{[\alpha + T(y) + T(\tilde{y})]d(\theta) - (\delta + 2)c(\theta)\} \, d\theta.$$

La demostración para la nueva muestra se lleva a cabo de manera análoga. ■

1.5.2 Distribuciones previa no informativas

Cuando no existe una base poblacional sobre el parámetro de interés o cuando existe total ignorancia de parte del investigador acerca del comportamiento de probabilístico del parámetro, es necesario definir distribuciones *previa* que sean no informativas. Es decir, definir distribuciones *previa* que jueguen un papel mínimo en términos de influencia en la distribución *posterior*. Una característica de estas

distribuciones es que su forma es vaga, plana o difusa, cumpliendo así el objetivo de no influenciar a la distribución *posterior*. Por tanto la pregunta de interés que surge en este instante es: ¿cómo seleccionar distribuciones *previa* no informativas[4] sobre el parámetro de interés?

En los anteriores términos, la distribución uniforme define una distribución *previa* que cumple con las características de no información en la mayoría de escenarios. Específicamente en aquellos problemas en donde el parámetro de interés está limitado a un espacio de muestreo. Por ejemplo, en la distribución Binomial, el parámetro de interés está limitado al espacio de muestreo $[0, 1]$. Sin embargo, no en todos los problemas encaja la distribución uniforme. Nótese, por ejemplo, que en el caso en que la distribución exponencial se acomode a los datos como candidata a verosimilitud, entonces el espacio de muestreo del parámetro de interés estaría dado por $(0, \infty)$ en cuyo caso la distribución uniforme no sería conveniente puesto que sería una distribución impropia en el espacio de muestreo del parámetro de interés. Es decir

$$\text{si } p(\theta) \propto kI_{\Theta}(\theta), \text{ entonces } \int_{\Theta} p(\theta)\ d(\theta) \longrightarrow \infty.$$

Por otro lado, una característica importante que debe tener una distribución *previa* no informativa es que sea invariante en términos de transformaciones matemáticas. Es decir, si el parámetro de interés es θ con distribución *previa* no informativa dada por $p(\theta)$, y sea $\phi = h(\theta)$ una función de θ, entonces la distribución *previa* de ϕ debería ser no informativa. Sin embargo, la teoría de probabilidad afirma que la distribución de probabilidad de una transformación está dada por

$$p(\phi) = p(\theta) \mid \frac{d\theta}{d\phi} \mid = p(\theta) \mid h'(\theta) \mid^{-1} \tag{1.5.12}$$

y claramente si la función h no es una función lineal, entonces los resultados encontrados por medio de este enfoque indicarían que la distribución *previa* $p(\phi)$ sería informativa contradiciendo los supuestos de $p(\theta)$.

Ejemplo 1.5.1. *Suponga que el parámetro de interés es θ y que está restringido a un espacio de muestreo dado por el intervalo $[0, 1]$. Si se supone completa ignorancia acerca del comportamiento del parámetro, entonces una buena opción, con respecto a la distribución* previa*, sería la distribución uniforme en el intervalo $[0, 1]$. Es decir, la distribución* previa *no informativa estaría dada por*

$$p(\theta) = I_{[0,1]}(\theta)$$

Suponga ahora que existe una transformación del parámetro de interés dada por $\phi = h(\theta) = \ln(\theta)$. Por tanto, siguiendo (1.5.12) se tiene que la distribución de ϕ

[4] Existen muchas denominaciones para las distribuciones uniformes que no son informativas. Por ejemplo, Box Tiao proponen el nombre de distribuciones localmente uniformes para asegurar que cumplan con las condiciones de función de densidad de probabilidad en un rango particular del espacio paramétrico. Sin embargo, en este texto vamos a utilizar la expresión ≪no informativa≫ al referirse a este tipo de distribuciones a previa.

estaría dada por

$$p(\phi) = I_{(-\infty,0)}(\phi)e^{\phi}$$

la cual es informativa con respecto al parámetro ϕ. Sin embargo, es el mismo problema y existe una contradicción en términos de que para θ se desconoce todo, pero para una función ϕ existe evidencia de que el parámetro se comporta de cierta manera.

Ejemplo 1.5.2. *Buscar otro ejemplo donde la uniforme no sirva como a previa y explicar XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*

Para palear las anteriores diferencias, es necesario encontrar una distribución *previa* no informativa que sea invariante a transformaciones matemáticas. La distribución *previa* no informativa de Jeffreys, definida a continuación, cuenta con esta agradable propiedad.

DefiniciÃ³n1.5.2. *Si la verosimilitud de los datos está determinada por un único parámetro θ, la distribución* previa *no informativa de Jeffreys tiene distribución de probabilidad dada por*

$$p(\theta) \propto (I(\theta))^{1/2} \tag{1.5.13}$$

con $I(\theta)$ la información de Fisher definida como

$$\begin{aligned} I(\theta) &= E\left\{\left[\frac{\partial}{\partial\theta}\log p(\mathbf{Y} \mid \theta)\right]^2\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{Y} \mid \theta)\right\} \end{aligned}$$

Si la verosimilitud de los datos está determinada por un vector de parámetros $\boldsymbol{\theta}$, la distribución previa *no informativa de Jeffreys tiene distribución de probabilidad dada por*

$$p(\theta) \propto |\mathbf{I}(\theta)|^{1/2} \tag{1.5.14}$$

donde $\mathbf{I}$ es la matriz de información de Fisher definida como

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta}) &= E\left\{\left[\frac{\partial}{\partial\theta}\log p(\mathbf{Y} \mid \theta)\right]^2\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{Y} \mid \theta)\right\} \end{aligned}$$

Nótese que si la verosimilitud de las observaciones pertenecen a la familia de distribuciones exponencial, entonces la distribución a previa de Jeffreys no es difícil de calcular. Por otro la do nótese que la distribución a previa no informativa de Jeffreys depende, de cierta manera, del mecanismo probabilístico que rige a los datos. Lo anterior hace que ciertos críticos de la estadística bayesiana critiquen este enfoque puesto que se supone que la formulación de la distribución a previa es independiente de los datos observados.

Resultado 1.5.8. *La distribución* previa *no informativa de Jeffreys es invariante a transformaciones uno a uno. Es decir, si* $\phi = h(\theta)$*, entonces* $p(\phi) \propto (I(\phi))^{1/2}$*.*

Prueba. En primer lugar nótese que

$$I(\theta) = \mathbf{J}(\phi) \mid \frac{d\phi}{d\theta} \mid^2$$

puesto que al utilizar la regla de la cadena del cálculo matemático se tiene que

$$\begin{aligned}
\mathbf{J}(\phi) = -E\left[\frac{d^2 \log p(\mathbf{Y} \mid \phi)}{d\phi^2}\right] &= -E\left[\frac{d}{d\phi}\left(\frac{d \log p(\mathbf{Y} \mid \phi)}{d\phi}\right)\right] \\
&= -E\left[\frac{d}{d\theta}\left(\frac{d \log p(\mathbf{Y} \mid \phi)}{d\phi}\right) \mid \frac{d\theta}{d\phi} \mid\right] \\
&= -E\left[\frac{d^2 \log p(\mathbf{Y} \mid \phi)}{d\theta^2} \mid \frac{d\theta}{d\phi} \mid^2\right] \\
&= -E\left[\frac{d^2 \log p(\mathbf{Y} \mid \theta = h^{-1}(\phi))}{d\theta^2} \mid \frac{d\theta}{d\phi} \mid^2\right] \\
&= I(\theta) \mid \frac{d\theta}{d\phi} \mid^2
\end{aligned}$$

Ahora, de la definición de función de distribución para una función y utilizando (1.4.11), se tiene que

$$p(\phi) = p(\theta) \mid \frac{d\theta}{d\phi} \mid \propto (I(\theta))^{1/2} \mid \frac{d\theta}{d\phi} \mid \propto I(\phi)^{1/2} \mid \frac{d\phi}{d\theta} \mid\mid \frac{d\theta}{d\phi} \mid = I(\phi)^{1/2}$$

■

En **?**, p. 59 citan una Tabla de resumen en donde se encuentran distribuciones a previa no informativas para las distribuciones probabilísticas más comunes. A continuación se exponen dos ejemplos que utilizan este enfoque.

Ejemplo 1.5.3. *Si* Y *es una variable aleatoria con distribución Binomial, entonces el espacio de muestreo del parámetro de interés será el intervalo* $[0, 1]$*; Sería conveniente utilizar la función de distribución uniforme sobre este intervalo como distribución* previa *no informativa. Con el enfoque de Jeffreys se llega a este mismo resultado puesto que la información de Fisher para la distribución binomial es* $J(\theta) = n/\theta(1-\theta)$ *dado que*

$$\log p(Y \mid \theta) = \log \binom{n}{y} + y \log(\theta) + (n - y) \log(1 - \theta)$$

y

$$\frac{d^2 \log p(Y \mid \theta)}{d\theta^2} = -\frac{y}{\theta^2} - \frac{n - y}{(1 - \theta)^2}$$

Por lo tanto al calcular la esperanza, y por consiguiente la información de Fisher, se tiene que

$$I(\theta) = -E\left[\frac{d^2 \log p(Y \mid \theta)}{d\theta^2}\right] = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Es decir, la distribución previa *no informativa para el parámetro de interés es proporcional a* $\theta^{-1/2}(1-\theta)^{-1/2}$*, la cual comparte la misma forma estructural de una distribución* $beta(1/2, 1/2)$ *que a su vez es idéntica a la distribución uniforme. En términos de la distribución* posterior *para el parámetro de interés, se tiene que*

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta)p(\theta) \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{-1/2}(1-\theta)^{-1/2} \\ &= \theta^{x+1/2-1}(1-\theta)^{n-x+1/2-1} \end{aligned}$$

Por tanto, la distribución de $\theta \mid Y$ *es* $Beta(x+1/2, n-x+1/2)$*. Por construcción, esta distribución no está alterada ni influenciada por la distribución* previa *pues la misma es no informativa.*

Ejemplo 1.5.4. *Si* $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ *es una muestra aleatoria de variables con distribución de Poisson, entonces el espacio de muestreo del parámetro de interés será el intervalo* $(0, \infty)$*; por tanto utilizar la distribución uniforme como distribución* previa *no informativa no es conveniente. Ahora, la información de Fisher para la distribución conjunta es* $I(\theta) = n/\theta$ *puesto que*

$$\log p(\mathbf{Y} \mid \theta) = -n\theta + \log(\theta)\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!)$$

y

$$\frac{d^2 \log p(\mathbf{Y} \mid \theta)}{d\theta^2} = -\frac{\sum_{i=1}^{n} y_i}{\theta^2}$$

Por lo tanto al calcular la esperanza, y por consiguiente la información de Fisher, se tiene que

$$I(\theta) = -E\left[\frac{d^2 \log p(\mathbf{Y} \mid \theta)}{d\theta^2}\right] = \frac{\sum_{i=1}^{n} E(y_i)}{\theta^2} = \frac{n}{\theta}$$

Es decir, la distribución previa *no informativa para el parámetro de interés es proporcional a* $\theta^{-1/2}$*. En términos de la distribución* posterior *para el parámetro de interés, se tiene que*

$$p(\theta \mid Y) \propto p(Y \mid \theta)p(\theta) \propto e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i}\theta^{-1/2} = e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i - 1/2}$$

Por tanto, la distribución de $\theta \mid \mathbf{Y}$ *es* $Gamma(\sum_{i=1}^{n} y_i + 1/2, n)$*. Por construcción, esta distribución no está alterada ni influenciada por la distribución* previa *pues la misma es no informativa.*

Ejemplo 1.5.5. *Suponga que* $\mathbf{Y} = \{Y_1 \ldots, Y_n\}$ *es una muestra aleatoria con distribución normal de parámetros* $(\theta, \sigma^2)'$. *Del ejemplo E.1.2, se tiene que la matriz de información de Fisher para el vector de parámetros está dada por*

$$\begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{1.5.15}$$

Por lo tanto, la distribución a previa no informativa de Jeffreys está dada por

$$p(\theta, \sigma^2) \propto 1/\sigma^6 \tag{1.5.16}$$

1.6 pruebas de hipótesis

A excepción del juzgamiento de hipótesis, las inferencias que hacen los estadísticos bayesianos, acerca de poblaciones normales, son muy similares a las que los estadísticos de la tradición frecuentista, de Neyman y Pearson, hacen. Consideremos la siguiente situación. Un instrumento mide la posición de un objeto con un determinado error. Éste error está distribuido de manera uniforme en el intervalo (-1cm, 1cm). Supongamos que el instrumento midió la posición de un objeto en +0.9999cm del origen. Planteamos la siguiente hipótesis nula, H: La posición real del objeto es exactamente el origen. Imagine que planteamos este problema de inferencia estadística a los profesores López (frecuentista clásico) y Cepeda (acérrimo bayesiano). Razonamiento del frecuentista: Si la hipótesis nula es verdadera, ha ocurrido un evento con una probabilidad (a dos colas) de ocurrencia de 0.0001 o menos. Mediante un criterio razonable (nivel de significación), este es un evento muy raro y por lo tanto rechaza H. Razonamiento del bayesiano: El bayesiano ve las cosas desde un punto de vista diferente. Dada una observación, la verosimilitud asociada con la posición del objeto en el intervalo -0.0001 y +1.9999 es la misma, 0.5. Fuera de esos límites la verosimilitud es nula. Ahora, el origen está dentro de la región en donde la verosimilitud es máxima; por lo tanto sea cual sea la distribución a previa asociada al parámetro de posición, la distribución a posterior tomara el valor cero en cualquier lugar fuera del intervalo -0.0001 y +1.9999. Así, con la observación disponible, no hay evidencia para el rechazo de H. Bajo esta paradoja, Brewer (2002) sugiere que ambos estadísticos tienen razón, pero a la vez están equivocados. El frecuentista tiene razón en afirmar que, con la evidencia disponible, ha ocurrido un evento extraordinariamente extraño o que la hipótesis nula es falsa. El bayesiano tiene razón en argumentar que, en términos de la situación, no hay evidencia en contra de la hipótesis nula. Esta paradoja se presenta porque los bayesianos tienden a trabajar dentro de la situación que ellos creen que existe (o al menos creen que ellos creen que existe) y la lógica bayesiana se mueve en ese marco de referencia. Los bayesianos hacen las inferencias en términos de la verosimilitud de los eventos observados, mientras que los frecuentistas hacen inferencias en términos de eventos que ni siquiera han ocurrido. .

1.6.1 Factor de Bayes

1.6.2 P-valor bayesiano

1.7 Criterios de información

Criterio DIC

El criterio de información de devianza (denotada por DIC por los iniciales en inglés) es una generalización del popular criterio AIC para los modelos jerárquicos, y se basa en el concepto de la devianza que se define como

$$D(y, \boldsymbol{\theta}) = -2 * \log(p(y|\boldsymbol{\theta})) \tag{1.7.1}$$

cuya media posterior es una medida usual del ajuste del modelo. **?** sugirió graficar la distribución posteriori de la devianza para observar el ajuste del modelo a los datos. Una estimación de esta media posterior está dada por

$$\hat{E}_D = \frac{1}{M} \sum_{m=1}^{M} D(y, \boldsymbol{\theta}^m)$$

En donde $\boldsymbol{\theta}^m$ es un valor simulado de la distribución posterior de $\boldsymbol{\theta}$ y M es la longitud de la cadena generada. Dado lo anterior, el DIC se define como

$$DIC = \hat{E}_D + p_D$$

Donde pD es el número efectivo de parámetros. Nótese que en la anterior formulación, el DIC se puede descomponer en dos partes: la parte de la bondad de ajuste del modelo, medido a través de E_D, y la parte que mide la complejidad del modelo pD. Otra formulación equivalente del DIC se obtiene teniendo en cuenta que

$$p_D = \hat{E}_D - \hat{D}$$

Donde $\hat{D} = -2 * \log(p(y|\hat{\boldsymbol{\theta}}))$ con $\hat{\boldsymbol{\theta}}$ denotando la media posterior de $\boldsymbol{\theta}$; es decir, $\hat{D}$ es la estimación de la devianza usando $\hat{\boldsymbol{\theta}}$, y p_D se puede ver como la mediposterior de la devianza menos la devianza de las medias posterior **?**. De esta forma, el DIC también se puede escribir como

$$DIC = \hat{D} + 2p_D$$

Interpretación de DIC: El modelo con el menor DIC es considerado como el modelo que mejor predice un conjunto de datos con la misma estructura que los datos observados. Al respecto se deben tener en cuenta las siguientes consideraciones:

- El DIC puede ser negativo puesto que $p(y|\theta)$ puede tomar valores mayores a 1 asociado a una devianza pequeña.
- pD, y por consiguiente DIC, no es invariante a parametrizaciones del modelo. Se sugiere en la práctica usar parametrizaciones que conducen a la normalidad en lposterior.

Criterio BIC

El criterio de información bayesiano BIC, también conocido como el criterio de Schwarz **?**, se define como

$$BIC = -2\log(p(y|\hat{\boldsymbol{\theta}})) + p\log(n)$$

Donde p es el número de parámetros en el modelo y n el número de datos observados. Cabe resaltar que en el criterio BIC hay una mayor penalización por el número excesivo de parámetros que en el criterio AIC, y en la práctica se prefieren los modelos con un BIC menor.

Nota: Se debe recalcar que los dos criterios tienen diferentes enfoques, el criterio BIC se enfoca en identificar el modelo verdadero, mientras que el criterio DIC enfoca en encontrar el modelo con mejor capacidad de predicción.

1.8 Acerca de la notación

Antes de empezar las próximas secciones, es necesario revisar la notación que se seguirá de ahora en adelante. Del teorema de Bayes resultan tres grandes definiciones que constituyen la base de la estadística Bayesiana y que a lo largo de este texto se mencionarán diferenciandolas por medio de la notación. El símbolo más importante de la estadística matemática es p, el cual indica que existe una distribución de probabilidad para los datos, para el vector de parámetros, condicional o no. De hecho todos las definiciones y resultados anteriores han estado supeditadas al uso de esta monótona notación. En el ámbito de la notación de investigación internacional es común diferenciar las distribuciones con el fin de hacer más ameno el estudio del enfoque Bayesiano. En este texto se seguirá esta distinción.

Andrew Gelman comenta que John Cook comenta que,

Existe sólo un símbolo importante en estadística, p. El mismo símbolo representa todo. Simplemente se debe usarlo y darse cuenta cuál p es cual, pues pueden provenir de diferentes contextos... Un ejemplo claro en donde p representa cuatro funciones distintas en una sola ecuación es el siguiente:

$p(\theta \mid x) = p(x \mid \theta)\frac{p(\theta)}{p(x)}$

Usualmente, la regla de Bayes no requiere de mucha explicación. Sin embargo, en la anterior ecuación cada p denota funciones que son totalmente diferentes, aunque compartan el mismo símbolo. Los autores prefieren este tipo de escritura para evadir la notación engorrosa que requeriría la escritura completamente explícita.

A veces, la sobrecarga de la decimonovena letra del alfabeto se convierte en un lastre y los estadísticos cambian de notación y de alfabeto y usan la contraparte griega π. Aunque esto a veces hace las cosas un poco más confusas.

En su libro, Bayesian Data Analysis, Andrew Gelman, explica por qué la notación simple, con el uso (a veces abuso) de la letra p es más rigurosa de lo que, a simple vista, pueda parecer y comenta que,

En realidad no me gusta la notación que la mayoría de los estadísticos usan... f, para distribuciones de muestreo, π, para distribuciones a previa y L, para verosimilitudes. Este estilo de notación se desvía de lo que realmente es importante. La notación no debería depender del orden en que las distribuciones son especificadas. Todas ellas son distribuciones de probabilidad, eso es lo realmente importante.

Esto tiene sentido, aún más cuando se estudian las propiedades estadísticas de los estimadores desde el punto de vista de la teoría de la medida. Siendo así, el símbolo p se refiere a una notación para una medida de probabilidad, quizás inducida por un elemento aleatorio. De hecho, en la ecuación que determina la regla de Bayes, cada una de las p son medidas de probabilidad que no comparten el mismo espacio de medida (ni la misma σ-álgebra, ni el mimo espacio muestral).

De hecho, todo queda claro al realizar un diagrama que permita ver el espacio de salida y el espacio de llegada de los elementos aleatorios que inducen (si es el caso), cada una de las distribuciones de probabilidad. Por otra parte, Bob Carpenter, concluye que

[Una vez resuelto el problema de identificación de los espacios] la notación estadística depende en gran manera del contexto y aunque la regla de Bayes no necesite de mucha explicación, es necesario conocerlo todo acerca del contexto para poder interpretar las funciones que la conforman... El problema se hace mucho más agudo para los estadísticos novatos, pero eso se resuelve con la práctica. Una vez que uno sabe lo que está haciendo, se vuelve obvia la referencia de la distribución p.

Por lo anterior, es natural que algunos de los textos clásicos de estadística matemática, parezcan olvidar el contexto de las diferentes medidas de probabilidad. En realidad no es que lo olviden, lo que pasa es que los autores no son novatos y asumen que el lector sigue la idea de la referencia de la p en cuestión. Sin embargo, y lo digo por mi y sólo por mí, sería mejor que no asumieran esa idea. De esta manera, el estudio de estos textos sería un poco menos denso.

2 Modelos uniparamétricos

Los modelos que están definidos en términos de un solo parámetro que pertenece al conjunto de los números reales se definen como modelos uniparamétricos. Este capítulo estudia modelos, discretos y continuos, que son comunes de implementar en la práctica. Dado que todos ellos son inducidos por familias de probabilidad conjugadas, entonces las estimaciones posteriores para los parámetros pueden hallarse sin necesidad de sofisticaciones computacionales. Es decir, con el uso de una simple calculadora de bolsillo, es posible realizar inferencia bayesiana propiamente dicha. Por lo tanto, en este capítulo, será menor el uso de software estadístico. Sin embargo, para cada modelo se incluye la sintaxis de `JAGS`, para un ejemplo práctico que permite la familiarización e interiorización del ambiente computacional de este software que será indispensable en el desarrollo de capítulos posteriores.

2.1 Bernoulli

Suponga que Y es una variable aleatoria con distribución Bernoulli, su distribución está dada por

$$p(Y \mid \theta) = \theta^y (1-\theta)^{1-y} I_{\{0,1\}}(y), \tag{2.1.1}$$

Como el parámetro θ está restringido al espacio $\Theta = [0,1]$, entonces es posible formular varias opciones para la distribución previa del parámetro. En particular, la distribución uniforme restringida al intervalo $[0,1]$ o la distribución Beta parecen ser buenas opciones. Dado que la distribución uniforme es un caso particular de la distribución Beta, entonces vamos a trabajar con ésta. Por lo tanto la distribución previa del parámetro θ está dada por

$$p(\theta \mid \alpha, \beta) = \frac{1}{Beta(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} I_{[0,1]}(\theta). \tag{2.1.2}$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.1.1. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid Y \sim Beta(y+\alpha, \beta - y + 1)$$

Prueba.

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \frac{I_{\{0,1\}}(y)}{Beta(\alpha,\beta)} \theta^y \theta^{\alpha-1}(1-\theta)^{\beta-1}(1-\theta)^{1-y} I_{[0,1]}(\theta) \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{\beta-y+1-1} I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Beta(y + \alpha, \beta - y + 1)$. ■

Del anterior resultado, podemos ver que la familia de distribución Beta es conjugada con respecto a la familia de distribución Bernoulli. Ahora consideramos cuál sería la distribución previa no informativa de Jeffreys para el parámetro θ. De acuerdo a la Definición 1.5.2, tenemos que

$$p(\theta) \propto I(\theta)^{1/2}$$

donde $I(\theta)$ es la información de Fisher, que en este caso está dada por

$$\begin{aligned} I(\theta) &= -E\left\{\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{Y} \mid \theta)\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta^2}\{Y\log\theta + (1-Y)\log(1-\theta)\}\right\} \\ &= E\left\{\frac{Y}{\theta^2} + \frac{1-Y}{(1-\theta)^2}\right\} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

De esta forma, tenemos que la distribución previa no informativa de Jeffreys debe ser proporcional a $\theta^{-1/2}(1-\theta)^{-1/2}$, el cual corresponde a la distribución $Beta(1/2, 1/2)$ cuya función de densidad se muestra en la figura **??** la cual asigna iguales pesos a los valores extremos del parámetro de interés y la no informatividad se representa en la simetría de la función alrededor del valor 0.5.

Resultado 2.1.2. *La distribución predictiva previa para una observación y está dada por*

$$p(Y) = \frac{Beta(y+\alpha, \beta-y+1)}{Beta(\alpha,\beta)} I_{\{0,1\}}(y), \tag{2.1.3}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(Y) &= \int p(Y \mid \theta)p(\theta \mid \alpha, \beta)\, d\theta \\ &= \int_0^1 \theta^y(1-\theta)^{1-y} I_{\{0,1\}}(y) \frac{1}{Beta(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta \\ &= \frac{Beta(y+\alpha,\beta-y+1)}{Beta(\alpha,\beta)} I_{\{0,1\}}(y) \int_0^1 \frac{\theta^{y+\alpha-1}(1-\theta)^{\beta-y+1-1}}{Beta(y+\alpha,\beta-y+1)}\, d\theta \\ &= \frac{Beta(y+\alpha,\beta-y+1)}{Beta(\alpha,\beta)} I_{\{0,1\}}(y) \end{aligned}$$

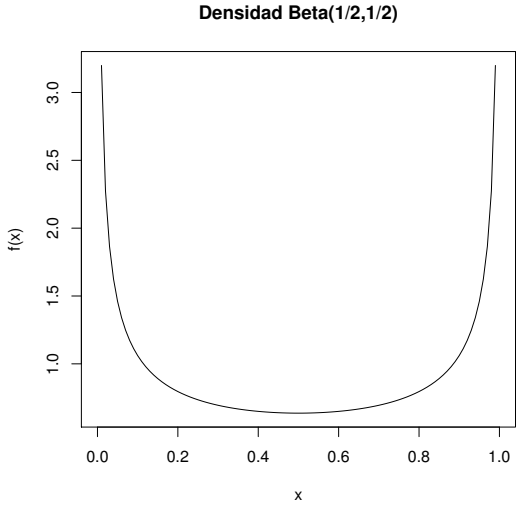

Figura 2.1: *Distribución previa no informativa de Jeffreys para el parámetro de una distribución Bernoulli*

Nótese que en la anterior demostración, la integral al lado derecho de la tercera igualdad es igual a la unidad, puesto que la expresión matemática dentro de la integral corresponde a la función de densidad de una variable aleatoria con distribucion $Beta$, que tiene rango en el intervalo $(0, 1)$. Por otro lado se deben verificar las dos condiciones de función de densidad. Es decir

1. $p(Y) > 0$ $(\forall y \in Y)$. Esta condición se tiene trivialmente puesto que la función matemática Beta siempre toma valores positivos.
2. $\int p(y)\ dx = 1$. En este caso, esta función es discreta definida en el conjunto $\{0, 1\}$. Por lo tanto esta condición es equivalente a

$$\sum_{y\in\{0,1\}} P(Y = y) = \sum_{y\in\{0,1\}} \frac{Beta(y + \alpha, \beta - y + 1)}{Beta(\alpha, \beta)} = 1$$

   Lo cual se verifica fácilmente teniendo en cuenta las propiedades de la función matemática Beta y de la función matemática Gamma.

■

La distribución predictiva dada en **??** está basada únicamente en la distribución previa del parámetro θ, una vez observada la variable Y se puede pensar en actualizar la distribución predictiva basando en la distribución posterior del parámetro, esta distribución se da en el siguiente resultado.

Resultado 2.1.3. *Después de la recolección de los datos, la distribución predictiva*

posterior para una nueva observación $\tilde{y}$ está dada por

$$p(\tilde{y} \mid Y) = \frac{Beta(\tilde{y} + y + \alpha, \beta - \tilde{y} - y + 2)}{Beta(y + \alpha, \beta - y + 1)} I_{\{0,1\}}(\tilde{y}), \tag{2.1.4}$$

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(\tilde{y} \mid Y) &= \int p(\tilde{y} \mid \theta) p(\theta \mid Y)\, d\theta \\ &= \int_0^1 \theta^{\tilde{y}} (1-\theta)^{1-\tilde{y}} I_{\{0,1\}}(\tilde{y}) \frac{\theta^{y+\alpha-1}(1-\theta)^{\beta-y+1-1}}{Beta(y+\alpha, \beta-y+1)}\, d\theta \\ &= \frac{Beta(\tilde{y} + y + \alpha, \beta - \tilde{y} - y + 2)}{Beta(y + \alpha, \beta - y + 1)} I_{\{0,1\}}(\tilde{y}) \\ &\qquad \times \int_0^1 \frac{\theta^{\tilde{y}+y+\alpha-1}(1-\theta)^{\beta-\tilde{y}-y+2-1}}{Beta(\tilde{y} + y + \alpha, \beta - \tilde{y} - y + 2)}\, d\theta \\ &= \frac{Beta(\tilde{y} + y + \alpha, \beta - \tilde{y} - y + 2)}{Beta(y + \alpha, \beta - y + 1)} I_{\{0,1\}}(\tilde{y}) \end{aligned}$$

■

Ahora, en la práctica rara vez se observa la realización de una única variable aleatoria Bernoulli Y, sino una muestra de variables aleatorias $Y_1, \cdots, Y_n$. En este caso, la distribución posterior del parámetro θ está dada en el siguiente resultado.

Resultado 2.1.4. *Cuando se tiene una muestra aleatoria $Y_1, \ldots, Y_n$ de variables con distribución Bernoulli de parámetro θ, entonces la distribución posterior del parámetro de interés es*

$$\theta \mid Y_1, \ldots, Y_n \sim Beta\left(\sum_{i=1}^n y_i + \alpha, \beta - \sum_{i=1}^n y_i + n\right)$$

La demostración se deja como ejercicio.

Ejemplo 2.1.1. *Es común en muchos países del mundo que se presenten encuestas de opinión electoral unas semanas antes de las elecciones presidenciales. Dentro de este tipo de encuestas se acostumbra a indagar acerca del favoritismo de los candidatos involucrados en la contienda electoral. Suponga que un candidato presidencial llamado José Pérez está interesado en conocer su intención de voto previa a las elecciones. Para esto, él contrata a una firma encuestadora para la realización de un muestreo probabilístico entre la población votante. El resultado de este estudio puede hacer cambiar o afirmar las estrategias publicitarias y la redefinición de la campaña electoral. La firma encuestadora decide implementar una estrategia de muestreo con un tamaño de muestra de doce mil personas. A cada respondiente se le realiza la siguiente pregunta:* ***Si las elecciones presidenciales fueran mañana. ¿Usted votaría por el candidato José Pérez?***

Las respuestas a esta pregunta son realizaciones de una muestra aleatoria de doce mil variables con densidad Bernoulli. Los resultados del estudio arrojan que

6360 personas de las personas entrevistadas, es decir un 53 por ciento, votarían por el suscrito candidato. Técnicamente se debe analizar esta cifra puesto que las implicaciones de ganar en una primera vuelta son grandes en el sentido económico, logístico y administrativo. Claramente, el dato 53 por ciento asegura una ventaja dentro de la muestra de doce mil personas. Sin embargo, es necesario realizar un estudio más profundo acerca de la caracterización estructural de la intención de voto del candidato en la población de todos los votantes.

Con base en lo anteriormente expuesto, se decide utilizar la inferencia bayesiana puesto que existe información previa de un estudio anterior, contratado por el mismo candidato unos meses atrás en donde se entrevistaron a mil personas, con un favoritismo que estaba alrededor del 35 por ciento. Esta situación conlleva a la utilización de la metodología bayesiana que incorpora la información pasada acerca del mismo fenómeno.

El estadístico de la firma encuestadora decide utilizar una distribución previa[1] $Beta(\alpha = 350, \beta = 650)$. *Utilizando el resultado 2.1.4, se contempla que la distribución posterior del parámetro de interés, que representa la probabilidad de éxito en las elecciones presidenciales, es* $Beta(6360 + 350, 650 - 6360 + 12000) = Beta(6710, 6290)$. *Por lo tanto, utilizando la distribución posterior, se estima que la intención de voto por el candidato es de* $\frac{6710}{6710+6290} = \frac{6710}{13000} = 0.516$ *y este valor equivale a la media de la distribución posterior. Este mismo análisis puede ejecutarse en* `JAGS`*, mediante el uso del siguiente código computacional*

Sin embargo, si no se tuviese información previa como la suministrada por el estudio de meses anteriores, el análisis bayesiano sugeriría trabajar con una distribución previa no informativa, que en este caso, correspondería a una $Beta(\alpha = 0.5, \beta = 0.5)$. *siguiendo el mismo análisis, se tiene que la distribución posterior es* $Beta(6360.5, 5640.5)$. *Finalmente, se estimaría que la intención de voto por el candidato es de* $\frac{6350.5}{12001} = 0.529$. *Las gráficas* **??** *y* **??** *muestran el comportamiento de las distribuciones previas y posteriores en ambos escenarios. Nótese que la distribución no informativa influye muy poco en el comportamiento de la distribución posterior.*

Utilizando el siguiente código en R, es posible conocer los intervalos de credibilidad para las dos distribuciones posteriores. Es posible concluir que en ambos escenarios, el candidato aventaja significativamente a sus contrincantes y, salvo algún cambio drástico en el comportamiento del electorado, ganará las elecciones. Lo anterior se deduce puesto que el intervalo de credibilidad al 95 % no contiene ningún valor menor a 0.5

```
> qbeta(c(0.025, 0.975),6710,6290)
[1] 0.5075614 0.5247415
> qbeta(c(0.025,0.975),6350.5,5640.5)
[1] 0.5206678 0.5385340
```

[1] Como se verá más adelante, es conveniente definir los parámetros de la distribución previa como α igual al número de votantes a favor y β igual al número de votantes en contra.

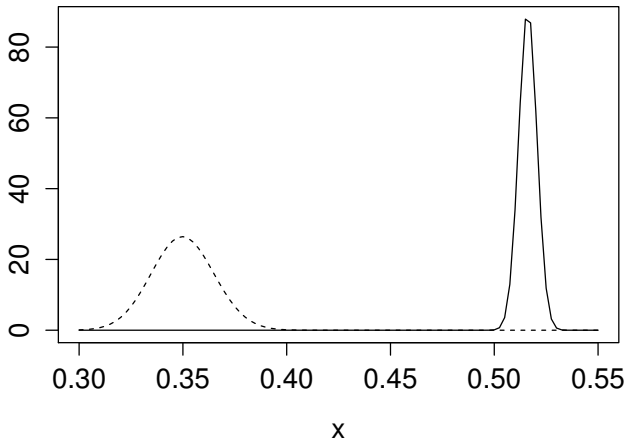

Figura 2.2: *Distribución previa informativa (línea punteada) y distribución posterior (línea continua) para el ejemplo de las encuestas electorales.*

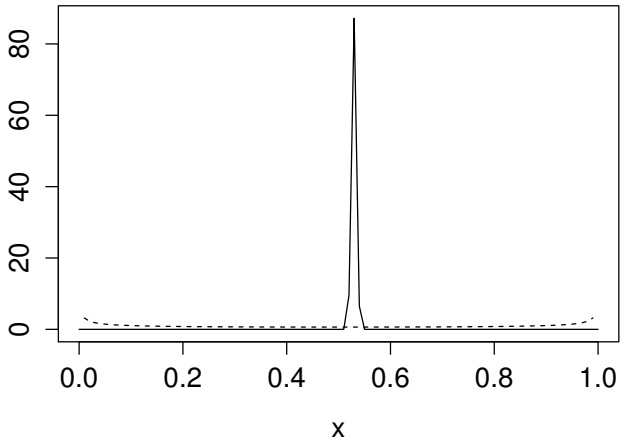

Figura 2.3: *Distribución previa no informativa (línea punteada) y distribución posterior (línea continua) para el ejemplo de las encuestas electorales.*

Por otro lado, el siguiente código en JAGS permite obtener el mismo tipo de inferencia creando tres cadenas de Markov cuya distribución de probabilidad coincide

con la distribución posterior del ejemplo.

```
#datos
y<-c(1, 0, 1,..., 0)
n<-length(y)

#modelo jags Bernoulli
bern.model <-function() {
for(i in 1:n)
{
y[i]~dbern(theta)
}
theta~dbeta(350, 650)#información previa
}

bern.data <- list("y","n")
bern.param <- c("theta")
bern.inits <- function(){list("theta"=c(0.5))}
set.seed(123)

bern.fit <- jags(data=bern.data, inits=bern.inits, bern.param,
                 n.chains=3, n.iter=10000, n.burnin=1000,
                 n.thin=10, model.file=bern.model)

print(bern.fit)
```

2.2 Binomial

Cuando se dispone de una muestra aleatoria de variables con distribución Bernoulli $Y_1, \ldots, Y_n$, la inferencia bayesiana se puede llevar a cabo usando la distribución Binomial, puesto que es bien sabido que la suma de variables aleatorias Bernoulli

$$S = \sum_{i=1}^{n} Y_i$$

sigue una distribución Binomial. Es decir:

$$p(S \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s} I_{\{0,1,\ldots,n\}}(s), \tag{2.2.1}$$

Nótese que la distribución binomial es un caso general para la distribución Bernoulli, cuando $n = 1$. Entonces, así como en la distribución Bernoulli, el parámetro θ está restringido al espacio $\Theta = [0, 1]$. Luego, es admisible proponer que θ siga una distribución Beta. Por tanto la distribución previa del parámetro θ está dada por la expresión (2.1.2). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.2.1. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid S \sim Beta(s+\alpha, \beta - s + n)$$

Prueba.

$$\begin{aligned} p(\theta \mid S) &\propto p(S \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \frac{\binom{n}{s} I_{\{0,1,\ldots,n\}}(s)}{Beta(\alpha, \beta)} \theta^s \theta^{\alpha-1} (1-\theta)^{\beta-1} (1-\theta)^{n-s} I_{[0,1]}(\theta) \\ &\propto \theta^{s+\alpha-1} (1-\theta)^{\beta-s+n-1} I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se llega a una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Beta(s+\alpha, \beta - s + n)$.

■

Del resultado anterior podemos ver que el estimador bayesiano de θ está dada por la media de la distribución posterior, dada por

$$\hat{\theta}_B = \frac{s+\alpha}{n+\alpha+\beta} \tag{2.2.2}$$

En la práctica, se acostumbra a escoger los hiperparámetros α y β de tal forma que correspondan al número de éxitos y fracasos obtenidos en los datos previa, respectivamente. De esta forma, $\hat{\theta}_P = \alpha/(\alpha+\beta)$ corresponde a la estimación previa del parámetro θ. Por otro lado, el estimador clásico de θ está dado por $\hat{\theta}_C = s/n$. Entonces es posible notar que el estimador bayesiano de θ en (**??**) de alguna forma combina el estimador clásico y el estimador previa. Más aún, se puede ver que $\hat{\theta}_B$ se puede escribir como un promedio ponderado entre la estimación clásica y la estimación previa. Puesto que

$$\begin{aligned} \hat{\theta}_B = \frac{s+\alpha}{n+\alpha+\beta} &= \frac{s}{n+\alpha+\beta} + \frac{\alpha}{n+\alpha+\beta} \\ &= \frac{n}{n+\alpha+\beta}\frac{s}{n} + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta} \\ &= \frac{n}{n+\alpha+\beta}\hat{\theta}_C + \frac{\alpha+\beta}{n+\alpha+\beta}\hat{\theta}_P \end{aligned}$$

De esta forma, queda en evidencia que la estimación bayesiana de θ siempre será un valor intermedio entre la estimación clásica y la estimación previa. La gráfica **??** da una ilustración acerca de la anterior afirmación, en donde se puede observar que para una distribución previa concentrada en 2/7 y una función de verosimilitud[2] con máximo en 8/10, se tiene una distribución posterior centrada en 10/17; es decir, la estimación bayesiana se encuentra situada entre la estimación previa y la estimación clásica.

[2]La función de verosimilitud es una función del parámetro y sólo se puede graficar una vez se hayan observado las realizaciones de la variable aleatoria.

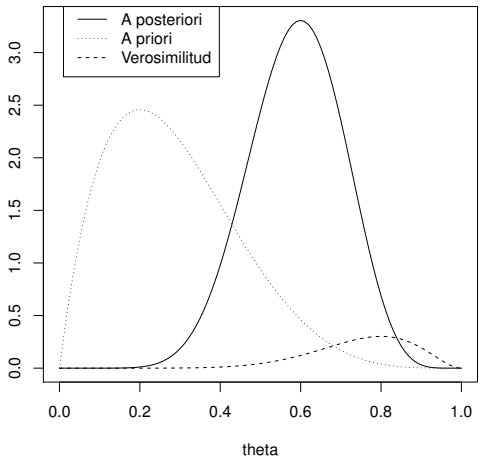

Figura 2.4: *Función de verosimilitud, función de densidad previa y posterior para* $\alpha = 2$, $\beta = 5$, $s = 8$ *y* $n = 10$.

Por otro lado, entre más grande sea el tamaño muestral n, más cercano estará $\hat{\theta}_B$ de $\hat{\theta}_C$ o equivalentemente la función de densidad posterior de θ estará más concentrada en s/n; mientras que entre mayor número de datos tenga la muestra de la distribución previa ($\alpha + \beta$=número de datos), más cercano estará $\hat{\theta}_B$ de $\hat{\theta}_P$ y la densidad posterior de θ estará más concentrada en $\alpha/(\alpha + \beta)$.

Para ilustrar lo anterior, suponga que la distribución previa de θ está dada con $\alpha = \beta = 5$, es decir la estimacion previa es 0.5, y suponga además que la estimación clásica es 0.33, pero el tamaño muestral n incrementa manteniendo constante la estimacion clásica. En la figura **??** se muestra la estimación posterior de θ, es evidente que a medida que el tamaño muestral n aumenta, la estimación posterior se acerca más a la estimación clásica.

Anteriormente, se comentó que se acostumbra a escoger los parámetros α y β que correspondan al número de éxitos y fracasos en la información previa, sin embargo, la información previa puede no presentarse de esta forma. Por ejemplo, en algunas situaciones, la información previa puede proveer el valor de θ, es decir, el valor de $\hat{\theta}_P$, y el valor de la desviación estándar de la estimación (comúnmente conocido como el error estándar). Por ejemplo, suponga que $\hat{\theta}_P = 0.5$ con un error estándar de 0.1, entonces podemos encontrar los valores de α y β de las expresiones $\frac{\alpha}{\alpha+\beta} = 0.5$ y $\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = 0.1$, de donde se tiene que $\alpha = 12$ y $\beta = 12$, y la distribución a priori correspondiente $Beta(12, 12)$ tiene una esperanza de 0.05 y una desviación estándar de 0.1. Se puede ver que entre mayor sea la desviación estándar, menores resultan los valores de α y β, que conducen a una distribución previa menos informativa.

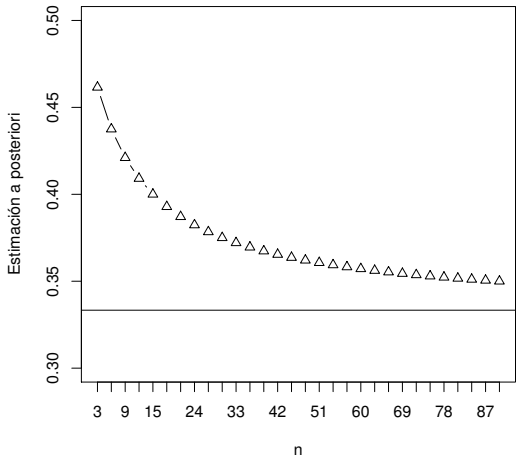

Figura 2.5: *Estimación posterior de θ para diferentes valores de n y s con $\alpha = \beta = 5$.*

Ahora, se vio anteriomente que la distribución previa no informativa de Jeffreys corresponde a la distribución $Beta(1/2, 1/2)$, la cual conduce a la distribución posterior $Beta(s + 1/2, n - s + 1/2)$, que a su vez nos lleva al estimador

$$\hat{\theta}_B = \frac{s + 1/2}{n + 1} \tag{2.2.3}$$

La anterior expresión es comparable con el estimador clásico $\hat{\theta_C} = \frac{s}{n}$, en el sentido de que los dos son aplicables cuando no se dispone de ninguna información previa. Podemos observar que aparte del alto grado de similitud que tienen los dos estimadores, es preferible usar el estimador (**??**) en situaciones donde el valor teórico de θ es muy pequeño, y como consecuencia en la muestra $s = 0$, por ejemplo, cuando θ representa el porcentaje de personas que esten infectados con algún virus poco común. En estos casos, el estimador clásico $\hat{\theta}_C = 0$ sugiriendo que ningún porcentaje de la población está infectado, conclusión que puede ser errónea; por otro lado, el estimador bayesiano $\hat{\theta}_B = \frac{0.5}{n+1}$, el cual tiende a ser un porcentaje muy pequeño a medida que aumente el tamaño muestral n, pero nunca llega a dar el valor 0 como la estimación de θ.

En el siguiente resultado, se encuentra la distribución predictiva previa para una variable binomial S.

Resultado 2.2.2. *La distribución predictiva previa para la observación particular de la suma de variables aleatorias Bernoulli, s, está dada por una distribución*

Beta-Binomial dada por

$$p(S) = \binom{n}{s} \frac{Beta(s+\alpha, \beta - s + n)}{Beta(\alpha, \beta)} I_{\{0,1,\ldots,n\}}(s). \tag{2.2.4}$$

Prueba. De la definición de función de distribución predictiva previa se tiene que

$$\begin{aligned} p(S) &= \int p(S \mid \theta) p(\theta \mid \alpha, \beta)\ d\theta \\ &= \int_0^1 \binom{n}{s} \theta^s (1-\theta)^{n-s} I_{\{0,1,\ldots,n\}}(s) \frac{1}{Beta(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}\ d\theta \\ &= \binom{n}{s} \frac{Beta(s+\alpha, \beta - s + n)}{Beta(\alpha, \beta)} I_{\{0,1,\ldots,n\}}(s) \\ &\qquad \times \int_0^1 \frac{\theta^{s+\alpha-1}(1-\theta)^{\beta-s+n-1}}{Beta(s+\alpha, \beta - s + n)}\ d\theta \\ &= \binom{n}{s} \frac{Beta(s+\alpha, \beta - s + n)}{Beta(\alpha, \beta)} I_{\{0,1,\ldots,n\}}(s) \end{aligned}$$

■

Una vez observados los valores muestrales, podemos encontrar la distribución predictiva posterior para una nueva variable binomial $\tilde{S}$ en una muestra de tamaño $\tilde{n}$. Esta distribución se encuentra en el siguiente resultado.

Resultado 2.2.3. *Después de la recolección de los datos $y_1, \cdots, y_n$, la distribución predictiva posterior para una nueva variable $\tilde{S}$ en una muestra del tamaño $\tilde{n}$ está dada por*

$$p(\tilde{s} \mid S) = \binom{\tilde{n}}{\tilde{s}} \frac{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}{Beta(s+\alpha, \beta - s + n)} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}), \tag{2.2.5}$$

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(\tilde{s} \mid S) &= \int p(\tilde{s} \mid \theta) p(\theta \mid S)\ d\theta \\ &= \int_0^1 \binom{\tilde{n}}{\tilde{s}} \theta^{\tilde{s}} (1-\theta)^{\tilde{n}-\tilde{s}} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}) \frac{\theta^{s+\alpha-1}(1-\theta)^{\beta-s+n-1}}{Beta(s+\alpha, \beta - s + n)}\ d\theta \\ &= \binom{\tilde{n}}{\tilde{s}} \frac{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}{Beta(s+\alpha, \beta - s + n)} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}) \\ &\qquad \times \int_0^1 \frac{\theta^{\tilde{s}+s+\alpha-1}(1-\theta)^{\beta-\tilde{s}-s+n+\tilde{n}-1}}{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}\ d\theta \\ &= \binom{\tilde{n}}{\tilde{s}} \frac{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}{Beta(s+\alpha, \beta - s + n)} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}) \end{aligned}$$

■

En la anterior distribución predictiva, se necesita calcular funciones Beta. Cuando los tamaños muestrales n, $\tilde{n}$ y/o los parámetros de la distribución previa α y β son muy grandes, puede presentar problemas numéricos al momento de calcular directamente estas funciones Beta. Por ejemplo, supongamos que $n = 1000$, $s = 650$, $\alpha = 200$, $\beta = 300$ y $\tilde{n} = 800$, de esta forma, los posibles valores para $\tilde{s}$ son $0, 1, \cdots, 800$, y se tiene que la probabilidad de que $\tilde{s}$ tome el valor 500 está dada por

$$Pr(\tilde{s} = 500|S) = \binom{800}{500} \frac{Beta(1350, 950)}{Beta(850, 650)} \tag{2.2.6}$$

y desafortunadamente, en R, se presenta error al intentar ejecutar `beta(1350,950)` o `beta(850,650)`. Planteamos la siguiente solución numérica cuando se quiere calcular la función predictiva (**??**) en muestras grandes. El problema central es el cómputo de $\frac{Beta(a,b)}{Beta(c,d)}$ con $a \geq c$ y $b \geq d$, valores enteros. Podemos ver que

$$\begin{aligned}
&\frac{Beta(a,b)}{Beta(c,d)} \\
&= \frac{(a-1)!(b-1)!(c+d-1)!}{(c-1)!(d-1)!(a+b-1)!} \\
&= \frac{(a-1)(a-2)\cdots(a-(a-c))(b-1)(b-2)\cdots(b-(b-d))}{(a+b-1)(a+b-2)\cdots(a+b-(a+b-c-d))} \\
&= \frac{a^{a-c}(1-\frac{1}{a})(1-\frac{2}{a})\cdots(1-\frac{a-c}{a})b^{b-d}(1-\frac{1}{b})(1-\frac{2}{b})\cdots(1-\frac{b-d}{b})}{(a+b)^{a+b-c-d}(1-\frac{1}{a+b})(1-\frac{2}{a+b})\cdots(1-\frac{a+b-c-d}{a+b})} \\
&= \underbrace{\left(\frac{a}{a+b}\right)^{a-c}}_{t1} \underbrace{\left(\frac{b}{a+b}\right)^{b-d}}_{t2} \underbrace{(1-\frac{1}{a})(1-\frac{2}{a})\cdots(1-\frac{a-c}{a})}_{t3} \\
&\quad \underbrace{(1-\frac{1}{b})(1-\frac{2}{b})\cdots(1-\frac{b-d}{b})}_{t4} \underbrace{(1-\frac{1}{a+b})(1-\frac{2}{a+b})\cdots(1-\frac{a+b-c-d}{a+b})}_{t5}
\end{aligned}$$

Calculando separadamente los términos $t1$, $t2$, $t3$, $t4$ y $t5$ podemos calcular $\frac{Beta(a,b)}{Beta(c,d)}$ para valores grandes de a, b, c y d. La siguiente función `prob` calcula la densidad (**??**) para un valor particular de $\tilde{s}$ usando la anterior técnica.

```
> prob<-function(s.mono,n.mono,s,n,alfa,beta){
+ a<-s.mono+s+alfa; b<-n.mono-s.mono+n-s+beta
+ c<-s+alfa; d<-n-s+beta
+ t1<-(a/(a+b))^(a-c); t2<-(b/(a+b))^(b-d)
+ t3<-prod(1-c(1:(a-c))/a); t4<-prod(1-c(1:(b-d))/b)
+ t5<-prod(1-c(1:(a+b-c-d))/(a+b))
+ if(a==c){resul<- t2*t4/t5}
+ if(b==d){resul<-t1*t3/t5}
+ if(a>c&b>d){resul<-choose(n.mono,s.mono)*t1*t2*t3*t4/t5}
+ resul
+ }
```

Si queremos examinar la distribución predictiva para todos valores de la variable $\tilde{S}$, podemos usar los siguientes códigos

```
> n<-1000; s<-650
> alfa<-200; beta<-300
> n.mono<-800
> res<-rep(NA,(1+n.mono))
> for(i in 1:length(res)){
+ res[i]<-prob(i-1,n.mono,s,n,alfa,beta)
+ }
```

y como resultado, `res` contiene las 801 probabilidades asociadas a todos los posibles valores de $\tilde{s}$.

Los resultados obtenidos con la anterior técnica es equivalente a lo obtenido usando la función `lbeta` que computa el logaritmo natural de la función beta. Así, para calcular la probabilidad en (**??**), simplemente usamos el siguiente código

```
> choose(800,500)*exp(lbeta(1350,950)-lbeta(850,650))
[1] 0.0005969157
```

Nótese que esta probabilidad es la misma contenido en `res`, puesto que

```
> res[501]
[1] 0.0005969157
```

Finalmente, se observa que la distribución predictiva (**??**) corresponde a una distribución Beta-binomial con parámetros $s + \alpha$ y $\beta - s + n$. Y el paquete `VGAM` en `R` **?** contiene funciones que calculan la función de densidad, función de distribución, percentiles, además de generar números aleatorios para la distribución Beta-binomial. Las probabilidades puntuales de $\tilde{s}$ se puede calcular con la función `dbetabinom`, teniendo en cuenta que los parámetros utilizados son $\mu = (s + \alpha)/(n + \alpha + \beta)$ y $\rho = 1/(1 + n + \alpha + \beta)$. Con el siguiente código, podemos calcular las probabilidades para todos los posibles valores de $\tilde{s}$.

```
> library(VGAM)
> mu<-(s+alfa)/(n+alfa+beta)
> rho<-1/(1+n+alfa+beta)
> res2<-rep(NA,(1+n.mono))
> for(i in 1:length(res2)){
+ res2[i]<-dbetabinom(i-1,size=n.mono,prob=mu,rho=rho)
+ }
```

Podemos ver que

```
> res2[501]
[1] 0.0005969157
```

que es idéntico a lo obtenido anteriormente. Adicionalmente, al escribir la distribución predictiva de (**??**) como la función de densidad de una distribución Beta-binomial, se puede encontrar la esperanza de esta distribución, la cual está dada por

$$E(\tilde{S}|S) = \tilde{n}\frac{s+\alpha}{n+\alpha+\beta}$$

Nótese que la esperanza en la anterior expresión corresponde simplemente al tamaño $\tilde{n}$ de la nueva muestra multiplicado por la estimación bayesiana del parámetro θ. Adicionalmente, la esperanza de $\tilde{S}$ también se puede obtener muliplicando todos los posibles valores de $\tilde{S}$ con su respectiva probabilidad, y sumand al final, como se muestra a continuacion.

```
> sum(res*c(0:n.mono))
[1] 453.3333
> n.mono*(s+alfa)/(n+alfa+beta)
[1] 453.3333
```

Retomando el ejemplo 2.1.1, suponga que la encuesta de opinión electoral se lleva a cabo en diferentes ciudades de un determinado país, en este caso, para cada ciudad se tiene una muestra de variables con distribución Bernoulli o equivalentemente una variable binomial; de esta forma, se dispone de una muestra de variables con distribución Binomial. La distribución posterior del parámetro θ para estos casos se encuentra en el siguiente resultado.

Resultado 2.2.4. *Cuando se tiene una sucesión de variables aleatorias $S_1, \ldots, S_i, \ldots, S_k$ independientes y con distribución Binomial(n_i, θ) para $i = 1, \ldots, k$, entonces la distribución posterior del parámetro de interés θ es*

$$\theta \mid S_1, \ldots, S_k \sim Beta\left(\sum_{i=1}^{k} s_i + \alpha, \beta + \sum_{i=1}^{k} n_i - \sum_{i=1}^{k} s_i\right)$$

Prueba.

$$\begin{aligned} p(\theta \mid S_1, \ldots, S_k) &\propto \prod_{i=1}^{k} p(S_i \mid \theta) p(\theta \mid \alpha, \beta) \\ &\propto \prod_{i=1}^{k} \theta^{\sum_{i=1} s_i} \theta^{\alpha-1} (1-\theta)^{\beta-1} (1-\theta)^{\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} s_i} I_{[0,1]}(\theta) \\ &= \theta^{\sum_{i=1}^{k} s_i + \alpha - 1} (1-\theta)^{\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} s_i + \beta} I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de densidad de la distribución $Beta\left(\sum_{i=1}^{k} s_i + \alpha, \beta + \sum_{i=1}^{k} n_i - \sum_{i=1}^{n} s_i\right)$. ■

Ejemplo 2.2.1. *El siguiente conjunto de datos fue estudiado inicialmente por* **?** *y se ha convertido en uno de los ejemplos prácticos más citados en la historia de la estadística moderna. Se trata de los porcentajes de bateo en una muestra de 18 jugadores profesionales en la temporada regular de béisbol en Estados Unidos en el año 1970.* **?** *establece que, en términos generales, este valor representa la razón entre la cantidad de* hits[3] *y el número de turnos al bate. La fórmula para calcular esta estadística es* s/n*, donde* s *es el número de* hits *y* n *es el total de turnos.*

Este conjunto de datos está disponible en el paquete `pscl` *(***?***) de* `R` *y se puede cargar mediante el siguiente código computacional.*

Código JAGS

```
library(pscl)
data(EfronMorris)
attach(EfronMorris)
```

	Nombre	s previo	n previo	s	n
1	Roberto Clemente	18	45	126	367
2	Frank Robinson	17	45	126	426
3	Frank Howard	16	45	143	521
4	Jay Johnstone	15	45	61	275
5	Ken Berry	14	45	114	418
6	Jim Spencer	14	45	125	466
7	Don Kessinger	13	45	154	586
8	Luis Alvarado	12	45	28	138
9	Ron Santo	11	45	137	510
10	Ron Swoboda	11	45	46	200
11	Del Unser	10	45	73	277
12	Billy Williams	10	45	69	270
13	George Scott	10	45	131	435
14	Rico Petrocelli	10	45	142	538
15	Ellie Rodriguez	10	45	42	186
16	Bert Campaneris	9	45	159	558
17	Thurman Munson	8	45	128	408
18	Max Alvis	7	45	14	70

En la primera columna se tiene el número del jugador, la segunda columna proporciona el nombre del jugador, la terce y cuarta columna representan el número de hits y el número de turnos al bate, respectivamente, luego de unas semanas de iniciada la temporada. La quinta y sexta columna representan el número de hits y el número de turnos al bate al final de la temporada regular.

[3] **?** afirma que se anota como *hit* la conexión efectuada por el bateador que coloca la pelota dentro del terreno de juego, permitiéndole alcanzar al menos una base, sin que se produzca un error de defensa del equipo contrario. Para lograr un hit, el bateador debe llegar a primera base antes de que ningún jugador defensivo lo toque con la bola en el trayecto del home a la inicial, o que el jugador de la defensa que tenga la bola pise la primera base antes que el bateador llegue a la misma.

Suponga que, partiendo de la muestra de los 18 jugadores, el objetivo es estimar el porcentaje de bateo, notado como θ, en toda la liga en el año de 1970. En primera instancia es plausible considerar que cada uno de los jugadores se comporta de manera independiente y que el porcentaje de bateo es común a todos, puesto que pertenecen a la misma liga profesional. Por lo tanto, es posible establecer que el número de hits s_i ($i = 1, \ldots, 18$) para cada jugador tiene la siguiente distribución

$$S_i \sim Binomial(n_i, \theta) \qquad i = 1, \ldots, 18.$$

Utilizando un enfoque bayesiano, es posible sacar provecho de la información recolectada al principio de la temporada, constituida por la tercera y cuarta columna del archivo de datos. En esta instancia, se tuvieron $18 + 17 + \cdots + 8 + 7 = 215$ hits para un total de $45 \times 18 = 810$ turnos al bate. Con esta información, se define la caracterización estructural de la distribución previa que, siguiendo las recomendaciones anteriores, está dada por una $Beta(\alpha = 215, \beta = 810 - 215) = Beta(\alpha = 215, \beta = 595)$. Del resultado 2.2.4, y teniendo en cuenta que al final de la temporada se obtuvieron $\sum S_i = 1818$ hits para un total de $\sum n_i = 6649$ turnos al bate, se tiene que la distribución posterior para este ejemplo es una $Beta(1818 + 215, 6649 - 1818 + 595) = Beta(2033, 5426)$. Por lo tanto, utilizando la distribución posterior, se estima que el porcentaje de bateo en la liga profesional en el año de 1970 es de $\frac{2033}{2033+5426} = \frac{2033}{7459} = 0.272$. Este valor corresponde a la media de la distribución posterior.

Nótese que los mismos resultados se encuentran cuando se analiza este conjunto de datos en JAGS, mediante el siguiente código computacional.

Código JAGS

```
model{
for(i in 1 : k)
{
s[i] ~ dbin(theta, n[i])
}
theta ~ dbeta(215,595)
}

DATA
list(k = 18, s = c(126, 126, 143,  61, 114, 125, 154,  28, 137,  46,  73,
69, 131, 142,  42, 159, 128,  14), n=c(367, 426, 521, 275, 418, 466, 586,
138, 510, 200, 277, 270, 435, 538, 186, 558, 408, 70))

INITS
list(theta=0.5)
```

La siguiente salida de JAGS permite conocer la estimación bayesiana posterior y los límites del intervalo de credibilidad al 95 %.

```
node   mean       sd    MC Error   2.5% median  97.5% start sample
```

```
theta 0.2726 0.005201 4.763E-5 0.2623 0.2726 0.2829     1  10000
```

La figura **??** *muestra el comportamiento de las distribuciones previa y posterior para este ejemplo. Nótese que, con un análisis frecuentista, se hubiese llegado a una estimación cercana de* $\frac{1818}{6649} = 0.273$*. Por otro lado, el mismo intervalo de credibilidad del 95 % correspondiente a* $(0.262, 0.282)$*, se puede hallar mediante el siguiente código computacional de* `R`*.*

Código R

```
> qbeta(c(0.025, 0.975),2033,5426)
[1] 0.2625105 0.2827183
```

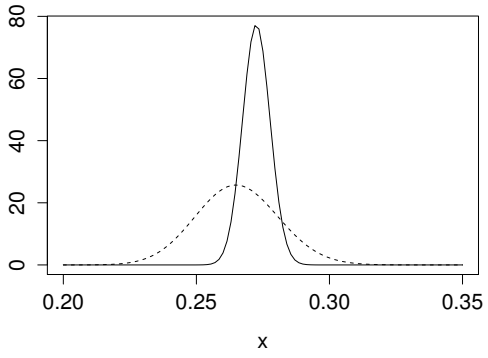

Figura 2.6: *Función de densidad previa (línea punteada) y función de densidad posterior (línea continua) para el ejemplo de bateo.*

Es posible analizar este conjunto de datos desde otra perspectiva al suponer que los jugadores no constituyen una muestra aleatoria y cada uno de ellos tiene un promedio de bateo diferente. Sin embargo, este análisis se deja como ejercicio en un capítulo posterior.

Ejemplo 2.2.2. *Continuando con el conjunto de datos de Efron y Morris, suponga que el entrenador de un equipo de las ligas inferiores está interesado en adquirir los servicios de Max Alvis. Este jugador no tuvo un buen promedio de bateo en la temporada y no tuvo muchos turnos al bate. El entrenador quiere conocer cuál será el número más probable de hits que anotará en la siguiente temporada. Teniendo en cuenta que es un jugador que viene de la liga profesional, lo más conveniente es que tenga muchos turnos al bate, digamos 400.*

Para resolver este cuestionamiento, es conveniente recurrir a la función predictiva posterior, dada en el resultado **??** *de la página* **??***. Para este análisis, se define la caracterización estructural de la distribución previa del jugador que está dada por*

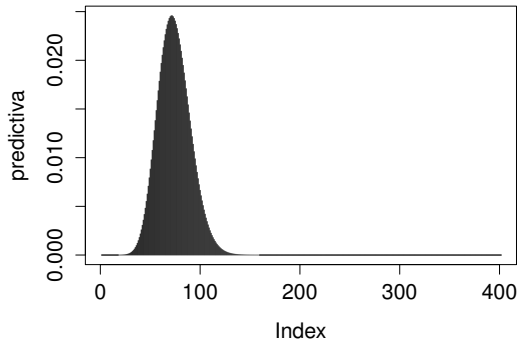

Figura 2.7: *Función de densidad predictiva posterior para el jugador Max Alvis.*

una $Beta(\alpha = 7, \beta = 38)$. *La siguiente función en* `R` *permite obtener la distribución predictiva para este jugador, que se muestra en la figura* **??**.

Código R

```
n <- 70
s<- 14
alp <-7
bet <- 38
n.ast <- 400

predictiva <- rep(NA,n.ast+1)
for(k in 0:n.ast)
{
predictiva[(k+1)] <-
choose(n.ast,k)*beta(k+s+alp,bet-k-s+n.ast+n)/beta(s+alp,bet-s+n)
}

sum(predictiva)
[1] 1
plot(predictiva, type="h")
which(predictiva==max(predictiva))
[1] 71
```

La última línea del código computacional permite concluir que lo más probable es que el jugador realice 71 hits en 400 turnos al bate, cifra que no convence al entrenador para adquirir los servicios del jugador.

2.3 Binomial negativa

La distribución binomial negativa describe el número de ensayos necesarios para alcanzar un número determinado y fijo de éxitos k en una secuencia independiente de experimentos tipo Bernoulli. Ésta distribución es particularmente útil cuando el porcentaje θ que se quiere estimar es muy pequeño, como la proporción de una población que padece de alguna enfermedad. La razón por la que no se utiliza la distribución binomial es que al fijar el número de ensayos n, como el porcentaje θ es muy pequeño, es muy probable que en en la muestra de tamaño n no se encuetre ningún paciente con la enfermedad; mientras que al utilizar la distribución binomial negativa, de antemano se garantiza que se obtendrá k pacientes con la enfermedad en la muestra.

Suponga que Y es una variable aleatoria cuya distribución es Binomial negativa que representa el número de ensayos necesarios y para alcanzar un número determinado y fijo de éxitos k en un experimento. La forma funcional de esta distribución es la siguiente

$$p(Y \mid \theta) = \binom{y-1}{k-1}\theta^k(1-\theta)^{y-k}I_{\{k,k+1,\ldots,\}}(y), \tag{2.3.1}$$

Así como en la distribución Bernoulli y Binomial, el parámetro θ está restringido al espacio $\Theta = [0, 1]$. Luego, es admisible proponer que θ siga una distribución Beta. Por tanto, la distribución previa del parámetro θ está dada por la expresión (**??**). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.3.1. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid Y \sim Beta(\alpha + k, \beta + y - k)$$

Prueba.

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta)p(\theta \mid \alpha, \beta) \\ &\propto \theta^{\alpha+k-1}(1-\theta)^{y+\beta-k-1}I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se llega a una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Beta(\alpha + k, \beta + y - k)$. ■

En algunas situaciones se puede encontrar una muestra de variables con distribución binomial negativa, por ejemplo, la entrevista de pacientes para encontrar cierta enfermedad puede llevarse a cabo en diferentes puntos de atención médica o en diferentes ciudades del país. Así en cada punto de atención, se tendrá el dato correspondiente a una variable con distribución binomial negativa. El procedimiento inferencial bayesiano para estas situaciones se describe a continuación:

Resultado 2.3.2. *Cuando se tiene una sucesión de variables aleatorias $Y_1, \ldots, Y_n$ independientes y con distribución $BinomialNegativa(k_i, \theta)$ $(i = 1, \ldots, n)$, entonces la distribución posterior del parámetro de interés es*

$$\theta \mid Y_1, \ldots, Y_n \sim Beta(\alpha + \sum_{i=1}^{n} k_i, \beta + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} k_i) \tag{2.3.2}$$

Prueba. Se deja como ejercicio para el lector. ■

Ejemplo 2.3.1. *Una franquicia de investigación farmacéutica ha desarrollado un nuevo tratamiento farmacológico sobre pacientes diabéticos que padezcan, a su vez, de enfermedades cardíacas, mejor conocidas como cardiopatías, entre las que se pueden encontrar la angina de pecho, infarto de miocardio, insuficiencia mitral, estenosis mitral, entre otras. Para evaluar el nuevo tratamiento, es necesario seleccionar una muestra, mediante el diseño de un experimento clínico, de pacientes que tienen estas características.*

Por otro lado, se sabe que la proporción de personas que padecen de diabetes y que además tienen algún tipo de condición cardiaca es muy baja y es necesario obtener una estimación precisa de la proporción de personas con estas condiciones. Con base en lo anteriormente expuesto, se puede pensar en seleccionar una grande de personas y utilizar un acercamiento binomial para estimar esta proporción. Sin embargo, dado que la prevalencia de esta condición es bastante baja, es posible que el número de personas en la muestra que presenten estas enfermedades sea nulo; por consiguiente, la estimación binomial no será, de ninguna forma, precisa.

Por lo tanto, el diseño clínico está supeditado al uso de la distribución Binomial Negativa, en donde se entrevistarán pacientes, de una base de datos de un hospital de la ciudad asociado con la franquicia, hasta conseguir una muestra de cinco pacientes que padezcan de estas condiciones. Después de varios meses de entrevistas, se encontró el quinto paciente en la entrevista número 1106.

Mediante el análisis bayesiano, suponiendo una distribución previa $Beta(0.5, 0.5)$, se llega a que la distribución posterior del parámetrtos θ es $Beta(0.5 + 5, 0.5 + 1106 - 5) = Beta(5.5, 1101.5)$. Por lo tanto, la estimación puntual del parámetro de interés, que corresponde a la media de la distribución posterior, es 0.0049, que equivale una proporción de 0.49 % de personas con estas enfermedades. El siguiente código computacional muestra cómo se puede llegar a las mismas conclusiones con JAGS

Código JAGS

```
model
{
y~dnegbin(theta,5)
theta~dbeta(0.5, 0.5)
}

DATA
list(y =1106)
```

```
INITS
list(theta=0.5)
```

Después de cinco mil iteraciones, la salida del anterior código muestra la estimación puntual dada por 0.004956 y un intervalo de credibilidad al 95 %, dado por $(0.001766, 0.009931)$.

Ejemplo 2.3.2. *Continuando con la temática del ejemplo anterior, suponga que la franquicia llevó a cabo la misma investigación en las 31 ciudades con mayor densidad poblacional de país. Como en la mayoría de los casos, debido al condicionamiento presupuestal, el experimento difirió en el número de éxitos en cada caso. En total, se tuvieron 29620 entrevistas para un total de éxitos de 152, tal como se muestra a continuación.*

Ciudad	y	k
BOGOTA	1001	4
MEDELLIN	978	6
CALI	999	5
BARRANQUILLA	860	4
CARTAGENA	1155	4
CUCUTA	585	6
BUCARAMANGA	1030	3
IBAGUE	960	5
SOLEDAD	1002	6
SANTA MARTA	763	7
SOACHA	1036	5
PASTO	779	5
MONTERIA	1158	4
VILLAVICENCIO	1017	5
BELLO	888	6
MANIZALES	977	4
VALLEDUPAR	1256	6
BUENAVENTURA	1349	6
NEIVA	1047	5
PALMIRA	1088	5
ARMENIA	649	3
POPAYAN	765	4
FLORIDABLANCA	699	5
SINCELEJO	1042	4
ITAGUI	1212	5
BARRANCABERMEJA	660	5
TULUA	671	5
ENVIGADO	835	6
DOSQUEBRADAS	997	5
RIOHACHA	1146	4

```
SINCELEJO            1016          5
```

Mediante el análisis bayesiano, suponiendo una distribución previa[4] *no informativa* $Beta(0.5, 0.5)$*, se llega a que la distribución posterior del parámetrtos* θ *es* $Beta(0.5 + 152, 0.5 + 29620 - 152) = Beta(152.5, 29468.5)$*. Por lo tanto, la estimación puntual del parámetro de interés, que corresponde a la media de la distribución posterior, es 0.0051, que equivale una proporción de 0.51 % de personas con estas enfermedades. El siguiente código computacional muestra cómo se puede llegar a las mismas conclusiones con JAGS*

Código JAGS

```
model
{
for(i in 1:31){
y[i]~dnegbin(theta,k[i])
}
theta~dbeta(0.5, 0.5)
}

DATA
list(y = c(1001, 978, 999, 860, 1155, 585, 1030, 960, 1002, 763, 1036,
779, 1158, 1017, 888, 977, 1256, 1349, 1047, 1088, 649, 765, 699, 1042,
1212, 660, 671, 835, 997, 1146, 1016), k = c(4, 6, 5, 4, 4, 6, 3, 5,
6, 7, 5, 5, 4, 5, 6, 4, 6, 6, 5, 5, 3, 4, 5, 4, 5, 5, 5, 6, 5, 4, 5))

INITS
list(theta=0.5)
```

Después de cinco mil iteraciones, la salida del anterior código muestra la estimación puntual dada por 0.005121 y un intervalo de credibilidad al 95 %, dado por $(0.004335, 0.005956)$*, mucho más estrecho que el intervalo de credibilidad del anterior ejercicio.*

Una vez observados los datos actuales y encontrada la distribución posterior, se puede encontrar la distribución predictiva posterior de una nueva variable con distribución binomial negativa. Es decir, se puede definir el mecanismo probabilístico para el número de ensayos necesarios para encontrar $\tilde{k}$ éxitos.

Resultado 2.3.3. *Después de la recolección de datos, la distribución predictiva posterior para una nueva variable* $\tilde{Y}$ *está dada por*

$$p(\tilde{Y}|Y_1, \cdots, Y_n) = \binom{\tilde{y}-1}{\tilde{k}-1} \frac{Beta(\alpha + \tilde{k} + \sum k_i, \beta + \tilde{y} - \tilde{k} + \sum y_i - \sum k_i)}{Beta(\alpha + \sum k_i, \beta + \sum y_i - \sum k_i)} I_{\{\tilde{k}, \tilde{k}+1, \cdots\}}(\tilde{y})$$

[4] Nótese que es posible también asignar una previa informativa $Beta(5.5, 1101.5)$, que da cuenta de la información del estudio del ejemplo anterior.

Prueba.

$$
\begin{aligned}
& p(\tilde{Y}|Y_1,\cdots,Y_n) \\
&= \int p(\tilde{Y}|\theta)p(\theta|Y_1,\cdots,Y_n)d\theta \\
&= \int_0^1 \binom{\tilde{y}-1}{\tilde{k}-1}\theta^{\alpha+\tilde{k}}(1-\theta)^{\beta+\tilde{y}-\tilde{k}} I_{\{\tilde{k},\tilde{k}+1,\cdots\}}(\tilde{y}) \frac{\theta^{\sum k_i-1}(1-\theta)^{\sum y_i-\sum k_i-1}}{Beta(\alpha+\sum k_i,\beta+\sum y_i-\sum k_i)}d\theta \\
&= \binom{\tilde{y}-1}{\tilde{k}-1}\frac{I_{\{\tilde{k},\tilde{k}+1,\cdots\}}(\tilde{y})}{Beta(\alpha+\sum k_i,\beta+\sum y_i-\sum k_i)}\int_0^1 \theta^{\alpha+\tilde{k}+\sum k_i-1}(1-\theta)^{\beta+\tilde{y}-\tilde{k}+\sum y_i-\sum k_i-1}d\theta \\
&= \binom{\tilde{y}-1}{\tilde{k}-1}\frac{Beta(\alpha+\tilde{k}+\sum k_i,\beta+\tilde{y}-\tilde{k}+\sum y_i-\sum k_i)}{Beta(\alpha+\sum k_i,\beta+\sum y_i-\sum k_i)} I_{\{\tilde{k},\tilde{k}+1,\cdots\}}(\tilde{y})
\end{aligned}
$$

■

Ejemplo 2.3.3. *Siguiendo con los datos del Ejemplo 2.3.2, suponga que se quiere recolectar información de 3 pacientes con cardiopatía en cierta ciudad, y se quiere conocer acerca del número de entrevistas necesarias. Utilizando la distribución previa* $Beta(0.5, 0.5)$ *y los datos de las 31 ciudades del ejemplo, se tiene que la distribución predictiva para el número de entrevistas necesarias para encontrar 3 pacientes está dada por*

$$
\begin{aligned}
& p(\tilde{Y}|Y_1,\cdots,Y_n) \\
&= \binom{\tilde{y}-1}{4}\frac{Beta(0.5+5+152,0.5+\tilde{y}-5+29620-152)}{Beta(0.5+152,0.5+29620-152)} I_{\{5,6,\cdots\}}(\tilde{y}) \\
&= \binom{\tilde{y}-1}{4}\frac{Beta(157.5,\tilde{y}+29463.5)}{Beta(152.5,29468.5)} I_{\{5,6,\cdots\}}(\tilde{y})
\end{aligned}
$$

Con los siguientes códigos se puede calcular la anterior función predictiva.

Código R

```
> predict<-function(y,alfa,beta,s,n,k){
+ choose(y-1,k-1)*exp(lbeta(alfa+k+s,beta+y-k+n-s)-lbeta(alfa+s,beta+n-s))
+ }
>
> alfa<-beta<-0.5
> s<-152;n<-29620;k<-5
> fun<-rep(NA)
>
> for(y in 5:5000){
+ fun[y-4]<-predict(y,alfa,beta,s,n,k)
+ }
> sum(fun)
[1] 0.9999994
> plot(fun,type="h",ylab="Predictiva")
```

Figura 2.8: *Distribución predictiva posterior para el número de entrevistas necesarias para encontrar 5 pacientes usando los datos del ejemplo 2.3.2*

Se puede ver que el número de entrevistas que tiene mayor probabilidad asociadas es el valor 772, usando el comando

```
> which(fun==max(fun))+4
[1] 772
```

También, se puede calcular la probabilidad de que en menos de 500 entrevistas se encuentren los 5 pacientes es de solo el 12 % usando el comando

```
> sum(fun[1:(500-4)])
[1] 0.1200985
```

2.4 Poisson

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es una muestra aleatoria de variables con distribución Poisson con parámetro θ, la función de distribución conjunta o la función de verosimilitud está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!} I_{\{0,1,\ldots\}}(y_i) \\ &= \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n) \end{aligned}$$

donde $\{0, 1 \ldots\}^n$ denota el producto cartesiano n veces sobre el conjunto $\{0, 1 \ldots\}$. Por otro lado, como el parámetro θ está restringido al espacio $\Theta = (0, \infty)$, entonces es posible formular varias opciones para la distribución previa del parámetro. Algunas opciones se encuentran considerando la distribución exponencial o la distribución chi-cuadrado o la distribución Gamma. Nótese que las dos primeras distribuciones son casos particulares de la última. Por lo tanto, la distribución previa del parámetro θ está dada por

$$p(\theta \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} I_{(0,\infty)}(\theta). \tag{2.4.1}$$

Bajo este marco de referencia se tienen el siguiente resultado con respecto a la distribucion posterior del parámetro de interés θ.

Resultado 2.4.1. *La distribución posterior del parámetro θ está dada por*

$$\theta \mid \mathbf{Y} \sim Gamma\left(\sum_{i=1}^{n} y_i + \alpha, n + \beta\right)$$

Prueba.

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^n y_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \theta^{\sum_{i=1}^n y_i} e^{-\beta\theta} e^{-n\theta} I_{(0,\infty)}(\theta) \\ &\propto \theta^{\sum_{i=1}^n y_i + \alpha - 1} e^{-(\beta+n)\theta} I_{(0,\infty)}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma(\sum_{i=1}^n y_i + \alpha, n + \beta)$. ■

Utilizando el resultado anterior, se tiene que la estimación Bayesiana del parámetro θ está dada por

$$\hat{\theta} = \frac{\sum_{i=1}^n y_i + \alpha}{n + \beta}.$$

La anterior expresión sugiere tomar los parámetros de la distribución previa α y β de la siguiente manera: β representa el número de observaciones en la información previa, mientras que α representa la suma de los datos de la información previa. De esta forma, α/β representa la estimación previa del parámetro θ. Y la estimación Bayesiana de θ se puede escribir como

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^n y_i + \alpha}{\beta + n} \\ &= \frac{n}{n+\beta} * \frac{\sum y_i}{n} + \frac{\beta}{n+\beta} * \frac{\alpha}{\beta} \\ &= \frac{n}{n+\beta} * \hat{\theta_C} + \frac{\beta}{n+\beta} * \hat{\theta_P} \end{aligned}$$

Es decir, la estimación Bayesiana de θ es un promedio ponderado entre la estimación clásica y la estimación previa del parámetro θ, donde los pesos dependen directamente del tamaño muestral de la información actual y de la información previa.

A continuación estudiamos las distribuciones predictivas previa y posterior de una nueva observación

Resultado 2.4.2. *La distribución predictiva previa para una observación* $\mathbf{y} = \{y_1, \ldots, y_n\}$ *de la muestra aleatoria está dada por*

$$p(\mathbf{Y}) = \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}} \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^n y_i!} \tag{2.4.2}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned}
p(\mathbf{Y}) &= \int p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta)\, d\theta \\
&= \int_0^\infty \frac{e^{-n\theta} \theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n) \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}\, d\theta \\
&= \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}} \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^n y_i!} \\
&\qquad \times \int_0^\infty \frac{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}}{\Gamma(\sum_{i=1}^n y_i + \alpha)} \theta^{\sum_{i=1}^n y_i + \alpha - 1} e^{-(\beta+n)\theta}\, d\theta \\
&= \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}} \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^n y_i!}
\end{aligned}$$

■

En el caso en que la muestra aleatoria estuviera constituida por una sola variable aleatoria, entonces $n = 1$ y si, en particular, los hiper-parámetros de la distribución previa fuesen $\alpha = \beta = 1$, entonces no es difícil ver, utilizando la definición de la función matemática Gamma, que la función de distribución predictiva (**??**) estaría dada por

$$\begin{aligned}
p(Y) &= \frac{\Gamma(y+1)}{\Gamma(1)} \frac{1}{2^{y+1}} \frac{I_{\{0,1,\ldots\}}(y)}{y!} \\
&= \frac{1}{2^{y+1}} I_{\{0,1,\ldots\}}(y)
\end{aligned} \tag{2.4.3}$$

Para chequear la convergencia de la anterior distribución es necesario recurrir a los resultados del análisis matemático (**?**, p. 361). Dado que el espacio de muestreo de la variable aleatoria Y es $\{0, 1, \ldots\}$, entonces la suma infinita converge a uno

lo que conlleva a que, en este caso particular, $P(Y)$ sea una auténtica función de densidad de probabilidad.

$$\sum_{y=0}^{\infty} p(Y=y) = \sum_{y=0}^{\infty} \left(\frac{1}{2}\right)^{y+1} = \frac{1}{2}\sum_{y=0}^{\infty}\left(\frac{1}{2}\right)^{y} = \frac{1}{2}\frac{1}{1-1/2} = 1$$

y podemos afirmar que la expresion (**??**) sí representa una función de densidad de una variable discreta. Ahora, consideramos la distribución predictiva poseterior de una muestra aleatoria, esta distribución se presenta en el siguiente resultado.

Resultado 2.4.3. *Después de la recolección de los datos, la distribución predictiva posterior para una nueva posible observación* $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{n^*}\}$*, de tamaño* n^**, está dada por*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{\Gamma(\sum_{i=1}^{n^*}\tilde{y}_i + \sum_{i=1}^{n} y_i + \alpha)}{\Gamma(\sum_{i=1}^{n} y_i + \alpha)} \frac{(\beta+n)^{\sum_{i=1}^{n} y_i + \alpha}}{(n^*+\beta+n)^{\sum_{i=1}^{n^*}\tilde{y}_i + \sum_{i=1}^{n} y_i + \alpha}} \times \frac{I_{\{0,1,\ldots\}^{n^*}}(\tilde{y}_1, \ldots, \tilde{y}_{n^*})}{\prod_{i=1}^{n^*}\tilde{y}_i!} \tag{2.4.4}$$

Prueba. De la definición de función de distribución predictiva, y haciendo uso del mismo razonamiento en la demostración del Resultado 2.1.8, se tiene la prueba inmediata. ■

La anterior distribución corresponde a una distribucion multivariada que nos permite calcular probabilidades predictivas para cualesquieras valores de $\tilde{y}_1, \cdots, \tilde{y}_{n^*}$; sin embargo, en algunas situaciones, como por ejemplo, cuando θ representa el número promedio de algún suceso en una región geográfica, entonces al momento de la predicción, podemos estar interesados en predecir el número total o el número promedio de sucesos en la nueva muestra aleatoria de regiones geográficos. Es decir, podemos estar más interesados en la distribución de $\sum_{y=1}^{n^*}\tilde{y}_i$ o de $\sum_{y=1}^{n^*}\tilde{y}_i/n^*$ en vez de la distribución conjunta de $\tilde{y}_1, \cdots, \tilde{y}_{n^*}$. La distribución predictiva de $\sum_{y=1}^{n^*}\tilde{y}_i$ se presenta en el siguiente resultado, y con esta distribución se puede obtener fácilmente probabilidades predictivas para $\sum_{y=1}^{n^*}\tilde{y}_i/n^*$.

Resultado 2.4.4. *Después de la recolección de los datos, la distribución predictiva posterior para la suma de un vector de observaciones nuevas* $(\tilde{y}_1, \ldots, \tilde{y}_{n^*})$*,* $\tilde{s} = \sum_{y=1}^{n^*}\tilde{y}_i$*, está dada por:*

$$p(\tilde{\mathbf{s}} \mid \mathbf{Y}) = \frac{\Gamma(\tilde{s} + \sum_{i=1}^{n} y_i + \alpha)}{\Gamma(\sum_{i=1}^{n} y_i + \alpha)} \frac{(n+\beta)^{\sum_{i=1}^{n} y_i + \alpha}}{(n^*+n+\beta)^{\tilde{s}+\sum_{i=1}^{n} y_i + \alpha}} \frac{(n^*)^{\tilde{s}} I_{\{0,1,\ldots\}}(\tilde{s})}{\tilde{s}!} \tag{2.4.5}$$

Prueba. Usando el hecho de que $\theta|\mathbf{Y} \sim Gamma(\sum_{i=1}^{n} y_i + \alpha, n+\beta)$ y $\tilde{s}|\theta \sim$

$Poisson(n^*\theta)$ se procede a calcular $\tilde{s}/p(\mathbf{y})$, así:

$$\begin{aligned}
&p(\tilde{s}|\mathbf{y}) \\
&= \int_\Omega p(\tilde{s}|\theta)p(\theta|\mathbf{y})d\theta \\
&= \int_\Omega \frac{(n^*\theta)^{\tilde{s}}e^{-n^*\theta}}{\tilde{s}!} I_{\{0,1,\ldots\}}(\tilde{s})(\beta+n)^{\sum_{i=1}^n y_i+\alpha} \frac{\theta^{\tilde{s}+\sum_{i=1}^n y_i+\alpha-1}}{\Gamma(\sum_{i=1}^n y_i+\alpha)} e^{-(\beta+n)\theta} I_{(0,\infty)}(\theta)d\theta \\
&= \frac{(n^*)^{\tilde{s}}(\beta+n)^{\sum_{i=1}^n y_i+\alpha}}{\tilde{s}!\Gamma(\sum_{i=1}^n y_i+\alpha)} I_{\{0,1,\ldots\}}(\tilde{s}) \int_0^\infty \theta^{\sum_{i=1}^n y_i+\alpha-1} e^{-(n^*+\beta+n)\theta} d\theta
\end{aligned}$$

Agrupando las constantes para obtener la integral de una distribución gamma con $\alpha = \tilde{s} + \sum_{i=1}^n y_i + \alpha$ y $\beta = n^* + n + \beta$ se obtiene el resultado. ■

Debido a la complejidad de la expresión en (**??**), es prácticamente imposible comprobar analíticamente $\sum_{i=0}^\infty p(\tilde{s}=i) = 1$, y también muy difícile encontrar la expresión matemática de la esperanza de $\tilde{s}$, sin embargo, en situaciones prácticas, se puede usar aproximaciones numéricas tal como se verá en el ejemplo al final de esta sección.

En el ejemplo 1.5.4, se consideró la situación cuando no se tiene ninguna información previa, la distribución previa que se debe usar está dada por

$$p(\theta) \propto \theta^{-1/2},$$

que corresponde a una distribución previa impropia, puesto que $\int_0^\infty \theta^{-1/2} = \infty$. Sin embargo, este hecho no afecta que la inferencia posterior se pueda llevar a cabo, puesto que la distribución posterior está dada por

$$\theta|\mathbf{Y} \sim Gamma(\sum y_i + 1/2, n)$$

y la estimación Bayesiana del parámetro θ viene dada por

$$\hat{\theta} = \frac{\sum y_i + 1/2}{n}.$$

la cual es muy similar a la estimación clásica de θ dada por $\bar{Y}$.

Cuando se utiliza la distribución previa no informativa de Jeffreys, la distribución predictiva para nuevas observaciones $\tilde{y} = \tilde{y}_1, \cdots, \tilde{y}_{n^*}$ y $\tilde{s} = \sum_{i=1}^{n^*} \tilde{y}_i$ están dadas por

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{\Gamma(\sum_{i=1}^{n^*} \tilde{y}_i + \sum_{i=1}^n y_i + 0.5)}{\Gamma(\sum_{i=1}^n y_i + 0.5)} \frac{n^{\sum_{i=1}^n y_i+0.5}}{(n^*+n)^{\sum_{i=1}^n \tilde{y}_i+\sum_{i=1}^n y_i+0.5}} \frac{I_{\{0,1,\ldots\}^{n^*}}(\tilde{y}_1,\ldots,\tilde{y}_{n^*})}{\prod_{i=1}^{n^*}\tilde{y}_i!} \tag{2.4.6}$$

y

$$p(\tilde{\mathbf{s}} \mid \mathbf{Y}) = \frac{\Gamma(\tilde{s} + \sum_{i=1}^n y_i + 0.5)}{\Gamma(\sum_{i=1}^n y_i + 0.5)} \frac{n^{\sum_{i=1}^n y_i+0.5}}{(n^*+n)^{\tilde{s}+\sum_{i=1}^n y_i+0.5}} \frac{I_{\{0,1,\ldots\}}(\tilde{s})}{\tilde{s}!} \tag{2.4.7}$$

Ejemplo 2.4.1. *Por políticas gubernamentales, los alcaldes las ciudades están obligados a realizar un seguimiento exhaustivo al comportamiento de la accidentalidad en las vías urbanas y medirlo en términos del número de accidentes de tránsito. Lo anterior es necesario para evaluar la gestión de las autoridades administrativas y evaluar las políticas públicas que el gobierno de la ciudad ha implementado para disminuir esta cifra.*

Suponga que la alcaldía de una ciudad quiere implementar una estrategia educativa para disminuir el número de accidentes de tránsito, generados por manejar en estado de embriaguez. Para esto, se registraron durante diez días 30 días el número de accidentes de tránsito por ebriedad del conductor. Los datos para cada uno de los días son 22, 9, 9, 20, 10, 14, 11, 14, 11, 11, 19, 12, 8, 9, 16, 8, 13, 8, 14, 12, 14, 11, 14, 13, 11, 14, 13, 11, 7, 12.

Es posible modelar la variable aleatoria número de accidentes de tránsito en un día mediante una distribución de Poisson puesto que el promedio muestral y la varianza muestral de los datos son semejantes. Para este conjunto de datos, el promedio equivale a 12.33, mientras que la varianza es de 12.51. El histograma de los valores observados se puede ver en la figura **??**.

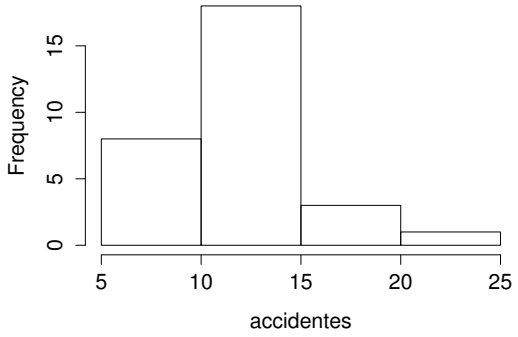

Figura 2.9: *Histograma para los datos de accidentes de tránsito.*

En primera instancia, es posible realizar un análisis no informativo, al formular una distribución previa de Jeffreys, utilizando el resultado del ejemplo **??** *de la página* **??**, *que indica que una distribución previa no informativa es proporcional a* $\theta^{-1/2}$, *para lo cual la distribución posterior* $Gamma(\sum_{i=1}^{n} yi + 1/2, n)$. *De esta manera, la distribución posterior del parámetro de interés es* $Gamma(370.5, 30)$. *Por lo tanto, un estimador de* θ *está dado por la media de la distribución posterior que es* $370.5/30 = 12.35$, *muy cercano al valor del estimador de máxima verosimilitud correspondiente al promedio muestral. La figura* **??** *(lado izquierdo) muestra el comportamiento de las distribuciones de Jeffreys y posterior para este ejemplo.*

Por otro lado, basándose en datos históricos, la alcaldía observó que, en el mismo periodo del año anterior, ocurrieron 37 accidentes en 9 días de observación.

Luego, una distribución previa informativa[5] *está dada por* $Gamma(\alpha = 38, \beta = 9)$. *Luego, apelando al resultado* **??**, *la distribución posterior corresponde a una* $Gamma(370 + 38, 30 + 9) = Gamma(408, 39)$. *Para este caso, un estimador de* θ *está dado por la media de la distribución posterior que es* $480/39 = 12.31$. *La figura* **??** *(lado derecho) muestra el comportamiento de las distribuciones previa (informativa) y posterior para este ejemplo.*

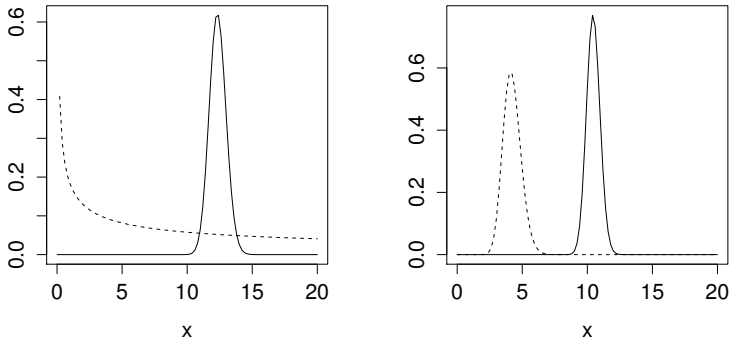

Figura 2.10: *Distribución previa (línea punteada) y distribución posterior (línea continua) para el ejemplo del tránsito.*

Ejemplo 2.4.2. *A continuación se grafica la distribución previa informativa posterior para una observación,* $\tilde{y}$*:*

2.5 Exponencial

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ corresponde a una muestra de variables aleatorias con distribución Exponencial. Luego, la función de distribución conjunta o verosimilitud está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{n} \theta e^{(-\theta y)} I_{(0,\infty)}(y_i) \\ &= \theta^n e^{(-\theta \sum_{i=1}^{n} y_i)} I_{(0,\infty)^n}(y_1, \ldots, y_n) \end{aligned} \tag{2.5.1}$$

Donde $\{0, 1 \ldots\}^n$ denota el producto cartesiano n veces sobre el intervalo $(0, \infty)$. Por otro lado, como el parámetro θ está restringido al espacio $\Theta = (0, \infty)$, entonces es posible formular varias opciones para la distribución previa del parámetro, al

[5] En la práctica, se recomienda que los valores de los hiperparámetros α y β correspondan a la suma del número de eventos más uno y número de observaciones, respectivamente.

Figura 2.11: *Distribución predictiva posterior para* $n* = 1$ *para el ejemplo del tránsito.*

igual que en la distribución Poisson. Así mismo, suponga que la distribución previa para el parámetro de interés es la distribución Gamma tal como aparece en la expresión (2.1.9). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.5.1. *La distribución posterior del parámetro* θ *sigue una distribución*

$$\theta \mid \mathbf{Y} \sim Gamma\left(\alpha + n, \beta + \sum_{i=1}^{n} y_i\right)$$

Prueba.

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \theta^n e^{(-\theta \sum_{i=1}^n y_i)} I_{(0,\infty)^n}(y_1, \ldots, y_n) \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)} I_{(0,\infty)}(\theta) \\ &\propto \theta^{\alpha+n-1} e^{-(\beta+\sum_{i=1}^n y_i)} I_{(0,\infty)}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma(\alpha + n, \beta + \sum_{i=1}^n y_i)$. ■

Resultado 2.5.2. *La distribución predictiva previa para una observación* $\mathbf{y} = \{y_1, \ldots, y_n\}$ *de la muestra aleatoria está dada por*

$$p(\mathbf{Y}) = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}} I_{(0,\infty)^n}(y_1, \ldots, y_n) \tag{2.5.2}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(\mathbf{Y}) &= \int p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta)\, d\theta \\ &= \int_0^\infty \theta^n e^{(-\theta \sum_{i=1}^n y_i)} I_{(0,\infty)^n}(y_1, \ldots, y_n) \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}\, d\theta \\ &= \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}} I_{(0,\infty)^n}(y_1, \ldots, y_n) \\ &\qquad \times \int_0^\infty \frac{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(n+\alpha)} \theta^{\alpha+n-1} e^{-(\beta+\sum_{i=1}^n y_i)\theta}\, d\theta \\ &= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}} I_{(0,\infty)^n}(y_1, \ldots, y_n) \end{aligned}$$

■

Por ejemplo, en el caso en que la muestra aleatoria estuviera constituida por una sola variable aleatoria, entonces no es difícil ver, utilizando la definición de la función matemática Gamma, que la función de distribución predictiva (2.2.2) estaría dada por

$$\begin{aligned} p(Y) &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta+y)^{\alpha+1}} I_{(0,\infty)}(y) \\ &= \frac{\alpha\beta^\alpha}{(\beta+y)^{\alpha+1}} I_{(0,\infty)}(y) \end{aligned}$$

Para chequear la convergencia de la anterior distribución es necesario recurrir a los resultados del cálculo integral. Dado que el espacio de muestreo de la variable

aleatoria Y es el intervalo $(0, \infty)$, entonces la integral a uno lo que conlleva a que, en este caso particular, $P(Y)$ sea una auténtica función de densidad de probabilidad.

$$\int_0^\infty p(Y)\ dy = \int_0^\infty \frac{\alpha\beta^\alpha}{(\beta+y)^{\alpha+1}}\ dy = \beta^\alpha \left[\frac{(\beta+y)^{-\alpha}}{-\alpha}\right]_0^\infty = 1$$

Volviendo al caso general en donde se tiene una muestra aleatoria, se tiene el siguiente resultado.

Resultado 2.5.3. *Después de la recolección de los datos, la distribución predictiva posterior para una conjunto de nuevas variables aleatorias* $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{n^*}\}$*, de tamaño* n^**, está dada por*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{\Gamma(n+\alpha+n^*)}{\Gamma(n+\alpha)} \frac{(\beta+\sum_{i=1}^n y_i)^{n+\alpha}}{(\sum_{i=1}^{n^*} \tilde{y}_i + \beta + \sum_{i=1}^n y_i)^{n^*+\alpha+n}} \times I_{(0,\infty)^{n^*}}(\tilde{y}_1, \ldots, \tilde{y}_n) \tag{2.5.3}$$

Prueba. De la definición de función de distribución predictiva, y haciendo uso del mismo razonamiento en la demostración del Resultado 2.2.2, se tiene la prueba inmediatamente. ■

El anterior resultado permite calcular la distribución predictiva conjunta de variables aleatorias por observar, en algunas situaciones, lo que se quiere pronosticar es el comportamiento probabilístico de promedio muestral de este conjunto de variables aleatorias, es decir, $\bar{Y}^* = \sum_{i=1}^{n^*} \tilde{Y}_i$. En el siguiente resultado se presenta la distribución predictiva de esta variable aleatoria.

Resultado 2.5.4. *Después de la recolección de los datos, la distribución predictiva posterior para el promedio muestral de un nuevo conjunto de variables aleatorias* $\bar{Y}^* = \sum_{i=1}^{n^*} \tilde{Y}_i$ *está dada por*

$$p(\bar{Y}^*) = \frac{n^*\Gamma(n^*+\alpha+n)}{\Gamma(n^*)\Gamma(\alpha+n)} \frac{(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{(n^*\bar{Y}^* + \beta + \sum y_i)^{n^*+\alpha+n}} (n^*\bar{Y}^*)^{n^*-1} I_{(0,\infty)}(\bar{Y}^*)$$

Prueba. En primer lugar se halla la distribución predictiva posterior de la variable $\tilde{S} = \sum_{i=1}^{n^*} \tilde{Y}_i$, teniendo en cuenta que $\tilde{S}|\theta \sim Gamma(n^*, \theta)$, de esta forma

$$\begin{aligned}
p(\tilde{S}|\mathbf{Y}) &= \int p(\tilde{S}|\theta)p(\theta|\mathbf{Y})\ d\theta \\
&= \int_0^\infty \frac{\theta^{n^*}}{\Gamma(n^*)} \tilde{S}^{n^*-1} e^{-\theta\tilde{S}} I_{(0,\infty)}(\tilde{S}) \frac{(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(\alpha+n)} \theta^{\alpha+n-1} e^{-(\beta+\sum y_i)\theta} d\theta \\
&= \frac{\tilde{S}^{n^*-1}(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(n^*)\Gamma(\alpha+n)} I_{(0,\infty)}(\tilde{S}) \int_0^\infty \theta^{n^*+\alpha+n-1} e^{-(\tilde{S}+\beta+\sum y_i)\theta}\ d\theta \\
&= \frac{\tilde{S}^{n^*-1}(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(n^*)\Gamma(\alpha+n)} \frac{\Gamma(n^*+\alpha+n)}{(\tilde{S}+\beta+\sum y_i)^{n^*+\alpha+n}} I_{(0,\infty)}(\tilde{S})
\end{aligned}$$

Al aplicar el teorema de transformación a la distribución predictiva, se puede hallar la distribución de $\bar{Y}^*$, dada por

$$p(\bar{Y}^*|\mathbf{Y}) = \frac{n^*\Gamma(n^*+\alpha+n)}{\Gamma(n^*)\Gamma(\alpha+n)} \frac{(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{(n^*\bar{Y}^*+\beta+\sum y_i)^{n^*+\alpha+n}} (n^*\bar{Y}^*)^{n^*-1} I_{(0,\infty)}(\bar{Y}^*)$$

■

Se puede ver que al utilizar la distribución previa no informativa de Jeffrey, la distribución predictiva posterior de $\bar{Y}^*$ está dada por

$$p(\bar{Y}^*|\mathbf{Y}) = \frac{n^*\Gamma(n^*+n)}{\Gamma(n^*)\Gamma(n)} \frac{(\sum_{i=1}^n y_i)^n}{(n^*\bar{Y}^*+\sum y_i)^{n^*+n}} (n^*\bar{Y}^*)^{n^*-1} I_{(0,\infty)}(\bar{Y}^*)$$

Ejemplo 2.5.1. ? *reportan un conjunto de datos que da cuenta de los tiempos de sobrevivencia de $n = 69$ miembros del programa de transplante de corazón de Stanford.Los tiempos se reportan en días después del transplante. Los datos pueden ser encontrados en el paquete* `survival` *(?) de* `R` *mediante la implementación del siguiente código computacional.*

Código R

```
require(survival)
data(heart)
attach(heart)
View(heart)

sobrevida <- stop[transplant==1]-start[transplant==1]
data.frame(heart[transplant==1,], sobrevida)
```

A continuación se muestran los primeros y últimos datos de este estudio. Se recuerda que el total de pacientes atendidos en este estudio fue de $n = 69$ y la suma de los tiempos de sobrevida es de $\sum_{i=1}^n y_i = 25998.5$.

id	start	stop	Sobrevida
3	1.0	16	15.0
4	36.0	39	3.0
7	51.0	675	624.0
...	...	...	...
97	21.0	131	110.0
98	96.0	109	13.0
100	38.0	39	1.0

Estos tiempos pueden ser modelados mediante una distribución exponencial. Además de inferir acerca del parámetro de esta distribución, también es posible

inferir acerca del tiempo promedio de sobrevivencia de un individuo sometido a este tipo de transplantes. Luego, dadas las implicaciones del estudio, se debe ser muy cuidadosos en la asignación de los parámetros de la distribución previa. Una forma de hacerlo es asignar valores muy pequeños a estos parámetros. Otra forma de hacerlo es utilizando la distribución previa de Jeffreys, que corresponde a una distribución impropia (ver ejercicio XXXXXXXXXX) y conduce a resultados muy cercanos a los del enfoque anterior.

Utilizando parámetros previos muy cercanos a cero, la distribución posterior del parámetro de interés es $Gamma(69, 25998.5)$. *Como es bien sabido, una estimación bayesiana para el parámetro* θ *está dada por la media de esta distribución posterior, la cual equivale a* $69/25998.5 = 0.0026$. *Ahora, como la esperanza de la distribución exponencial es* $1/\theta$, *entonces el tiempo promedio de sobrevivencia es de* $1/0.0026 = 376.78$ *días. Sin embargo, el promedio no es una medida de escala válidas en este tipo de análisis, en donde se presentan datos atípicos, puesto que no es una medida robusta y se prefiere la utilización de la mediana. El siguiente código computacional en JAGS puede ser usado para realizar inferencias sobre el parámetro* θ, *sobre el tiempo promedio y el tiempo mediano. De la misma forma, es posible obtener intervalos de credibilidad para estos parámetros.*

Código JAGS

```
model
{
for(i in 1:n)
{
y[i] ~ dexp(theta)
}
theta ~ dgamma(0.1,0.1)
mean <- 1/theta
}

DATA
list(n=69, y=c(15, 3, 624, 46, 127, 64, 1350, 280, 23, 10, 1024, 39, 730,
136, 1775, 1, 836, 60, 1536, 1549, 54, 47, 0.5, 51, 1367, 1264, 44, 994, 51,
1106, 897, 253, 147, 51, 875, 322, 838, 65, 815, 551, 66, 228, 65, 660, 25,
589, 592, 63, 12, 499, 305, 29, 456, 439, 48, 297, 389, 50, 339, 68, 26, 30,
237, 161, 14, 167, 110, 13, 1))

INITS
list( theta=0.5)
```

Después de diez mil iteraciones, los resultados de este código muestran una estimación para θ *de 0.0026 con un intervalo de credibilidad de (0.002065, 0.00332). Para la media* $1/\theta$, *se tiene una estimación puntual de 382.1 con un intervalo de credibilidad de (301.2, 484.3). La mediana se estimó en 378 días de sobrevivencia.*

Suponga ahora se va a realizar el transplante de corazón a 5 pacientes, y se quiere conocer el comportamiento probabilístico del tiempo promedio de sobrevida

en estos 5 pacientes. Aplicando la distribución predictiva obtenida en el resultado 2.5.4, usando la distribución previa no informativa de Jeffrey, se tiene que

$$p(\bar{Y}^*|\mathbf{Y}) = \frac{5\Gamma(5+69)}{\Gamma(5)\Gamma(69)} \frac{25998.5^{69}}{(5\bar{Y}^* + 25998.5)^{5+69}} (5\bar{Y}^*)^4$$

El cálculo de esta función predictiva se puede llevar a cabo con el siguiente código en R*, además de comprobar que la integral de la función es 1.*

Código R

```
> pred_exp<-function(x){
+ n.mono*((s/(s+x*n.mono))^n)*((x*n.mono/(s+x*n.mono))^n.mono)/(x*n.mono*beta(n,n.mono))
+ }
>
> alfa<-beta<-0
> s<-25998.5
> n<-69
> n.mono<-5
> integrate(pred_exp,0.0001,10000)
1 with absolute error < 3.2e-10
```

La distribución predictiva de esta función se puede visualizar en la figura Podemos

Figura 2.12: *Distribución predictiva posterior para los datos de tiempo de sobrevidad de transplante de corazón.*

ver que la probabilidad de que en promedio los cinco pacientes sobrevivan más de 800 días es de 2.6 %, usando el comando

```
> integrate(pred_exp,800,10000)
0.02644847 with absolute error < 2.6e-09
```

2.6 Normal con media desconocida y varianza conocida

Suponga que Y es una variable aleatoria con distribución $Normal(\theta, \sigma^2)$ con θ desconcido pero σ^2 conocido. Luego, la función de distribución de los datos está dada por

$$p(Y \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\} I_{\mathbb{R}}(y) \tag{2.6.1}$$

Como el parámetro θ está restringido a los reales, entonces es posible asignarle una distribución previa

$$\theta \sim Normal(\mu, \tau^2).$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.6.1. *La distribución posterior del parámetro de interés sigue una distribución*

$$\theta \sim Normal(\mu_1, \tau_1^2).$$

En donde

$$\mu_1 = \frac{\frac{1}{\sigma^2}Y + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \tag{2.6.2}$$

Prueba.

$$\begin{aligned}
p(\theta \mid Y) &\propto p(Y \mid \theta) p(\theta \mid \mu, \tau^2) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} - \frac{(\theta-\mu)^2}{\tau^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{\theta^2}{\sigma^2} - \frac{2\theta y}{\sigma^2} + \frac{\theta^2}{\tau^2} - \frac{2\theta\mu}{\tau^2}\right]\right\} \\
&= \exp\left\{-\frac{\theta^2}{2}\left[\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right] + \theta\left[\frac{y}{\sigma^2} + \frac{\mu}{\tau^2}\right]\right\} \\
&= \exp\left\{-\frac{\theta^2}{2\tau_1^2} + \frac{\theta\mu_1}{\tau_1^2}\right\} \\
&= \exp\left\{-\frac{1}{2\tau_1^2}(\theta^2 - 2\theta\mu_1)\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau_1^2}(\theta^2 - 2\theta\mu_1 + \mu_1^2)\right\} = \exp\left\{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_1, \tau_1^2)$. ■

Nótese que en el caso en donde se desconozca el comportamiento estructural de θ, entonces es posible hacer su distribución tan plana y vaga como sea posible. Para esto, basta con hacer tender al parámetro de precisión de la distribución previa hacia infinito. Es decir $\tau^2 \longrightarrow \infty$, y por lo tanto la distribución posterior tendería a una $Normal(y, \sigma^2/n)$.

Resultado 2.6.2. *La distribución predictiva previa para una observación y es*

$$y \sim Normal(\mu, \tau^2 + \sigma^2)$$

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(Y) &= \int p(Y \mid \theta) p(\theta \mid \mu, \tau^2)\, d\theta \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\} d\theta \end{aligned}$$

? desarrolló las siguientes igualdades

$$\begin{aligned} &\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(y-\theta)^2}{\sigma^2}\right] \\ &= \frac{1}{2}\left[\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\theta^2 - 2\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)\theta + \left(\frac{\mu^2}{\tau^2}+\frac{y^2}{\sigma^2}\right)\right] \\ &= \frac{1}{2\tau_1^2}\left[\theta^2 - 2\tau_1^2\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)\theta + \tau_1^4\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)^2\right] + \frac{1}{2}\left(\frac{\mu^2}{\tau^2}+\frac{y^2}{\sigma^2}\right) - \frac{\tau_1^2}{2}\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)^2 \\ &= \frac{1}{2\tau_1^2}\left[\theta - \tau_1^2\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)\right]^2 + \frac{1}{2}\left[\left(\frac{1}{\sigma^2}-\frac{\tau_1^2}{\sigma^4}\right)y^2 - 2\frac{\mu\tau_1^2}{\tau^2\sigma^2}y + \left(\frac{\mu^2}{\tau^2}-\frac{\mu^2\tau_1^2}{\tau^4}\right)\right] \\ &= \frac{1}{2\tau_1^2}\left[\theta-\mu_1\right]^2 + \frac{1}{2}\left[\frac{1}{\sigma^2+\tau^2}y^2 - 2\frac{\mu}{\sigma^2+\tau^2}y + \frac{\mu^2}{\sigma^2+\tau^2}\right] \\ &= \frac{1}{2\tau_1^2}\left[\theta-\mu_1\right]^2 + \frac{1}{2(\sigma^2+\tau^2)}(y-\mu)^2. \end{aligned}$$

Entonces

$$\begin{aligned} p(Y) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma\tau} \exp\left\{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right\} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}(y-\mu)^2\right\} d\theta \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\tau_1^2}}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}(y-\mu)^2\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left\{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right\} d\theta \\ &= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}(y-\mu)^2\right\} \end{aligned}$$

■

Resultado 2.6.3. *La distribución predictiva posterior para una nueva observación* $\tilde{y}$ *es*

$$\tilde{y} \mid Y \sim Normal(\mu_1, \tau_1^2 + \sigma^2)$$

Prueba. Análogamente a la demostración del anterior resultado. ■

Por otro lado, suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ conforma una muestra aleatoria de variables con distribución $Normal(\theta, \sigma^2)$. Una vez más, con θ desconcido pero σ^2 conocido. Luego, la función de distribución conjunta o verosimilitud de los datos está dada por

$$p(\mathbf{Y} \mid \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\} I_{\mathbb{R}^n}(\mathbf{y}) \tag{2.6.3}$$

Bajo el mismo marco de referencia en donde el parámetro θ está restringido a los reales y su distribución previa seguía una distribución $Normal(\mu, \tau^2)$, se tienen los siguientes resultados

Resultado 2.6.4. *La distribución posterior del parámetro de interés sigue una distribución* $Normal(\mu_n, \tau_n^2)$. *En donde*

$$\mu_n = \frac{\frac{n}{\sigma^2}\bar{Y} + \frac{1}{\tau^2}\mu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \tag{2.6.4}$$

Prueba. En primer lugar nótese que

$$\sum_{i=1}^{n}(y_i - \theta)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \theta)^2 \tag{2.6.5}$$

puesto que

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \theta)^2 &= \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \theta)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \theta)\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - \theta)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \theta)^2\end{aligned}$$

Por lo tanto, se tiene que

$$\begin{aligned}
p(\theta \mid Y) &\propto p(Y \mid \theta)p(\theta \mid \mu, \tau^2) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{n(\bar{y}-\theta)^2}{\sigma^2} - \frac{(\theta-\mu)^2}{\tau^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{n\theta^2}{\sigma^2} - \frac{2n\theta\bar{y}}{\sigma^2} + \frac{\theta^2}{\tau^2} - \frac{2\theta\mu}{\tau^2}\right]\right\} \\
&= \exp\left\{-\frac{\theta^2}{2}\left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right] + \theta\left[n\frac{\bar{y}}{\sigma^2}+\frac{\mu}{\tau^2}\right]\right\} \\
&= \exp\left\{-\frac{\theta^2}{2\tau_n^2} + \frac{\theta\mu_n}{\tau_n^2}\right\} \\
&= \exp\left\{-\frac{1}{2\tau_n^2}(\theta^2 - 2\theta\mu_n)\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau_n^2}(\theta^2 - 2\theta\mu_n + \mu_n^2)\right\} = \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_n, \tau_n^2)$. ■

En caso de que se desconoce el comportamiento estructural de θ, se puede utilizar la distribución previa no informativa de Jeffrey. Es bien conocido en la teoría clásica que la información de Fisher contenida en una variable aleatoria Y acerca de θ está dada por $I(\theta) = 1/\sigma^2$, de esta forma, se puede concluir que la distribución previa no informativa de Jeffrey corresponde a una función constante $p(\theta) \propto k$, la cual no es una distribución de probabilidad propiamente dicho; sin embargo, al utilizar esta distribución previa, se puede encontrar que

$$\begin{aligned}
p(\theta|Y) &\propto p(Y|\theta) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2\right\} \\
&\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\}
\end{aligned}$$

la cual corresponde a la distribución $N(\bar{y}, \sigma^2/n)$. De esta forma, concluimos que el estimador Bayesiano de θ es $\bar{y}$ cuando no hay información previa, la cual coincide con el estimador clásico. Adicionalmente, un intervalo de credibilidad de θ se pueden obtener calculando los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución

$N(\bar{y}, \sigma^2/n)$, los cuales son idénticos a los valores $\bar{y} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$. Es decir, cuando se utiliza la previa no informativa de Jeffreys para θ, el intervalo de credibilidad coincide con el intervalo de confianza de la inferencia clásica. Lo anterior también se puede obtener haciendo la varianza de la distribución previa infinitamente grande, es decir $\tau^2 \to \infty$, entonces es posible hacer su distribución tan plana y vaga como sea posible, y tenemos que $\mu_n^2 = \bar{y}$ y $\tau_n^2 = \sigma^2/n$.

Resultado 2.6.5. *La distribución predictiva previa para una observación* $\mathbf{y}$ *es*

$$\mathbf{y} \sim Normal(\mu, \tau^2 + \sigma^2/n)$$

Prueba. Por la suficiencia del estimador $\bar{Y}$, se tiene que la función de verosimilitud de la muestra $\mathbf{y}$ es igual a la función de verosimilitud de $\bar{Y}$. Entonces tenemos que

$$\begin{aligned} p(\bar{Y}) &= \int p(\bar{Y} \mid \theta) p(\theta \mid \mu, \tau^2)\, d\theta \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{1}{2\sigma^2/n}(\bar{y}-\theta)^2\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\} d\theta \end{aligned}$$

Por otro lado, tenemos que

$$\begin{aligned} &\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(\bar{y}-\theta)^2}{\sigma^2/n}\right] \\ &= \frac{1}{2}\left[\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\theta^2 - 2\left(\frac{\mu}{\tau^2} + \frac{n\bar{y}}{\sigma^2}\right)\theta + \left(\frac{\mu^2}{\tau^2} + \frac{n\bar{y}^2}{\sigma^2}\right)\right] \\ &= \frac{1}{2\tau_n^2}\left[\theta^2 - 2\tau_n^2\left(\frac{\mu}{\tau^2} + \frac{n\bar{y}}{\sigma^2}\right)\theta + \tau_n^4\left(\frac{\mu}{\tau^2} + \frac{n\bar{y}}{\sigma^2}\right)^2\right] + \frac{1}{2}\left(\frac{\mu^2}{\tau^2} + \frac{n\bar{y}^2}{\sigma^2}\right) - \frac{\tau_n^2}{2}\left(\frac{\mu}{\tau^2} + \frac{n\bar{y}}{\sigma^2}\right)^2 \\ &= \frac{1}{2\tau_n^2}\left[\theta - \tau_n^2\left(\frac{\mu}{\tau^2} + \frac{n\bar{y}}{\sigma^2}\right)\right]^2 + \frac{1}{2}\left[\left(\frac{n}{\sigma^2} - \frac{n^2\tau_n^2}{\sigma^4}\right)\bar{y}^2 - 2\frac{n\mu\tau_n^2}{\tau^2\sigma^2}\bar{y} + \left(\frac{\mu^2}{\tau^2} - \frac{\mu^2\tau_n^2}{\tau^4}\right)\right] \\ &= \frac{1}{2\tau_n^2}\left[\theta - \mu_n\right]^2 + \frac{1}{2}\left[\frac{1}{\tau^2+\sigma^2/n}\bar{y}^2 - 2\frac{\mu}{\tau^2+\sigma^2/n}\bar{y} + \frac{\mu^2}{\tau^2+\sigma^2/n}\right] \\ &= \frac{1}{2\tau_n^2}\left[\theta - \mu_n\right]^2 + \frac{1}{2(\tau^2+\sigma^2/n)}(\bar{y}-\mu)^2. \end{aligned}$$

Entonces

$$\begin{aligned} p(\bar{Y}) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{\sigma^2/n}\sqrt{\tau^2}} \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right\} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2/n)}(\bar{y}-\mu)^2\right\} d\theta \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{n\tau_n^2}}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2/n)}(\bar{y}-\mu)^2\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\tau_n^2}} \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right\} d\theta \\ &= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2/n)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2/n)}(y-\mu)^2\right\} \end{aligned}$$

■

Resultado 2.6.6. *La distribución predictiva posterior para una nueva observación* $\tilde{\mathbf{y}}$ *es*

$$\tilde{\mathbf{y}} \mid \mathbf{Y} \sim Normal(\mu_n, \tau_n^2 + \sigma^2/n)$$

Prueba. Análogamente a la demostración del anterior resultado. ■

Cuando se utiliza la distribución previa no informativa de Jeffreys, se puede encontrar que la distribución predictiva del promedio de una nueva muestra de tamaño n^* está dada por $N(\bar{y}, \sigma^2(\frac{1}{n} + \frac{1}{n^*}))$. (Ejercicio XXXXXX)

Ejemplo 2.6.1. *La Superintendencia Nacional de Salud en uso de sus facultades constitucionales y legales controla y vigila los diferentes entes prestadores de Salud en todo el territorio colombiano. Por facultades que le otorga la ley esta entidad pretende evaluar la cantidad de consultas realizadas durante un mes por parte del personal de salud. Para esto, se obtuvo información sobre el desarrollo de las consultas de los profesionales durante un mes objetivo del año 2011.*

Debido a las quejas que se presentan a diario en la Superintendencia por cuestiones de solicitud y atención de citas, es de interés conocer la productividad mensual real de los profesionales de la salud. Puesto que el fin principal de la investigación es tomar acciones al respecto, la Superintendencia visitó, sin previo aviso, a un hospital de la ciudad y recolectó los datos de la atención real (en minutos) para una muestra de $n = 25$ *consultas de los profesionales que laboran allí.*

La hipótesis de la Superintendencia de Salud es que, para maximizar los recursos, el hospital no se toma el debido tiempo de atención a sus pacientes. Luego, el ente regulador comparará la atención real con la atención ideal (que por ley no debe ser menor a 18 minutos) de tiempo en consultas programadas, para así tomar las debidas medidas sancionatorias en caso de ser cierto. Los datos recolectados (minutos) fueron los siguientes: 15.6, 16.5, 19.5, 9.6, 19.6, 7.5, 10.0, 4.8, 10.8, 12.7, 20.1, 14.0, 3.2, 19.7, 7.5, 14.3, 8.0, 12.1, 4.4, 13.4, 11.8, 7.9, 7.7, 17.4 y 10.4. Por estudios previos, se conoce que la desviación estándar de las consultas debe ser menor a 5 minutos. Según la información recolectada, el promedio de tiempo en consulta fue de 11.9 minutos con una desviación estándar de 5.02.

Para este escenario, y dado las consecuencias del estudio, es conveniente utilizar una distribución previa no informativa, cuyos parámetros serán $\mu = 0$ *y* $\tau^2 = 10000$. *Recurriendo al resultado XXXXXXXX, la distribución posterior para el parámetro de interés,* θ, *es* $Normal(11.9, 1^2)$. *Para esta distribución, un intervalo de credibilidad del 99 % es (9.32, 14.47). Lo cual indica que efectivamente sí hay indicios de que el tiempo promedio de las consultas es menor al establecido por la ley. El siguiente código puede ser utilizado para realizar esta inferencia con el programa* `JAGS`.

Código JAGS

```
model
{
for(j in 1 : J) {
y[j] ~ dnorm(theta, 0.04)
```

```
}
theta ~ dnorm(0,0.0001)
}

DATA
list(J = 25, y = c(15.6, 16.5, ... , 10.4))
```

2.7 Normal con varianza desconocida y media conocida

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ conforma una muestra aleatoria de variables con distribución $Normal(\theta, \sigma^2)$, con θ conocido y σ^2 desconocido. Entonces, la función de distribución conjunta o verosimilitud de los datos está dada por

$$p(\mathbf{Y} \mid \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} n S_\theta^2\right\} I_{\mathbb{R}^n}(\mathbf{y}) \tag{2.7.1}$$

con $nS_\theta^2 = \sum_{i=1}^n (y_i - \theta)^2$. Como σ^2 debe ser positivo, entonces es posible asignarle al parámetro σ^2 una distribución previa $Inversa - Gamma(\alpha, \beta)$ dada por

$$p(\sigma^2 \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{-\beta/(\sigma^2)} I_{(0,\infty)}(\sigma^2). \tag{2.7.2}$$

Resultado 2.7.1. *La distribución posterior del parámetro σ^2 sigue una distribución*

$$\sigma^2 \mid \mathbf{Y} \sim Inversa - Gamma(\alpha + n/2, \beta + nS_\theta^2/2).$$

Prueba.

$$\begin{aligned} p(\sigma^2 \mid Y) &\propto p(Y \mid \sigma^2) p(\sigma^2 \mid \alpha, \beta) \\ &\propto (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} n S_\theta^2} (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2} \\ &= (\sigma^2)^{-(\alpha+n/2)-1} e^{-\left(\beta + nS_\theta^2/2\right)/\sigma^2} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Inversa - Gamma(\alpha + n/2, \beta + nS_\theta^2/2)$. ■

Resultado 2.7.2. *La distribución predictiva previa para una muestra $\mathbf{y} = \{y_1, \ldots, y_n\}$ está dada por*

$$p(\mathbf{Y}) = \frac{\Gamma(\alpha + n/2)}{(2\pi)^{n/2}\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + nS_\theta^2/2)^{\alpha+n/2}} I_{\mathbb{R}^n}(\mathbf{y}) \tag{2.7.3}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned}
p(\mathbf{Y}) &= \int p(\mathbf{Y} \mid \sigma^2) p(\sigma^2 \mid \alpha, \beta)\ d\sigma^2 \\
&= \int_0^\infty \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} n S_\theta^2\right\} I_{\mathbb{R}^n}(\mathbf{y}) \\
&\qquad \times \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{-\beta/(\sigma^2)} I_{(0,\infty)}(\sigma^2)\ d\sigma^2 \\
&= \frac{1}{(2\pi)^{n/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} I_{\mathbb{R}^n}(\mathbf{y}) \int_0^\infty (\sigma^2)^{-(\alpha+n/2)-1} e^{-\left(\beta + n S_\theta^2/2\right)/\sigma^2}\ d\sigma^2 \\
&= \frac{1}{(2\pi)^{n/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + n/2)}{(\beta + n S_\theta^2/2)^{\alpha+n/2}} I_{\mathbb{R}^n}(\mathbf{y}) \\
&\qquad \times \int_0^\infty \frac{(\beta + n S_\theta^2/2)^{\alpha+n/2}}{\Gamma(\alpha + n/2)} (\sigma^2)^{-(\alpha+n/2)-1} e^{-\left(\beta + n S_\theta^2/2\right)/\sigma^2}\ d\sigma^2 \\
&= \frac{1}{(2\pi)^{n/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + n/2)}{(\beta + n S_\theta^2/2)^{\alpha+n/2}} I_{\mathbb{R}^n}(\mathbf{y})
\end{aligned}$$

■

Resultado 2.7.3. *Después de la recolección de los datos, la distribución predictiva posterior para una nueva muestra* $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{n^*}\}$*, de tamaño* n^**, está dada por*

$$\begin{aligned}
p(\mathbf{Y}) &= \frac{\Gamma(\alpha + n/2 + n^*/2)}{(2\pi)^{n^*/2}\Gamma(\alpha + n/2)} \\
&\qquad \times \frac{(\beta + n S_\theta^2/2)^{\alpha+n/2}}{(\beta + n^* S_{\theta^*}^2/2 + n S_\theta^2/2)^{\alpha+n/2+n^*/2}} I_{\mathbb{R}^n}(\tilde{\mathbf{y}}) \qquad (2.7.4)
\end{aligned}$$

con $S_{\theta^*}^2 = \sum_{i=1}^{n^*} (\tilde{y}_i - \theta)^2$.

Prueba. De la definición de función de distribución predictiva, y haciendo uso del mismo razonamiento en la demostración del anteriori resultado, se tiene la prueba inmediatamente. ■

Nótese que en la definición de la distribución previa del parámetro de interés σ^2 se utilizaron dos hiper-parámetros α y β cualesquiera. Sin embargo, **?** afirma que una hiper-parametrización conveniente está dada por la escogencia de $\alpha = n_0/2$ y $\beta = n_0\sigma_0^2/2$, donde n_0 denota el número de datos previos y σ_0^2 una estimación previa del parámetro de interés σ^2. En este caso, la distribución posterior de σ^2 toma la siguiente forma

$$\sigma^2|\mathbf{Y} \sim \ Inversa - Gamma(\frac{n_0 + n}{2}, \frac{n_0\sigma_0^2 + n S_\theta^2}{2})$$

De esta forma, la estimación bayesiana de σ^2 viene dada por

$$\hat{\sigma^2}_B = \frac{n_0\sigma_0^2 + n S_\theta^2}{n_0 + n - 2}$$

Obviando en el denominador el término -2, el cual es un valor pequeño para valores grandes de n_0 y n, podemos ver que

$$\begin{aligned}\hat{\sigma^2}_B &\approx \frac{n_0\sigma_0^2 + nS_\theta^2}{n_0 + n} \\ &= \frac{n_0}{n_0 + n}\sigma_0^2 + \frac{n}{n_0 + n}S_\theta^2\end{aligned}$$

que es un promedio ponderado entre la estimación previa σ_0^2 y la estimación clásica S_θ^2 con pesos proporcionales a los tamaños n_0 y n. En caso de que no exista ninguna información previa para ser incorporada en los procedimientos bayesianos, se puede hallar la distribución previa no informativa de Jeffreys. Para eso, tenemos que

$$\log p(Y|\sigma^2) = \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - \theta)^2$$

De esta forma, la información de Fisher está dada por

$$I(\sigma^2) = \frac{1}{2\sigma^4}$$

Y podemos concluir que la distribución previa no informativa de Jeffreys está dada por

$$p(\sigma^2) \propto (\sigma^2)^{-1}$$

la cual es una distribución impropia, sin embargo, por medio de operaciones algebraicas similares a las del Resultado 2.7.1, se encuentra que la resutante distribución posterior de σ^2 está dada por

$$\sigma^2|\mathbf{Y} \sim \ Inversa - Gamma(\frac{n}{2}, \frac{nS_\theta^2}{2})$$

Y podemos que la estimación bayesiana de σ^2 está dad por $\sigma_0^2 = \frac{nS_\theta^2}{n-2}$, similar, aunque siempre mayor, a la estimación frecuentista dada por S_θ^2.

Ejemplo 2.7.1. *En (?, pg. 73), se analizó un grupo de datos que corresponden al grosor de 12 láminas de una línea de producción de vidrio templado de 3cm de grosor. Es claro que la calidad de las láminas producidas está íntimamente ligada con la varianza del proceso, puesto que una varianza grande implica que hay alta disparidad con respecto al grosor, es por esta razón que se desea estimar el parámetro varianza teniendo en cuenta que la media μ es conocida y toma el valor de 3 cm.*

Dado que no disponemos de alguna información previa, utilizaremos una distribución previa no informativa de Jeffreys para el parámetro σ^2. Después de la recolección de los datos, se encontró que el promedio muestral $\bar{Y} = 3.18$ y además $S_\theta^2 = 0.1309$. De esta forma, la distribución posterior para σ^2 está dada por una distribución $Inversa - Gamma(6, 0.7853)$, y la estimación bayesiana de σ^2 sería 0.15706. Y por consiguiente, la estimación bayesiana de la desviación estándar viene dada por 0.4cm.

En JAGS, la parametrización de una distribución normal está dada por la media y por la dispersión. Este último parámetro está definido como el inverso multiplicativo de la varianza. Como se puede ver en el apéndice XXXXXXXXXXXXX, dado que σ^2 tiene distribución previa Inversa-gamma, entonces este parámetro de dispersión tiene distribución previa Gamma. De esta forma, una buena escogencia de distribución previa para el parámetro de dispersión es $Gamma(0.001, 0.001)$, la cual está centrada en uno y tiene varianza mil; es decir está bien dispersa y puede considerase como una distribución no informativa. El siguiente código arroja una estimación puntual de 0.1572 y un intervalo de credibilidad del 95 % de $(0.0665, 0.3665)$.

Código JAGS

```
model{
for(j in 1 : J) {
y[j] ~ dnorm(3, disp)
}
disp ~ dgamma(0.001, 0.001)
sigma <-1 /disp
}

DATA
list(J = 12, y = c(3.56, 3.36, 2.99, 2.71, 3.31,
3.68, 2.78, 2.95, 2.82, 3.45, 3.42,3.15))

INITS
list(disp=1)
```

Ejemplo 2.7.2. *Continuando con el ejemplo anterior, suponga que se tiene conocimiento acerca del comportamiento de un lote de láminas de tamaño 15 del mes anterior. Para este lote, se obtuvo que el grosor promedio es de 2.89 cm, y la desviación estándar es de 0.4 cm. Por consiguiente, en el análisis bayesiano podemos incorporar estos datos como la información previa escogiendo* $n_0 = 15$ *y* $\sigma_0^2 = 0.4^2 = 0.16$. *Desde luego, los parámetros de la distribución previa serán* $\alpha = n_0/2 = 15/2 = 7.5$ *y* $\beta = n_0\sigma_0^2/2 = 7.5 \times 0.16 = 1.2$. *De esta forma, encontramos que la distribución posterior del parámetro* σ^2 *es* $Inversa - Gamma(13.5, 1.9853)$ *y encontramos que la estimación bayesiana de* σ^2 *está dada por 0.1588.*

2.8 Ejercicios

1. Verifique las identidades (**??**) y (**??**).
2. Demuestre que para una muestra aleatoria $y_1, \cdots, y_n|\theta \sim N(\theta, \sigma^2)$ con σ^2 conocida, cuando se utiliza la distribución previa no informativa de Jeffreys para θ, la distribución predictiva del promedio de una nueva muestra de tamaño n^* está dada por $N(\bar{y}, \sigma^2(\frac{1}{n} + \frac{1}{n^*}))$.

3 Computación bayesiana

3.1 Introducción

`R` es un software de uso libre que maneja un ambiente de programación enfocado al manejo de matrices y por lo tanto muy apropiado en un curso de Análisis bayesiano. Usando una serie de ejemplos, se describe cómo `R` puede ser usado como una herramienta efectiva; estos ejemplos tienen un énfasis especial en tópicos bayesianos, específicamente en el uso de las cadenas de Markov para simular distribuciones posterior. En la segunda sección se tratan tópicos bayesianos a través de las cadenas de Markov vía Monte Carlo. `R` es un software de libre distribución disponible en `http://cran.r-project.org`.

3.2 Monte Carlo vía cadenas de Markov

El enfoque bayesiano requiere que las inferencias estadísticas estén basadas en la distribución posterior; per, a veces el tratamiento algebraico, o incluso numérico, de esta distribución es complicado. Sin embargo, si estuviéramos en la capacidad de simular de la distribución posterior, podríamos usar un enfoque de Monte carlo para hacer inferencia. Por ejemplo, si estuviéramos interesados en la distribución posterior de $\theta = \phi(\alpha)$ donde ϕ es una función conocida, entonces podemos simular valores $\alpha_1, \alpha_2, ..., \alpha_N$ de $p(\alpha \mid x)$, la posterior de α dado x y entonces construir $\theta_1, ..., \theta_N$, donde $\theta_i = \phi(\alpha_i)$. Los valores $\theta_1, ..., \theta_N$ son valores de la distribución posterior de θ dado x - los intervalos para θ pueden ser generados de esta muestra, entre más grande N, más preciso el intervalo.

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

Ahora, deseamos simular de una distribución $p(\alpha \mid x)$, hacerlo directamente puede resultar un poco dificil. En vez de esto, la idea MCMC es fijar una cadena de Markov que tenga una distribució estacionaria igual a la distribución de interés, $p(\alpha \mid x)$. Luego simulamos desde esta y la iteramos hasta convergencia (en distribución), y entonces usamos los valores simulados como observaciones de $p(\alpha_| x)$.

Existen dos dificultades: En primer lugar, cómo fijar tal cadena de Markov; y

en segundo lugar, conocer el momento en el cual converge. Hay dos métodos para fijar la cadena de Markov: el muestreador de Gibbs y el algoritmo de Metropolis-Hastings. Por otro lado la convergencia de las cadenas de markov es motivo de investigación actual, por lo tanto aquí solo examinaremos los métodos anteriormente mencionados.

Notación: nuestro propósito es correr una cadena de Markov cuyas observaciones son $\alpha_1, \alpha_2, \cdots$, donde la distribución estacionaria de la cadena es $p(\alpha \mid x)$. Suponga que α es d-dimensional. Denotamos por α_i el i-ésimo componente de α y $\alpha_i^{(1)}, \alpha_i^{(2)}, \cdots, \alpha_i^{(d)}$ los d componentes del vector. Sea

$$P(\alpha^{(i)} \mid \alpha^{(1)}, \cdots, \alpha^{(i-1)}, \alpha^{(i+1)}, \cdots, \alpha^{(d)}, x \propto p(\alpha \mid x)$$

la distribución univariada del i-ésimo componente del vector condicionada a las otras componentes y a las observaciones de x.

3.3 El muestreador de Gibbs

≪*Este procedimiento es apropiado para obtener muestras de una distribución conjunta cuando es fácil muestrear de las distribuciones condicionadas.*≫ Peña pg 333. El procedimiento se implementa como sigue:

Sean $\alpha_i = (\alpha_i^{(1)}, ..., \alpha_i^{(d)})$ los valores actuales de α. Entonces α_{i+1} se obtiene así:

- Generar $\alpha_{i+1}^{(1)}$ de $p(\alpha^{(1)} \mid \alpha_i^{(2)}, ..., \alpha_i^{(d)}, x)$
- Generar $\alpha_{i+1}^{(2)}$ de $p(\alpha^{(2)} \mid \alpha_{i+1}^{(1)}, \alpha_i^{(3)}, ..., \alpha_i^{(d)}, x)$

  $\vdots$

- Generar $\alpha_{i+1}^{(d)}$ de $p(\alpha^{(d)} \mid \alpha_{i+1}^{(1)}, \alpha_{i+1}^{(2)}, ..., \alpha_{i+1}^{(d-1)}, x)$

La idea de este esquema es renovar cada componente por medio de la simulación de la correspondiente distribución condicional.

Importante: Una vez que la cadena converge, se tiene que los valores de α corresponden observaciones de la distribución posterior requerida, $p(\alpha \mid x)$. Sin embargo, en general, no se garantiza una muestra independiente de $p(\alpha \mid x)$, dado que el esquema del muestreador de Gibbs usa el valor actual para construir el siguiente valor. La secuencia de valores que se obtiene estará correlacionada.

Por ejemplo, se puede implementar el muestreador de Gibbs para generar una secuencia de observaciones con densidad conjunta $f(x, y)$ que corresponde a $(x, y) \sim$

$N_2\left(0, \begin{pmatrix} \rho & 0 \\ 0 & \rho \end{pmatrix}\right)$, por tanto el procedimiento es generar

$$x_{t+1} \mid y_t \sim N(\rho y_t, 1 - \rho^2)$$

$$y_{t+1} \mid x_{t+1} \sim N(\rho x_{t+1}, 1 - \rho^2)$$

Este resultado es inmediato, ya que la media de ambas variables es cero y su varianza uno; por tanto la covarianza entre ambas variables es ρ; esto se encuentra en **?**, p. 286

```
  bivariate.gibbs <- function (n, rho, x, y)
{
mat <- matrix(ncol = 2, nrow = n)
mat[1,] <- c(x, y)
for (i in 2:n){
x <- rnorm(1, rho * y, sqrt(1 - rho^2))
y <- rnorm(1, rho * x, sqrt(1 - rho^2))
mat[i, ] <- c(x, y)
}
mat<-as.data.frame(mat)
mat
}

  biv<-bivariate.gibbs(n=10000, rho=0.5, 0,0)
colMeans(biv)
var(biv)
cor(biv)
plot(biv)
```

Un problema común es el de descartar los m primeros valores, puesto que el algoritmo puede demorar en obtener convergencia; esto se puede resolver en forma empírica utilizando las medias y varianzas acumuladas y traficándolas se puede tomar una decisión acerca del valor m.

```
  g.diag <- function(sample){
n <- length(sample); res <- matrix(nrow=2, ncol=n)
for(i in 1:n){
res[1,i] <- mean(sample[1:i])
res[2,i] <- var(sample[1:i])
}
res
}
biv<-bivariate.gibbs(n=1000, rho=0.5, 0,0)
m1 <- g.diag(biv[,1])
m2 <- g.diag(biv[,2])
```

```
par(mfcol=c(2,1))
plot(m1[1,], type='l', ylim=c(-0.6,0.6), col=4)
lines(m2[1,], lty=2,col=2)
title("Diagnostico: Media acumulada")
plot(m1[2,], type='l', ylim=c(0.5,1.5),col=4)
lines(m2[2,], lty=2,col=2)
title("Diagnostico: Varianza acumulada")
```

Figura 3.1: *Convergencia de la media y varianza usando Gibbs*

Pero el muestrador de Gibbs también funciona en una "segunda fase", cuando queremos seleccionar una muestra de $f(\theta \mid x)$, es decir, la distribución de los parámetros dada la información observada x. En el siguiente ejemplo x tiene distribución $N(\mu, \sigma^2 = 1/\phi)$ y queremos obtener una muestra del vector aleatorio $\theta = (\mu, 1/\phi)$. Para este caso supongamos que conocemos las distribuciones previa; para la media μ uniforme y para la varianza ϕ una Gamma con parámetros a y b.

La distribución posterior de (μ, σ^2) satisface:

$$p(\mu, \phi \mid x) \propto (\phi)^{n/2} exp\left\{-\phi \frac{\sum_{j=1}^{n}(x_j - \mu)^2}{2}\right\} (\phi)^{a-1} exp(-b/\phi),$$

donde la primera parte después del signo de proporcionalidad, corresponde a la verosimilitud de la información observada y la segunda parte corresponde a la distribución posterior de ϕ; la distribución posterior de μ no aparece pues es una constante. Por tanto, ésta se puede escribir como:

$$\propto (\phi)^{n/2+a-1} exp\left\{-\phi\Big(\frac{\sum_{j=1}^{n}(x_j - \mu)^2}{2} + b\Big)\right\}$$

La distribución condicional de la varianza σ^2 dado $\{\mu, x\}$ es Gamma inversa con parámetros $\alpha + n/2$ y $\sum_{j=1}^{n}(x_j - \mu)^2/2 + b$, es decir,

$$\sigma^2 \mid \mu, x \sim GI\bigg(\alpha + n/2, \sum_{j=1}^{n}(x_j - \mu)^2/2 + b\bigg).$$

Análogamente, tenemos que la distribución de μ dado $\{\sigma^2, x\}$ es normal con media $\bar{x}$ y varianza σ^2/n, es decir,

$$\mu \mid \sigma^2, x \sim N(\bar{x}, \sigma^2/n).$$

Entonces, para implementar el muestreador de Gibbs, primero debemos escoger valores apropiados para a y b, con el propósito de representar correctamente la distribución previa, y luego

- Tomar un valor inicial para la media y la varianza, (μ_0, σ_0^2)
- Generar $(\mu_{i+1}, \sigma_{i+1}^2)$ simulando μ_{i+1} de $N(\bar{x}, \sigma_i^2/n)$ y luego σ_{i+1}^2 de $GI\Big(\alpha + n/2, \sum_{j=1}^{n}(x_j - \mu_{i+1})^2/2\Big)$
- Repetir para obtener $(\mu_0, \sigma_0^2), (\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \cdots$,
- Descartar los m primeros valores, es decir, suponga que el algoritmo converge después de m iteraciones.
- Entonces $(\mu_{m+1}, \sigma_{m+1}^2), (\mu_{m+2}, \sigma_{m+2}^2), \cdots$, es una muestra (correlacionada) de $p(\mu, \sigma^2 \mid x)$

```
  normal2 <- function(datos, a, b, N,inicial)
{
ndat <- length(datos)
xbar <- mean(datos)
mu.now <- inicial[1]
var.now <- inicial[2]
dummy <- matrix(ncol=2,nrow=N)
dummy[1,1] <- mu.now
dummy[1,2] <- var.now
for (i in 2:N) {
mu.next <- rnorm(1,xbar,sqrt(var.now/ndat))
var.next <- 1/(rgamma(1,a+ndat/2)/(b+sum((datos-mu.next)^2)/2))
dummy[i,1] <- mu.next
dummy[i,2] <- var.next
mu.now <- mu.next
var.now <- var.next
}
answer <- dummy
answer
}
observed <- rnorm(10,2,1)
mc1.vals <- normal2(observed,10,10,1000,c(5,5))
mc1.vals <- mc1.vals[101:1000,]
colMeans(mc1.vals)
par(mfcol=c(2,1))
plot(mc1.vals[,1],type='l',ylab = 'mean')
plot(mc1.vals[,2],type='l',ylab = 'variance')
windows()
par(mfcol=c(2,1))
```

```
hist(mc1.vals[,1],prob=T,xlab='mean',breaks=20)
lines(density(mc1.vals[,1],kernel='gaussian',adjust=1.8))
hist(mc1.vals[,2],prob=T,xlab='variance',breaks=20)
lines(density(mc1.vals[,2],kernel='gaussian',adjust=1.8))
```

3.4 El algoritmo de Metropolis

3.5 El algoritmo de Metropolis-Hastings

Este algoritmo se basa en proponer un nuevo punto de acuerdo a una función de densidad de probabilidad arbitraria y aceptar este valor propuesto con una probabilidad que depende del punto actual, el nuevo punto y la densidad de la cual fue propuesto el nuevo punto.

Suponga que deseamos simular de la posterior multivariada $p(\alpha \mid x)$. Sea $q(\alpha, \alpha')$ una función de densidad de probabilidad arbitraria que describe la probabilidad de aceptación de α' dado que la posición actual es α. La densidad q puede ser escogida. El algoritmo de Metropolis-Hastings está dado por:

- Sea el valor actual α_i, generar un valor candidato α' donde α' se obtiene como una observación de la densidad $q(\alpha_i, \alpha')$.
- Calcule

$$T(\alpha_i, \alpha') = \begin{cases} min[\frac{p(\alpha'|x)q(\alpha',\alpha_i)}{p(\alpha_i|x)q(\alpha_i,\alpha')}] & \text{si} p(\alpha_i \mid x)q(\alpha_i, \alpha') > 0, \\ 1 & \text{si} p(\alpha_i \mid x)q(\alpha_i, \alpha') = 0 \end{cases}$$

- La probabilidad de aceptar el candidato es $T(\alpha_i, \alpha')$. Fijar $\alpha_{i+1} = \alpha'$. De otra forma, rechazar el candidato y fijar $\alpha_{i+1} = \alpha_i$.
- Repita el paso anterior para obtener la secuencia $\alpha_0, \alpha_1, ...$, donde α_0 denota un valor arbitrario de arranque.
- Descarte los primeros m valores obtenidos. Entonces $\alpha_{m+1}, \alpha m + 2, ...$ es una secuencia (correlacionada) de la distribución posterior que se requería.

En principio, puede ser usada cualquier densidad q, pero si ésta es escogida ingenuamente, la eficiencia de la cadena puede ser muy pobre. Nótese que los valores de salida pueden ser los mismos que loas valores iniciales.

Siguiendo con el segundo ejemplo del apartado de Gibbs, en donde contábamos con 10 datos que tenían una distribución normal con media 2 y varianza 1, se ha escogido usar como q (distribuciones propuestas), normales centradas en el actual parámetro. Dadas las distribuciones propuestas, algunos valores de la varianza pueden ser negativos; este no es un problema porque la posterior toma el valor cero, por tanto este valor será aceptado con un probabilidad de cero. En el libro de Robert y Casella hay muchos ejemplos sobre las densidades q. En la figura 6, se observa una lenta y convergencia.

La relación más importante entre el muestreados de Gibbs y el algorítmo de Metropolis-Hastings, está dada como un teorema en el libro de Robert y Casella (pg 296).

Resultado 3.5.1. *El muestreador de Gibbs es equivalente al algorítmo de Metropolis-Hastings, con la probabilidad de aceptación igual a uno para todos los puntos propuestos.*

Lo anterior implica que la convergencia para ambos métodos no es la misma. Para cerrar la sección de cadenas de Markov vía Monte Carlo, es importante hacernos la siguiente pregunta: ¿Son independientes las muestras simuladas?. En principio no se puede hablar de independencia, pues es claro que la observación $\{i+1\}$ depende de la observación $\{i\}$. Dado que las observaciones resultantes se encuentran en estricto orden de medición, podríamos utilizar algunos criterios como ACF y PACF, para averiguar sobre la correlación entre observaciones.

La figura 7 meustra el gráfico de autocorrelación parcial de los datos para los dos métodos, en general no hay una estructura definida de autocorrelación parcial. La figura 8 muestra que para el caso del algoritmo Metropolis-Hastings, si existe una fuerte autocorrelación entre las primeras 100 observaciones, mientras que para el muestreador de Gibbs tal estructura no está definida. Este podría ser otro criterio de selección de la observación m-ésima.

3.6 Algoritmos híbridos

MEtropolis withing Gibbs

3.7 Diagnóstico de convergencia

3.8 Cadenas de Markov reversibles

3.9 Códigos en R

En esta sección se implementan cada uno de los algorítmos utilizados en los ejemplos desarrollados.

Algoritmo de METROPOLIS HASTING

```
  met.hast.1 <- function(datos, a, b, iter, ini, v.mean, v.var) {
mu0<- ini[1]; var0<- ini[2]
resul <- matrix(ncol=2,nrow=iter)
resul[1,1] <- mu0; resul[1,2] <- var0
for (i in 2:iter) {
```

```
mu.prop <- rnorm(1, mu0, sqrt(v.mean)) # Propuesta para la media
var.prop <- rnorm(1, var0, sqrt(v.var))# Propuesta para la varianza
if (var.prop <= 0) {
# Vigila varianzas negativas
T.val <- 0 # Nunca acepta }
else {
T1 <- prod(dnorm(datos,mu.prop,sqrt(var.prop)))
* dgamma(b/var.prop,a) * b* ( (1/var.prop)2̂ )
* dnorm(mu0, mu.prop,sqrt(v.mean)) * dnorm(var0,var.prop,sqrt(v.var))
T2 <- prod(dnorm(datos,mu0,sqrt(var0))) * dgamma(b/var0,a) * b
* ( (1/var0)2̂ ) * dnorm(mu.prop, mu0, sqrt(v.mean))
* dnorm(var.prop, var0,sqrt(v.var))
T.val <- min(1, T1/T2 ) }
u <- runif(1) if (u <= T.val) {
resul[i,1] <- mu.prop; resul[i,2] <- var.prop }
else {
resul[i,1] <- mu0; resul[i,2] <- var0 }
mu0<- resul[i,1]; var0<- resul[i,2]; }
final <- resul final }
mc2 <- met.hast.1(datos,10,10,1000,c(5,5),0.15,0.08)
```

Independencia entre selecciones

```
  mc1[,1]->Mu_Gibbs; mc1[,2]->Var_Gibbs
mc2[,1]->Mu_MH; mc2[,2]->Var_MH
par(mfrow=c(2,2))
pacf(Mu_Gibbs,1000)
pacf(Mu_MH,1000, ylim=c(-0.1,0.2))
pacf(Var_Gibbs,1000)
pacf(Var_MH,1000, ylim=c(-0.1,0.2))
par(mfrow=c(2,2))
acf(Mu_Gibbs,1000,ylim=c(-0.1,0.2))
acf(Mu_MH,1000 )
acf(Var_Gibbs,1000,ylim=c(-0.1,0.2))
acf(Var_MH,1000)
```

3.10 Metodología

En los próximos apartados el lector podrá identificar los pasos que se deberán seguir para la consecución de los objetivos señalados al principio de esta propuesta

doctoral.

3.10.1 Identificación

En primer lugar, cada modelo propuesto debe ser totalmente identificable en el sentido de **?**, p. 60. Lo anterior implica que para cada escenario planteado, las matrices de información auxiliar deben estar correctamente estructuradas tanto en los componentes fijos, como en los componentes aleatorios. Una vez, asegurada esta propiedad, se utilizará el enfoque bayesiano de proponer distribuciones previa conjugadas siempre que sea posible. Dado que este acercamiento a la modelación conjunta requiere un arduo trasfondo computacional, cuando las distribuciones posterior no parezcan tener una forma conocida o cerrada, se utilizará el enfoque propuesto mediante la implementación del algoritmo híbrido.

3.10.2 Computación

En total armonía con lo anterior, los parámetros del modelo (efectos fijos, efectos aleatorios y efectos de varianza) estarán dados por un vector por bloques. Para la estimación de estas cantidades, es necesario utilizar el enfoque propuesto mediante la implementación de sendas cadenas simuladas desde las distribuciones posterior y con la técnica de Gibbs, en caso que sean conocidas; o mediante la utilización de distribuciones de salto para la implementación de un algoritmo de tipo Metrópolis. De esta manera, se debe utilizar en cada iteración un algoritmo híbrido que permita la actualización de las cadenas paso a paso. En términos pragmáticos, el autor de esta propuesta doctoral utilizará el software estadístico R (**?**) para la implementación de tales algoritmos en cada uno de los escenarios propuestos.

3.10.3 Estimación

Aunque no será el quid de la investigación, y tampoco se abordará a profundidad las reglas de decisión derivadas de la teoría de decisión (**?**), es importante resaltar que en este trabajo doctoral, se utilizará el criterio de mínima pérdida cuadrática para la realización de cualquier tipo de inferencia ya sea de tipo estimativo o predictivo. Con base en lo anteriormente expuesto, se tiene que, siempre y cuando sea posible realizar la integración, las estimaciones puntuales estarán dadas por la esperanza del parámetro basada en la distribución posterior inducida por el análisis respectivo. Para los parámetros cuya distribución posterior no tiene una forma conocida o cerrada, se utilizarán métodos numéricos que deban ser llevados a cabo hasta obtener convergencia. Estos métodos convergen a la moda, el valor que hace máxima la distribución posterior, y este será el criterio de estimación en estos casos. A continuación se hace un repaso del tratamiento involucrado en la construcción de las cadenas que se generan a partir de los algoritmos utilizados en la estimación de los parámetros.

Métodos MCMC

Dado que una gran parte del desarrollo de este proyecto está ligada a la programación e implementación de métodos de Monte Carlo para realizar inferencias posterior de los parámetros de interés, se seguirá el razonamiento y recomendaciones de **?**, que puede ser resumido en los siguientes ítemes para cada parámetro de interés:

- Simulación tres o más cadenas de forma paralela. Los valores iniciales de cada cadena deben estar dispersos entre sí.
- Comprobación de la convergencia de las cadenas mediante el descarte de la primera mitad de los valores generados en las cadenas. Esta etapa se conoce como *burning stage*.
- Una vez que las cadenas converjan, mezclar los tres conjuntos de valores generados por las cadenas. Esto garantiza, en primera instancia, que las cadenas no estén auto-correlacionadas.
- Además de realizar esta mezcla, descartar valores intermedios mediante un entero. Esta etapa se conoce como *thining stage*. Al final se recomienda almacenar una mediana cantidad de valores simulados.
- Calibración del algoritmo de muestreo si la convergencia no se presenta rápidamente. Para los algoritmos de Metropolis-Hastings, escoger una distribución de salto acorde con la distribución de la cual se desea simular. Por ejemplo, **?** presentan dos distribuciones de salto para el problema de la modelación de la varianza (cada una de las propuestas presenta tasas de aceptación diferentes).
- Comparación y contraste de los resultados con modelos simples que permitan examinar posibles discrepancias y corregir errores de programación.

En términos de inferencia bayesiana, se tienen dos tipos de procesos en este trabajo de investigación. El primero y más común, se trata de realizar inferencias acerca de un vector de parámetros de interés $\boldsymbol{\theta}$. El segundo, trata con los momentos del parámetro, por ejemplo su esperanza. Nótese que el primer proceso se presenta con seguridad en los ejercicios empíricos simulados; sin embargo, el segundo se presenta en los ejercicios prácticos con datos reales, en donde se quiere contrastar alguna hipótesis.

Las anteriores dos opciones tienen tratamientos muy diferentes en términos de la cantidad de simulaciones requeridas. Por ejemplo, si el objetivo es inferir acerca de $\boldsymbol{\theta}$, para conocer su comportamiento estructural, basta con realizar una simulación que genere una cantidad mediana de valores y que se resumen en un promedio y una desviación estándar. Por otro lado, si el objetivo es inferir acerca de $E(\boldsymbol{\theta})$, se requieren muchas más simulaciones para obtener una buena precisión. Siguiendo a **?**, una vez terminado el proceso de *burning* y *thining*, se dividan los valores simulados en las cadenas paralelas y se forman k grupos; de esta forma, una estimación de $E(\boldsymbol{\theta})$ será la gran media de las medias muestrales de cada grupo y el error estándar será su desviación estándar dividida en $\sqrt{k}$.

4 Modelos multiparamétricos

4.1 Normal con media y varianza desconocida

Cuando se desconoce tanto la media como la varianza de la distribución, es necesario plantear diversos enfoques y situarse en el más conveniente, según el contexto del problema planteado. En términos de la asignación de las distribuciones previa para θ y σ^2. En estos términos, es posible:

- Suponer que la distribución previa $p(\theta)$ es independiente de la distribución previa $p(\sigma^2)$ y que ambas distribuciones son informativas. Luego, utilizar un análisis de simulación condicional conjunta para extraer muestras provenientes de las respectivas distribuciones posterior.
- Suponer que la distribución previa $p(\theta)$ es independiente de la distribución previa $p(\sigma^2)$ y que ambas distribuciones son no informativas.
- Suponer que la distribución previa para θ depende de σ^2 y escribirla como $p(\theta \mid \sigma^2)$, mientras que la distribución previa de σ^2 no depende de θ y se puede escribir como $p(\sigma^2)$. El análisis posterior de este enfoque encuentra la distribución posterior de $\sigma^2 \mid \mathbf{Y}$ y con ésta se encuentra la distribución posterior de $\theta \mid \sigma^2, \mathbf{Y}$.

4.1.1 Parámetros independientes

Otro enfoque para el análisis de los parámetros de interés en una verosimilitud normal es suponer que las distribuciones previa de cada uno de los parámetros son independientes pero al mismo tiempo son informativas. (**?**) afirma que este supuesto de independencia es atractivo en problemas para los cuales la información previa para θ no toma la forma de un número fijo de observaciones con varianza σ^2. Cabe resaltar que este enfoque no es conjugado con respecto a la verosimilitud normal; es más aunque se asume independencia previa, en las distribuciones posterior resultantes θ y σ^2 son dependientes y la distribución posterior conjunta no tiene una forma estructural cerrada.

En este orden de ideas, y siguiendo la argumentación del capítulo anterior, la distribución previa para el parámetro θ es

$$\theta \sim Normal(\mu, \tau^2)$$

y la distribución previa para el parámetro σ^2 es

$$\sigma^2 \sim Inversa - gamma(n_0/2, n_0\sigma_0^2/2)$$

Asumiendo independencia previa, la distribución previa conjunta resulta estar dada por

$$p(\theta, \sigma^2) \propto (\sigma^2)^{-n_0/2-1} \exp\left\{-n_0\sigma_0^2/(2\sigma^2)\right\} \exp\left\{\frac{1}{2\tau^2}(\theta-\mu)^2\right\} \tag{4.1.1}$$

Una vez que se conoce la forma estructural de la distribución previa conjunta, es posible establecer la distribución posterior conjunta puesto que la verosimilitud de los datos, $p(\mathbf{Y} \mid \theta, \sigma^2)$, está dada por la expresión (3.1.7) y

$$p(\theta, \sigma^2 \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \theta, \sigma^2) p(\theta, \sigma^2)$$

Resultado 4.1.1. *La distribución posterior conjunta de los parámetros de interés está dada por*

$$\begin{aligned} p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto (\sigma^2)^{-(n+n_0)/2-1} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + n(\bar{y}-\theta)^2\right] - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \end{aligned} \tag{4.1.2}$$

Prueba. Utilizando la expresión (3.1.8), se tiene que

$$\begin{aligned} p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta, \sigma^2) p(\theta, \sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\} \\ &\qquad \times (\sigma^2)^{-n_0/2-1} \exp\left\{-n_0\sigma_0^2/(2\sigma^2)\right\} \exp\left\{\frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\ &\propto (\sigma^2)^{-(n+n_0)/2-1} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + n(\bar{y}-\theta)^2\right] - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \end{aligned}$$

■

Nótese que la distribución posterior conjunta no tiene una forma estructural conocida y por lo tanto no es posible realizar el método de integración analítica para obtener una constante de integración (**?**). Sin embargo, sí es posible obtener las distribuciones marginales posterior mediante un argumento similar al de la sección anterior, notando que

$$p(\theta \mid \sigma^2, \mathbf{Y}) \propto p(\theta, \underbrace{\sigma^2}_{fijo} \mid \mathbf{Y}) \qquad \text{y} \qquad p(\sigma^2 \mid \theta, \mathbf{Y}) \propto p(\underbrace{\theta}_{fijo}, \sigma^2 \mid \mathbf{Y})$$

Es decir, para encontrar la distribución posterior marginal de θ dado σ^2, se utiliza la distribución posterior conjunta y los términos que no dependan de θ se incorporan en la constante de proporcionalidad. El mismo razonamiento se hace para el parámetro σ^2.

Resultado 4.1.2. *La distribución posterior condicional de θ es*

$$\theta \mid \sigma^2, \mathbf{Y} \sim Normal(\mu_n, \tau_n^2)$$

En donde las expresiones para μ_n y τ_n^2 están dadas por (2.6.4) XXXXXXX.

Prueba. Acudiendo a la distribución posterior conjunta e incorporando los términos que no dependen de θ en la constante de proporcionalidad, se tiene que

$$p(\theta \mid \sigma^2, \mathbf{Y}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right\}$$

Completando los cuadrados y siguiendo el razonamiento de la demostración del Resultado 2.6.4, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_n, \tau_n^2)$. ■

Resultado 4.1.3. *La distribución posterior condicional de σ^2 es*

$$\sigma^2 \mid \theta, \mathbf{Y} \sim Gamma - Inversa\left(\frac{n_0+n}{2}, \frac{v_0}{2}\right)$$

En donde $v_0 = n_0\sigma_0^2 + (n-1)S^2 + n(\bar{y}-\theta)^2$

Prueba. Acudiendo a la distribución posterior conjunta e incorporando los términos que no dependen de θ en la constante de proporcionalidad, se tiene que

$$p(\sigma^2 \mid \theta, \mathbf{Y}) \propto (\sigma^2)^{-(n+n_0)/2-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma - Inversa(\frac{n_0+n}{2}, \frac{v_0}{2})$. ■

La estimación de los parámetros se realiza acudiendo a la teoría de los procesos estocásticos mediante métodos de Monte Carlo, la cual consiste en BLA BLA

Ejemplo 4.1.1. ? *plantea el siguiente conjunto de datos que muestran la función renal de 15 individuos que se sometieron a una prueba médica exhaustiva en un hospital. Los resultados de la prueba renal están en un intervalo de -6 puntos a 4 puntos. Entre más alto sea el resultado, se concluye que el riñón del individuo es más sano. Nótese que estas pruebas son importantes para predecir el comportamiento de un riñón donado a un paciente con problemas renales. Los datos son extraídos de la siguiente página WEB (*`http://www.stat.stanford.edu/~omkar/monograph/data`*) y para este ejemplo sólo se utilizaron los primeros 15 datos del archivo.*

En principio, es de interés para el investigador conocer la media y la dispersión de estos datos, para poder analizar a fondo la situación de los pacientes que esperan un transplante.

Dado que se trata de una primera aproximación, se prefiere utilizar distribuciones previas no informativas para los parámetros de la media y varianza. Lo anterior

se logra en WinBugs definiendo las distribuciones previas de `mu ~ dnorm(0,0.001)` *y de* `prec ~ dgamma(0.001,0.001)`. *El lector debe recordar que* `prec` *corresponde al parámetro de precisión que resulta ser el inverso de la varianza. De esta forma, la distribución previa de mu está centrada en cero, pero con una varianza muy grande al igual que la distribución de la varianza.*

El siguiente código en WinBugs muestra cómo se lleva a cabo la inferencia.

Código WinBugs

```
model {
for (i in 1:n)
{
y[i] ~ dnorm(theta,prec);
}
theta ~ dnorm(0,0.001);
prec ~ dgamma(0.001,0.001);
sigma2<-1/sqrt(prec)
}

Data
list(y=c(1.69045085, -1.41076082, -0.27909483, -0.91387987,
3.21868429, -1.47282460, -0.96524353, -2.45084934, 1.03838153,
1.79928679, 0.97826621, 0.67463830, -1.08665864,
-0.00509027, 0.43708128), n=15)

Inits
list(theta=0, prec= 1)
```

Después de cinco mil iteraciones, la salida del anterior código muestra una estimación puntual para la esperanza de Y de -0.02382 con un intervalo de credibilidad del 95 % dado por (-0.4665, 0.4318). Por otro lado, la estimación puntual de la varianza de Y es de 2.667 con un intervalo de credibilidad del 95 % dado por (1.237, 5.67).

Nótese que el acondicionamiento sucesivo lleva a que la inferencia Bayesiana deba utilizar métodos de Monte Carlo para generar las cadenas de las distribuciones posteriores. En particular, el siguiente código de R muestra un algoritmo de Gibbs para los datos del ejemplo. Se recalca que se utiliza la librería `MCMCpack` *(**?**) para generar las realizaciones de la distribución inversa gamma, cuya distribución no informativa se muestra en la figura **??**.*

Código R

```
y=c(1.69045085, -1.41076082, -0.27909483, -0.91387987, 3.21868429,
-1.47282460, -0.96524353, -2.45084934, 1.03838153, 1.79928679,
0.97826621, 0.67463830, -1.08665864, -0.00509027, 0.43708128)

n<-length(y)
```

```
#Parámetros previos de theta
mu<-0
tau2<-1000
#Parámetros previos de sigma2
a<-0.001
b<-0.001

nsim<-1000
theta.pos<-rep(NA,nsim)
sigma2.pos<-rep(NA,nsim)

theta.pos[1]<-0
sigma2.pos[1]<-1

########################
# Muestreador de Gibbs #
########################

for(i in 1:nsim){
#Parámetros posteriores de theta
mu.n<-(mean(y)*(n/sigma2.pos[i])+(theta.pos[i]/n))/((n/sigma2.pos[i])+(1/tau2))
tau2.n<-1/((n/sigma2.pos[i])+(1/tau2))
#simulación de la distribución posterior condicional de theta
theta.pos[i+1]<-rnorm(1,mean=mu.n, sd=sqrt(tau2.n))
#Parámetros posteriores de sigma2
a.n<-a+n/2
b.n<-b+((n-1)*var(y)+n*(mean(y)-theta.pos[i]))/2
#simulación de la distribución posterior condicional de theta
sigma2.pos[i+1]<-rinvgamma(1, a.n, b.n)
}
```

4.1.2 Parámetros dependientes

En algunas situaciones es muy útil asumir una distribución previa conjugada en vez de una distribución previa no informativa. En este escenario, no es posible establecer que los parámetros tengan distribuciones previa independientes. Bajo esta situación, la inferencia posterior de los parámetros de interés debe ser llevada a cabo en dos etapas: En la primera, se debe establecer la distribución previa conjunta para ambos parámetros siguiendo la sencilla regla que afirma que

$$p(\theta, \sigma^2) = p(\sigma^2)p(\theta \mid \sigma^2)$$

En la segunda etapa ya es posible analizar posterior propiamente cada uno de los parámetros de interés siguiendo otra sencilla regla que afirma que

$$p(\theta, \sigma^2 \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \theta, \sigma^2)p(\theta, \sigma^2)$$

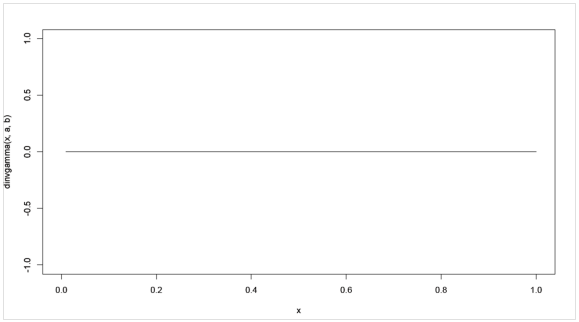

Figura 4.1: *Distribución previa no informativa para la varianza de los datos de funcionamiento del riñón.*

La anterior formulación conlleva a asignar una distribución previa para θ dependiente del parámetro σ^2. Esto quiere decir que en la distribución $p(\theta \mid \sigma^2)$ el valor de σ^2 se considera una constante fija y conocida. Siguiendo los argumentos de la Sección 2.6, es plausible que la distribución previa de θ sea

$$p(\theta \mid \sigma^2) \sim Normal(\mu, \sigma^2/c_0)$$

Donde c_0 es una constante. Por otro lado, y siguiendo los argumentos de la sección 2.7, una posible opción para la distribución previa de σ^2, que no depende de θ, corresponde a

$$p(\sigma^2) \sim Gamma - inversa(n_0/2, n_0\sigma_0^2/2)$$

Resultado 4.1.4. *La distribución conjunta previa de los parámetros θ y σ^2 está dada por una distribución*

$$\theta, \sigma^2 \sim Normal - Gamma(\mu, c_0, n_0, \sigma_0^2).$$

Prueba.

$$\begin{aligned} p(\theta, \sigma^2) &= p(\sigma^2)p(\theta \mid \sigma^2) \\ &\propto (\sigma^2)^{-n_0/2-1} \exp\left\{-n_0\sigma_0^2/(2\sigma^2)\right\} (\sigma^2)^{-1/2} \exp\left\{-\frac{c_0}{2\sigma^2}(\theta-\mu)^2\right\} \\ &= (\sigma^2)^{-n_0/2-1/2-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + c_0(\theta-\mu)^2\right]\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal - Gamma(\mu, c_0, n_0, \sigma_0^2)$. ■

Resultado 4.1.5. *La distribución posterior conjunta de los parámetros θ y σ^2 está dada por*

$$p(\theta, \sigma^2 \mid \mathbf{Y}) \propto (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \times \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + (c_0+n)(\theta-\mu_n)^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2\right]\right\} \tag{4.1.3}$$

Donde

$$\mu_n = \frac{\frac{n}{\sigma^2}\bar{Y} + \frac{c_0}{\sigma^2}\mu}{\frac{n}{\sigma^2} + \frac{c_0}{\sigma^2}} = \frac{n\bar{Y} + c_0\mu}{n + c_0} \tag{4.1.4}$$

Prueba. En primer lugar, siguiendo la expresión (3.1.4), la verosimilitud de la muestra está dada por

$$p(\mathbf{Y} \mid \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\} \tag{4.1.5}$$

Por otro lado, se tiene que

$$\begin{aligned} p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta, \sigma^2) p(\theta, \sigma^2) \\ &\propto (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + c_0(\theta-\mu)^2 + (n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\} \\ &= (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + (c_0+n)(\theta-\mu_n)^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2\right]\right\} \end{aligned}$$

puesto que

$$c_0(\theta-\mu)^2 + n(\bar{y}-\theta)^2 = (c_0+n)(\theta-\mu_n)^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2$$

■

Para encontrar las distribuciones marginales posterior de cada uno de los parámetros de interés se utilizan argumentos distintos puesto que θ depende de σ^2 pero este último no depende del primero. De esta manera se tiene que:

1. Para hallar la distribución posterior condicional de $\theta \mid \sigma^2$, dada por $P(\theta \mid \sigma^2, \mathbf{Y})$, se debe considerar que σ^2 es una constante fija y conocida tal como se consideró al principio de esta sección. Basado en lo anterior, es posible utilizar la siguiente regla de probabilidad

$$P(\theta \mid \sigma^2, \mathbf{Y}) = \frac{p(\theta, \sigma^2 \mid \mathbf{Y})}{p(\sigma^2, \mathbf{Y})} p(\mathbf{Y}) \propto p(\theta, \sigma^2 \mid \mathbf{Y})$$

Lo anterior sugiere que la distribución marginal posterior de θ, $p(\theta \mid \sigma^2, \mathbf{Y})$, se encuentra utilizando la distribución posterior conjunta, $p(\theta, \sigma^2 \mid \mathbf{Y})$, suponiendo que todas las expresiones que involucren al valor σ^2 se pueden incluir en la constante de proporcionalidad

2. Dado que σ^2 no depende de ningún otro parámetro entonces, utilizando la distribución posterior conjunta, es posible encontrar su distribución marginal posterior de la siguiente forma

$$p(\sigma^2 \mid \mathbf{Y}) = \int p(\theta, \sigma^2 \mid \mathbf{Y}) d\theta$$

Lo propio es posible hacer con θ, utilizando la distribución posterior conjunta, es posible encontrar su distribución marginal posterior de la siguiente forma

$$p(\theta \mid \mathbf{Y}) = \int p(\theta, \sigma^2 \mid \mathbf{Y}) d\sigma^2$$

Resultado 4.1.6. *La distribuciónposterior de θ condicional a sigma$^2, \sigma^2, \mathbf{Y}$ está dada por*

$$\theta \mid \sigma^2, \mathbf{Y} \sim Normal(\mu_n, \sigma^2/(n + c_0))$$

con $\mu_n = \frac{n\bar{y}+c_0\mu}{n+c_0}$.

Prueba. Acudiendo a la distribución posterior conjunta e incorporando los términos que no dependen de θ en la constante de proporcionalidad, se tiene que

$$\begin{aligned} p(\theta \mid \sigma^2, \mathbf{Y}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[c_0(\theta-\mu)^2 + n(\bar{y}-\theta)^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(n+c_0)\theta^2 - 2(n\bar{y}+c_0\mu)\theta\right]\right\} \\ &= \exp\left\{-\frac{n+c_0}{2\sigma^2}\left[\theta^2 - 2\frac{n\bar{y}+c_0\mu}{n+c_0}\theta\right]\right\} \\ &\propto \exp\left\{-\frac{n+c_0}{2\sigma^2}\left[\theta - \frac{n\bar{y}+c_0\mu}{n+c_0}\right]^2\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_n, \sigma^2/(n+c_0))$. ■

En el anterior resultado, la media de la distribución condicional posterior μ_n se puede escribir como $\mu_n = \frac{n}{n+c_0}\bar{y} + \frac{c_0}{n+c_0}\mu$, promedio ponderado entre la estimación clásica $\bar{y}$ y la estimación previa μ. Observando la forma que toman los pesos $\frac{n}{n+c_0}$ y $\frac{c_0}{n+c_0}$, se puede pensar a c_0 como el número de observaciones en la información previa, y así, los pesos de la estimación clásica y la estimación previa dependen directamente de los tamaños muestrales respectivos.

Resultado 4.1.7. *La distribución marginal posterior del parámetro* σ^2 *es*

$$\sigma^2 \mid \mathbf{Y} \sim Gamma - inversa\left(\frac{n+n_0}{2}, \frac{(n+n_0)\sigma_n^2}{2}\right)$$

Donde $(n+n_0)\sigma_n^2 = n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2$ *corresponde a una suma ponderada de la varianza previa, la varianza muestral y la distancia entre la media muestral y la media previa.*

Prueba. De la distribución posterior conjunta (3.1.5) e integrando con respecto a θ, se tiene que

$$\begin{aligned}
p(\sigma^2 \mid \mathbf{Y}) &= \int p(\theta, \sigma^2 \mid \mathbf{Y})\, d\theta \\
&\propto (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2\right]\right\} \\
&\qquad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{n+c_0}{2\sigma^2}(\theta-\mu_n)^2\right\}\, d\theta \\
&\propto (\sigma^2)^{-\frac{n_0+n}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2\right]\right\} \\
&\qquad \times \int_{-\infty}^{\infty} \frac{\sqrt{n+c_0}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{n+c_0}{2\sigma^2}(\theta-\mu_n)^2\right\}\, d\theta \\
&\propto (\sigma^2)^{-\frac{n_0+n}{2}-1} \exp\left\{-\frac{(n+n_0)\sigma_n^2}{2\sigma^2}\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma - inversa(\frac{n+n_0}{2}, \frac{(n+n_0)\sigma_n^2}{2})$. ■

Resultado 4.1.8. *La distribución marginal posterior del parámetro* θ *es*

$$\theta \mid \mathbf{Y} \sim t_{n+n_0}\left(\mu_n, \frac{\sigma_n^2}{c_0+n}\right)$$

Prueba. Esta demostración sigue los lineamentos de la demostración del Resultado 3.1.1 y partiendo de la distribución posterior conjunta e integrando con respecto a σ^2, se tiene que

$$\begin{aligned}
p(\theta \mid \mathbf{Y}) &= \int_0^{\infty} p(\theta, \sigma^2 \mid \mathbf{Y})\, d\sigma^2 \\
&\propto \int_0^{\infty} \left(\frac{1}{\sigma^2}\right)^{\frac{n_0+n+1}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\left[(n_0+n)\sigma_n^2 + (c_0+n)(\theta-\mu_n)^2\right]\right\}\, d\sigma^2
\end{aligned}$$

Haciendo un cambio de variable tal que

$$z = \frac{A}{2\sigma^2}, \qquad \text{donde} \quad A = (n_0+n)\sigma_n^2 + (c_0+n)(\theta-\mu_n)^2$$

por tanto

$$d\sigma^2 = -\frac{A}{2z^2}\, dz$$

Entonces, volviendo a la integral en cuestión, se tiene que

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto \left(\frac{1}{A}\right)^{\frac{n_0+n+1}{2}+1} \int_{\infty}^{0} \frac{-A}{2z^2}(2z)^{\frac{n_0+n+1}{2}+1} e^{-z}\, dz \\ &\propto A^{-\frac{n_0+n+1}{2}} \underbrace{\int_0^{\infty} z^{\frac{n_0+n+1}{2}-1} e^{-z}\, dz}_{Gamma\left(\frac{n_0+n+1}{2},1\right)} \\ &\propto A^{-\frac{n_0+n+1}{2}} = \left[(n_0+n)\sigma_n^2 + (c_0+n)(\theta-\mu_n)^2\right]^{-\frac{n_0+n+1}{2}} \\ &\propto \left[1 + \frac{(c_0+n)(\theta-\mu_n)^2}{(n_0+n)\sigma_n^2}\right]^{-\frac{n_0+n+1}{2}} \\ &= \left[1 + \frac{1}{n_0+n}\left(\frac{\theta-\mu_n}{\sigma_n/\sqrt{c_0+n}}\right)^2\right]^{-\frac{n_0+n+1}{2}} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $t_{n-1}(\bar{y}, \frac{\sigma_n^2}{c_0+n})$. ■

En términos de simulación de densidades se debe primero simular distribución $p(\sigma^2 \mid \mathbf{Y})$ y encontrar un valor estimado para este parámetro. Luego, se debe utilizar este valor para simular la distribución $p(\theta \mid \sigma^2, \mathbf{Y})$ y encontrar un valor estimado para este parámetro.

Ejemplo 4.1.2. *Para los datos de función renal (**?**) que se muestran en el Ejemplo 4.1.1, suponga que la información previa está contenida en la medición de función renal en una muestra de 12 pacientes dadas por: -1.3619, -1.1116, -0.4744, -0.5663, 2.2056, 0.9491, 0.2298, -0.7933, 1.0198, -0.9850, 3.5679 y -1.9504. La media y la varianza muestral de estas 12 observaciones corresponden a 0.060775 y 2.598512, así, se toma* $\mu = 0.060775$, $\sigma_0^2 = 2.598512$ *y* $c_0 = n_0 = 12$. *Por otro lado, la media y la varianza muestral de las 15 pacientes en la información actual son* $\bar{y} = 0.08349249$ *y* $S^2 = 2.301684$. *De esta forma, los parámetros de las distribuciones marginales posterior de* θ *y* σ^2 *se pueden calcular como* $\mu_n = \frac{15}{15+12}0.08349249 + \frac{12}{15+12}0.060775 = 0.07339583$ *y* $\sigma_n^2 = \frac{12*2.598512+14*2.301684+6.666667*(0.060775-0.08349249)^2}{15+12} = 2.348487$. *En conclusión, las distribuciones marginales posterior de* θ *y* σ^2 *están dadas por*

$$\theta|\mathbf{Y} \sim t_{27}(0.07339583, 2.348487/27 = 0.086981)$$

y

$$\sigma|\mathbf{Y} \sim IG(27/2 = 13.5, 27 * 2.348487/2 = 31.70457)$$

Así, la estimación Bayesiana de θ *es* $\mu_n = 0.07339583$ *y un intervalo de credibilidad de 95 % para* θ *se puede calcular como* $0.07339583 \pm t_{27,0.975} * \sqrt{0.086981} =$

$(-0.5317, 0.6785)$. *Por otro lado, la estimación Bayesiana de* σ^2 *está dada por* $31.70457/(13.5-1) = 2.536366$, *y un intervalo de credibilidad de* $95\,\%$ *para* σ^2 *se puede calcular como los percentiles* $2.5\,\%$ *y* $97.5\,\%$ *de la distribución* $IG(13.5, 31.70457)$, *dado por* $(1.467991, 4.351024)$.

Otra forma de estimar los parámetros θ *y* σ^2 *es utilizando las distribuciones* $p(\sigma^2|\mathbf{Y})$ *y* $p(\theta|\sigma^2, \mathbf{Y})$, *simulando valores para* σ^2, *y posteriormente reemplazarlos en* $p(\theta|\sigma^2, \mathbf{Y})$ *para simular valores de* θ. *Resultados obtenidos así para* θ *y* σ^2 *con 1000 iteraciones se muestran en la figura (**??**).*

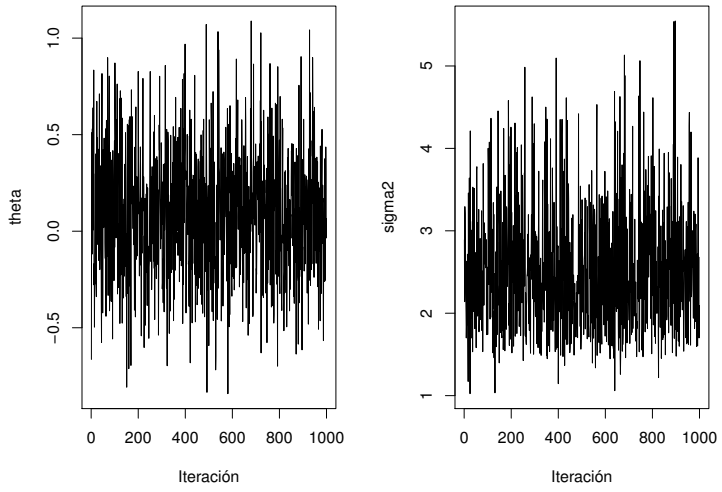

Figura 4.2: *Valores simulados para* θ *y* σ^2 *del Ejemplo 4.1.2*

La estimación e intervalo de credibilidad para θ *calculada desde los valores simulados corresponden a 0.07277454 y (-0.5292, 0.7863); mientras que para* σ^2, *están dadas por 2.527097 y (1.48856, 4.323065). Podemos ver que los resultados obtenidos en los dos enfoques son muy similares.*

4.1.3 Parámetros no informativos

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ corresponde a una muestra de variables aleatorias con distribución $Normal(\theta, \sigma^2)$. Luego, la función de distribución conjunta o verosimilitud está dada por (2.6.3)

$$p(\mathbf{Y} \mid \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} I_{\mathbb{R}^n}(\mathbf{y}) \tag{4.1.6}$$

En primer lugar suponga que los parámetros tienen distribuciones previa inde-

pendientes y en esta primera etapa se realizará el análisis suponiendo que estas distribuciones son no informativas[1]. Lo anterior implica que la distribución previa conjunta de los parámetros de interés está dada por

$$p(\theta, \sigma^2) = p(\theta)p(\sigma^2) \tag{4.1.7}$$

Como la distribución previa de θ es normal, es fácil verificar que ésta empieza a tener las características propias de una distribución no informativa cuando la varianza de la misma se vuelve muy grande, sin importar el valor de la media. Cuando esto sucede, entonces la forma de la distribución previa de θ se torna plana y es lógico pensar que puede ser acercada mediante una distribución constante, tal que

$$p(\theta) \propto cte$$

Por otro lado, (**?**) afirma que la distribución Gamma-inversa, la cual es la distribución previa para el parámetro σ^2, se vuelve no informativa cuando los hiperparámetros toman valores muy cercanos a cero. De esta forma haciendo tender $\alpha \longrightarrow 0$ y $\beta \longrightarrow 0$, entonces la expresión (2.7.2) se convierte en

$$p(\sigma^2) \propto 1/\sigma^2$$

Por lo anterior, la distribución previa no informativa conjunta estaría dada por

$$p(\theta, \sigma^2) \propto 1/\sigma^2 \tag{4.1.8}$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 4.1.9. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid \mathbf{Y} \sim t_{n-1}\left(\bar{y}, \frac{S^2}{n}\right).$$

Donde $(n-1)S^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Esta distribución también puede expresarse como

$$\frac{\theta - \bar{y}}{S/\sqrt{n}} \mid \mathbf{Y} \sim t_{n-1}$$

Prueba. En primer lugar nótese que la distribución posterior conjunta de los

[1] Sin tener en cuenta que puedan resultar impropias

parámetros de interés es

$$\begin{aligned} p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto p(\theta, \sigma^2) p(\mathbf{Y} \mid \theta, \sigma^2) \\ &\propto \frac{1}{\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\} \\ &= \frac{1}{\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \theta)^2 \right] \right\} \\ &= \frac{1}{\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[(n-1)S^2 + n(\bar{y} - \theta)^2 \right] \right\} \end{aligned} \tag{4.1.9}$$

Ahora, para hallar la distribución marginal posterior de θ es necesario integrar la anterior expresión con respecto a σ^2. Con esto, se tiene que

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &= \int_0^\infty p(\theta, \sigma^2 \mid \mathbf{Y}) \, d\sigma^2 \\ &\propto \int_0^\infty \left(\frac{1}{\sigma^2} \right)^{n/2+1} \exp\left\{ -\frac{1}{2\sigma^2} \left[(n-1)S^2 + n(\bar{y} - \theta)^2 \right] \right\} \, d\sigma^2 \end{aligned}$$

Haciendo un cambio de variable tal que

$$z = \frac{A}{2\sigma^2}, \qquad \text{donde} \quad A = (n-1)S^2 + n(\bar{y} - \theta)^2$$

por tanto

$$d\sigma^2 = -\frac{A}{2z^2} \, dz$$

Entonces, volviendo a la integral en cuestión, se tiene que

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto \left(\frac{1}{A} \right)^{n/2+1} \int_\infty^0 \frac{-A}{2z^2} (2z)^{n/2+1} e^{-z} \, dz \\ &\propto A^{-n/2} \underbrace{\int_0^\infty z^{n/2-1} e^{-z} \, dz}_{Gamma(n/2)} \\ &\propto A^{-n/2} = [(n-1)S^2 + n(\bar{y} - \theta)^2]^{-n/2} \\ &\propto \left[1 + \frac{n(\bar{y} - \theta)^2}{(n-1)S^2} \right]^{-n/2} = \left[1 + \frac{1}{n-1} \left(\frac{\bar{y} - \theta}{S/\sqrt{n}} \right)^2 \right]^{-\frac{(n-1)+1}{2}} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $t_{n-1}(\bar{y}, S^2/n)$. ■

Resultado 4.1.10. *La distribución posterior del parámetro σ^2 sigue una distribución*

$$\sigma^2 \mid \mathbf{Y} \sim Gamma - inversa((n-1)/2, (n-1)S^2/2).$$

Prueba. Utilizando el mismo argumento de la demostración anterior y recurriendo a la expresión (3.2.4), se tiene que

$$\begin{aligned}
p(\sigma^2 \mid \mathbf{Y}) &= \int_{-\infty}^{\infty} p(\theta, \sigma^2 \mid \mathbf{Y})\, d\theta \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\} d\theta \\
&= \frac{\sqrt{2\pi\sigma^2/n}}{\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\} \\
&\qquad \times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\} d\theta \\
&\propto (\sigma^2)^{-n/2-1/2} \exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\} \\
&= (\sigma^2)^{-\frac{n-1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma - inversa((n-1)/2, (n-1)S^2/2)$. ■

Análisis condicional

Nótese que si en un principio las distribuciones previa hubiesen sido tales que para el parámetro θ su distribución previa dependiese del otro parámetro σ y si este último no dependiese de ningún otro, entonces el análisis posterior debe cambiar significativamente. Con base en lo anterior, manteniendo la distribución previa conjunta dada por la expresión (3.1.3) y la la distribución posterior conjunta dada por la expresión (3.1.4), entonces la distribución marginal posterior de σ se mantendría de acuerdo al Resultado 3.1.2, pero la distribución posterior de θ condicional a $\sigma, \mathbf{Y}$ cambiaría de la forma en que lo afirma el siguiente resultado.

Resultado 4.1.11. *La distribución posterior de θ condicional a $\sigma, \mathbf{Y}$ es*

$$\theta \mid \sigma, \mathbf{Y} \sim Normal(\bar{y}, \sigma^2/n)$$

Prueba. De la expresión (3.1.4), suponiendo que σ^2 es fijo e incorporando los términos que no dependen de θ a la constante de proporcionalidad, se tiene que

$$\begin{aligned}
p(\theta \mid \sigma^2, \mathbf{Y}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y}-\theta)^2\right]\right\} \\
&= \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\bar{y}, \sigma^2/n)$. ■

En términos de simulación de densidades se debe primero simular distribución $p(\sigma^2 \mid \mathbf{Y})$ y encontrar un valor estimado para este parámetro. Luego, se debe utilizar este valor para simular la distribución $p(\theta \mid \sigma^2, \mathbf{Y})$ y encontrar un valor estimado para este parámetro.

4.2 Normal multivariante con media desconocida y varianza conocida

Cuando la verosimilitud de los datos es una distribución normal multivariante, las técnicas de inferencia no se distancian mucho del caso univariado. Se debe tener en cuenta el manejo matricial de las formas cuadráticas y las propiedades básicas del cálculo de matrices. Los desarrollos y resultados derivados de esta sección redundarán en el análisis de los modelos lineales con el enfoque bayesiano.

Sea $\mathbf{Y} = (Y_1, \ldots, Y_p)$ un vector aleatorio cuya distribución es normal multivariante dada por

$$p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\theta})\right\} \tag{4.2.1}$$

en donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ es la media de cada uno de los componentes del vector y $\boldsymbol{\Sigma}$ es una matriz de varianzas y covarianzas de orden $p \times p$, simétrica y definida positiva. La verosimilitud para una muestra de n vectores aleatorios independientes e idénticamente distribuidos está dada por

$$p(\mathbf{Y}_1 \ldots, \mathbf{Y}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\theta})\right\} \tag{4.2.2}$$

En el caso más general, suponga que el vector de medias θ sigue una distribución previa normal multivariante informativa y parametrizada por los hiper parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Gamma}$

$$p(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}) \propto \mid \boldsymbol{\Gamma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right\}$$

Nótese que esta distribución se hace no informativa cuando $\mid \boldsymbol{\Gamma}^{-1} \mid \longrightarrow 0$ sin importar el valor del vector de medias previa $\boldsymbol{\mu}$.

Resultado 4.2.1. *La distribución posterior del vector de parámetros $\boldsymbol{\mu}$ sigue una distribución normal multivariante*

$$\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}_n, \boldsymbol{\Gamma}_n).$$

En donde

$$\boldsymbol{\mu}_n = \left(\boldsymbol{\Gamma}^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu} + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}\right) \tag{4.2.3}$$

$$\boldsymbol{\Gamma}_n = \left(\boldsymbol{\Gamma}^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{4.2.4}$$

Prueba. En primer lugar, nótese la siguiente identidad

$$\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta})=\sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}})+n(\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta}) \tag{4.2.5}$$

que resulta puesto que

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta}) &= \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}}+\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}}+\bar{\mathbf{Y}}-\boldsymbol{\theta})\\ &= \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}})\\ &\quad+\sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta})\\ &\quad+(\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\\ &\quad+\sum_{i=1}^{n}(\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta})\\ &= \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}})+n(\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta})\end{aligned}$$

Por otro lado, de la definición de distribución previa, se tiene que

$$\begin{aligned}p(\boldsymbol{\theta}\mid\mathbf{Y},\boldsymbol{\Sigma}) &\propto p(\mathbf{Y}\mid\boldsymbol{\theta},\boldsymbol{\Sigma})p(\boldsymbol{\theta},\boldsymbol{\Sigma})\\ &= \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\theta})+(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}\left[n(\bar{\mathbf{y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}}-\boldsymbol{\theta})+(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}\left[-n\bar{\mathbf{y}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}-n\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}+n\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}-\boldsymbol{\theta}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}\right]\right\}\\ &= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'(\boldsymbol{\Gamma}^{-1}+n\boldsymbol{\Sigma}^{-1})\boldsymbol{\theta}-\boldsymbol{\theta}'(2\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}+2n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}})\right]\right\}\\ &= \exp\left\{-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n\right\}\\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n-\frac{1}{2}\boldsymbol{\mu}_n\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n\right\}\\ &= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\theta}-2\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n+\boldsymbol{\mu}_n\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n\right]\right\}\\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)'\boldsymbol{\Gamma}_n^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)\right\}\end{aligned}$$

■

Resultado 4.2.2. *La distribución posterior marginal de un subconjunto de parámetros, digamos $\boldsymbol{\theta}^{(1)}$ es también normal multivariante con media igual a la del subvector de medias apropiado, $\boldsymbol{\mu}_n^{(1)}$ y similar matriz de varianzas.*

Prueba. Este resultado se sigue inmediatamente de las propiedades de la distribución normal multivariante ■

Resultado 4.2.3. *La distribución posterior condicional de un subconjunto de parámetros, digamos $\boldsymbol{\theta}^{(1)}$, dado $\boldsymbol{\theta}^{(2)}$ es también normal multivariante dada por*

$$\boldsymbol{\theta}^{(1)} \mid \boldsymbol{\theta}^{(2)} \sim N_p\left(\boldsymbol{\mu}_n^{(1)} + \boldsymbol{\Gamma}_n^{(12)}\left(\boldsymbol{\Gamma}_n^{(22)}\right)^{-1}\left(\theta^{(2)} - \mu_n^{(2)}\right), \boldsymbol{\Gamma}_n^{(1|2)}\right).$$

En donde

$$\boldsymbol{\Gamma}_n^{(1|2)} = \boldsymbol{\Gamma}_n^{(11)} - \boldsymbol{\Gamma}_n^{(12)}\left(\boldsymbol{\Gamma}_n^{(22)}\right)^{-1}\boldsymbol{\Gamma}_n^{(21)} \tag{4.2.6}$$

Prueba. Este resultado se sigue inmediatamente de las propiedades de la distribución normal multivariante ■

4.3 Normal multivariante con media y varianza desconocida

Al igual que en las secciones anteriores, cuando se desconoce tanto la media como la varianza de la distribución, es necesario plantear diversos enfoques y situarse en el más conveniente.[2] Suponiendo que el número de observaciones en la muestra aleatoria sea suficiente, existe otra situación que se debe surtir y es la asignación de las distribuciones previa para $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$. En estos términos, es posible

- Suponer que la distribución previa $p(\boldsymbol{\theta})$ es independiente de la distribución previa $p(\boldsymbol{\Sigma})$ y que ambas distribuciones son no informativas.
- Suponer que la distribución previa para $\boldsymbol{\theta}$ depende de $\boldsymbol{\Sigma}$ y escribirla como $p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$, mientras que la distribución previa de $\boldsymbol{\Sigma}$ no depende de $\boldsymbol{\theta}$ y se puede escribir como $p(\boldsymbol{\Sigma})$. El análisis posterior de este enfoque encuentra la distribución posterior de $\boldsymbol{\Sigma} \mid \mathbf{Y}$ y con esta se encuentra la distribución posterior de $\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}$.
- Suponer que la distribución previa $p(\boldsymbol{\theta})$ es independiente de la distribución previa $p(\boldsymbol{\Sigma})$ y que ambas distribuciones son informativas. Luego, utilizar un análisis de simulación condicional conjunta para extraer muestras provenientes de las respectivas distribuciones posterior.

[2] Nótese que en términos de parámetros, existen p parámetros correspondientes al vector de medias $\boldsymbol{\theta}$ y $\binom{p}{2} = \dfrac{p(p+1)}{2}$ parámetros correspondientes a la matriz de varianzas $\boldsymbol{\Sigma}$. Pensando en la gran cantidad de parámetros que se deben modelar, es necesario tener en cuenta que el número de datos en la muestra aleatoria sea lo suficientemente grande.

4.3.1 Parámetros independientes

En este último enfoque se supone que las distribuciones previa para los parámetros de interés son independientes e informativas. Para lograr que las resultantes distribuciones posterior sean conjugadas, debe escribirse la verosimilitud de la muestra aleatoria $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ justo como los presenta la expresión (B.3.5) tal que

$$(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}})\right\} \tag{4.3.1}$$

Donde $\mathbf{S}_{\boldsymbol{\theta}} = \sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})'$. Para el vector de medias $\boldsymbol{\theta}$ es posible asignar una distribución previa normal tal que

$$\boldsymbol{\theta} \sim Normal_p(\boldsymbol{\mu}, \boldsymbol{\Gamma})$$

Por otro lado, la distribución para la matriz de varianzas $\boldsymbol{\Sigma}$ es

$$\boldsymbol{\Sigma} \sim Wishart - inversa_v(\boldsymbol{\Lambda}^{-1})$$

Asumiendo independencia previa, la distribución previa conjunta resulta estar dada por

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) &= p(\boldsymbol{\theta})p(\boldsymbol{\Sigma}) \\ &\propto \mid \boldsymbol{\Sigma} \mid^{-(v+p+1)/2} \\ &\times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\} \end{aligned} \tag{4.3.2}$$

Una vez que se conoce la forma estructural de la distribución previa conjunta, es posible establecer la distribución posterior conjunta puesto que la verosimilitud de los datos, $p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})$, está dada por la expresión (3.3.1), entonces y acudiendo a la simetría de las matrices $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}$ y $\mathbf{S}_{\boldsymbol{\theta}}$, se tiene que

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) &= p(\boldsymbol{\theta}, \boldsymbol{\Sigma})p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \\ &\propto \mid \boldsymbol{\Sigma} \mid^{-(v+n+p+1)/2} \\ &\times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\} \\ &\propto \mid \boldsymbol{\Sigma} \mid^{-(v+n+p+1)/2} \\ &\times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\} \end{aligned} \tag{4.3.3}$$

Dado que la distribución posterior conjunta no tiene una forma estructural conocida, no es posible utilizar el método de integración analítica. Sin embargo, es posible obtener las distribuciones condicionales de cada uno de los parámetros suponiendo fijos los restantes y teniendo en cuenta que

$$p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}) \propto p(\boldsymbol{\theta}, \underbrace{\boldsymbol{\Sigma}}_{fijo} \mid \mathbf{Y}) \qquad \text{y} \qquad p(\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \mathbf{Y}) \propto p(\underbrace{\boldsymbol{\theta}}_{fijo}, \boldsymbol{\Sigma} \mid \mathbf{Y})$$

Resultado 4.3.1. *La distribución posterior de la matriz de parámetros* $\boldsymbol{\Sigma}$ *condicional a* $\boldsymbol{\theta}, \mathbf{Y}$ *es*

$$\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \mathbf{Y} \sim inversa - Wishart_{v+n}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})$$

Prueba. La prueba es inmediata notando que

$$\begin{aligned}\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \mathbf{Y} \propto & \mid \boldsymbol{\Sigma} \mid^{-(v+n+p+1)/2} \\ & \times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\}\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $inversa - Wishart_{v+n}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})$. ■

Resultado 4.3.2. *La distribución posterior del vector de parámetros* $\boldsymbol{\theta}$ *condicional a* $\boldsymbol{\Sigma}, \mathbf{Y}$ *es*

$$\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y} \sim Normal_p(\boldsymbol{\mu}_n, \boldsymbol{\Gamma}_n)$$

donde $\boldsymbol{\mu}_n$ *y* $\boldsymbol{\Gamma}_n$ *están dadas por las expresiones (3.2.3) y (3.2.4), respectivamente.*

Prueba. Utilizando el resultado (B.3.5), se tiene que la verosimilitud para los datos observados también se puede escribir como

$$p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})\right\}$$

Por lo tanto, reemplazando la verosimilitud en (3.3.3), se tiene que la distribución posterior conjunta también se puede escribir como

$$\begin{aligned}p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) \propto & \mid \boldsymbol{\Sigma} \mid^{-(v+n+p+1)/2} \\ & \times \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\}\end{aligned}$$

Y fijando la matriz de parámetros $\boldsymbol{\Sigma}$, se encuentra que la distribución posterior de $\boldsymbol{\theta}$ condicional a $\boldsymbol{\Sigma}, \mathbf{Y}$ es tal que

$$p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}) \propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\}$$

La anterior expresión tiene la misma forma estructural que la expresión principal en la demostración del resultado 3.2.1. Luego, siguiendo el mismo razonamiento se llega fácilmente a la prueba. ■

4.3.2 Parámetros dependientes

Al igual que en el caso univariado, la inferencia posterior de los parámetros de interés debe ser llevada a cabo en dos etapas: En la primera, se debe establecer la distribución previa conjunta para ambos parámetros mediante

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = p(\boldsymbol{\Sigma})p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$$

Luego, en la segunda etapa es posible analizar posterior propiamente cada uno de los parámetros de interés puesto que

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

Al igual que en el caso univariado, la anterior formulación conlleva a asignar una distribución previa para $\boldsymbol{\theta}$ dependiente del parámetro $\boldsymbol{\Sigma}$. Esto quiere decir que en la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$ el valor de $\boldsymbol{\Sigma}$ se considera una constante fija y conocida. Siguiendo los linemientos de la Sección 3.2, una distribución previa para $\boldsymbol{\theta}$ condicional a $\boldsymbol{\Sigma}$ es

$$p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}) \sim Normal_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/c_0)$$

Donde c_0 es una constante. Por otro lado, y siguiendo los argumentos de la sección anterior, una posible opción para la distribución previa de $\boldsymbol{\Sigma}$, corresponde a

$$p(\boldsymbol{\Sigma}) \sim Wishart - inversa_{v_0}(\boldsymbol{\Lambda}_0^{-1})$$

Resultado 4.3.3. *La distribución previa conjunta de los parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ está dada por*

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-(v_0+p)/2+1} \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1}) + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\}$$

Prueba. La prueba es inmediata al multiplicar las densidades y asignar los términos que no dependen de los parámetros de interés a la constante de proporcionalidad. ■

Para encontrar las distribuciones posterior de cada uno de los parámetros de interés se utilizan argumentos similares a los de la sección 3.1.2.

Resultado 4.3.4. *La distribución posterior de $\boldsymbol{\theta}$ condicional a $\boldsymbol{\Sigma}, \mathbf{Y}$ está dada por*

$$\theta \mid \sigma^2, \mathbf{Y} \sim Normal_p(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/(n + c_0))$$

donde

$$\mu_n = \frac{n\bar{\mathbf{Y}} + c_0\boldsymbol{\mu}}{n + c_0}$$

Resultado 4.3.5. *La distribución marginal posterior de la matriz de parámetros $\boldsymbol{\Sigma}$ es*

$$\sigma^2 \mid \mathbf{Y} \sim Whishart - inversa_{n+v_0}(\boldsymbol{\Lambda}_n^{-1})$$

Donde

$$\boldsymbol{\Lambda}_n = \boldsymbol{\Lambda}_0 + \mathbf{S} + \frac{c_0 n}{c_0 + n}(\boldsymbol{\mu} - \bar{\mathbf{y}})(\boldsymbol{\mu} - \bar{\mathbf{y}})'$$

Resultado 4.3.6. *La distribución marginal posterior del parámetro* $\boldsymbol{\theta}$ *es* $t-$ *student multivariante tal que*

$$\boldsymbol{\theta} \mid \mathbf{Y} \sim t_{n+v_0-p+1}\left(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Lambda}_n}{n+v_0-p+1}\right)$$

En términos de simulación de densidades se debe primero simular distribución $p(\boldsymbol{\Sigma} \mid \mathbf{Y})$ y encontrar un valor estimado para esta matriz de parámetros. Luego, se debe utilizar este valor para simular la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y})$ y encontrar un valor estimado para este vector de parámetros.

4.3.3 Parámetros no informativos

(**?**) afirma que la distribución previa no informativa de Jeffreys conjunta para $\boldsymbol{\theta}, \boldsymbol{\Sigma}$, en este caso está dada por la siguiente expresión

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-(p+1)/2}$$

Por lo tanto la distribución posterior conjunta para $\boldsymbol{\theta}, \boldsymbol{\Sigma}$ está dada por

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) \propto \mid \boldsymbol{\Sigma} \mid^{-(p+n+1)/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta})\right\}$$

Resultado 4.3.7. *La distribución posterior del vector de parámetros* $\boldsymbol{\theta}$ *condicional a* $\boldsymbol{\Sigma}, \mathbf{Y}$ *es*

$$\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y} \sim Normal_p(\bar{\mathbf{y}}, \boldsymbol{\Sigma}/n)$$

Prueba. Utilizando la identidad (3.2.5) se tiene que

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta})\right\} \\ &\propto \exp\left\{-\frac{n}{2}(\boldsymbol{\theta}-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\bar{\mathbf{Y}})\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal_p(\bar{y}, \boldsymbol{\Sigma}/n)$. ■

Resultado 4.3.8. *La distribución marginal posterior de la matriz de parámetros* $\boldsymbol{\Sigma}$ *es*

$$\boldsymbol{\Sigma} \mid \mathbf{Y} \sim Whishart - inversa_{n-1}(\mathbf{S})$$

donde $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})'$

Prueba. En primer lugar nótese que

$$\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{S} + n(\boldsymbol{\theta}-\bar{\mathbf{y}})(\boldsymbol{\theta}-\bar{\mathbf{y}})'$$

Por otro lado, recurriendo a las propiedades del operador $traza$, e integrando la distribución posterior conjunta con respecto a $\boldsymbol{\theta}$, se tiene que

$$
\begin{aligned}
p(\boldsymbol{\Sigma} \mid \mathbf{Y}) &= \int p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y})\, d\boldsymbol{\theta} \\
&= \mid \boldsymbol{\Sigma} \mid^{-(p+n+1)/2} \int \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_i - \boldsymbol{\theta}) \right\} d\boldsymbol{\theta} \\
&= \mid \boldsymbol{\Sigma} \mid^{-(p+n+1)/2} \int \exp\left\{ -\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1} \mathbf{S}_{\boldsymbol{\theta}}) \right\} d\boldsymbol{\theta} \\
&= \mid \boldsymbol{\Sigma} \mid^{-(p+n+1)/2} \int \exp\left\{ -\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1} (\mathbf{S} + n(\boldsymbol{\theta} - \bar{\mathbf{y}})(\boldsymbol{\theta} - \bar{\mathbf{y}})')) \right\} d\boldsymbol{\theta} \\
&= \mid \boldsymbol{\Sigma} \mid^{-(p+n)/2} \exp\left\{ -\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1} \mathbf{S}) \right\} \\
&\qquad \times \int \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{ -\frac{n}{2} traza(\boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \bar{\mathbf{y}})(\boldsymbol{\theta} - \bar{\mathbf{y}})') \right\} d\boldsymbol{\theta} \\
&= \mid \boldsymbol{\Sigma} \mid^{-(p+n)/2} \exp\left\{ -\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1} \mathbf{S}) \right\} \\
&\qquad \times \int \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{ -\frac{n}{2} traza((\boldsymbol{\theta} - \bar{\mathbf{y}})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \bar{\mathbf{y}})) \right\} d\boldsymbol{\theta} \\
&= \mid \boldsymbol{\Sigma} \mid^{-(p+n)/2} \exp\left\{ -\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1} \mathbf{S}) \right\} \\
&\qquad \times \int \underbrace{\mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{ -\frac{n}{2} (\boldsymbol{\theta} - \bar{\mathbf{y}})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \bar{\mathbf{y}}) \right\}}_{Normal_p(\bar{\mathbf{y}}, \boldsymbol{\Sigma}/n)} d\boldsymbol{\theta} \\
&= \mid \boldsymbol{\Sigma} \mid^{-(p+n)/2} \exp\left\{ -\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1} \mathbf{S}) \right\}
\end{aligned}
$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Whishart - inversa_{n-1}(\mathbf{S})$. ■

4.4 Multinomial

Suponga que $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ es un vector aleatorio con distribución Multinomial. Luego, su distribución está parametrizada por el vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ y está dada por la siguiente expresión

$$
p(\mathbf{Y} \mid \boldsymbol{\theta}) = \binom{n}{y_1, \ldots, y_p} \prod_{i=1}^{p} \theta_i^{y_i} \qquad \theta_i > 0 \;\; , \;\; \sum_{i=1}^{n} y_i = n \;\; \text{y} \;\; \sum_{i=1}^{p} \theta_i = 1 \qquad (4.4.1)
$$

donde

$$\binom{n}{y_1,\ldots,y_p} = \frac{n!}{y_1!\cdots y_p!}.$$

Como cada parámetro θ está restringido al espacio $\Theta = [0,1]$, entonces es posible asignar a la distribución de Dirichlet como la distribución previa del vector de parámetros. Por lo tanto la distribución previa del vector de parámetros $\boldsymbol{\theta}$, parametrizada por el vector de hiperparámetros $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_p)'$, está dada por

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1+\cdots+\alpha_p)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_p)} \prod_{i=1}^{p} \theta_i^{\alpha_i-1} \qquad \alpha_i > 0 \text{ y } \sum_{i=1}^{p} \theta_i = 1 \tag{4.4.2}$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 4.4.1. *La distribución posterior del parámetro $\boldsymbol{\theta}$ sigue una distribución Dirichlet$(y_1+\alpha_1,\ldots,y_p+\alpha_p)$*

Prueba.

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \\ &= \binom{n}{y_1,\ldots,y_p} \prod_{i=1}^{p} \theta_i^{y_i} \frac{\Gamma(\alpha_1+\cdots+\alpha_p)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_p)} \prod_{i=1}^{p} \theta_i^{\alpha_i-1} \\ &\propto \prod_{i=1}^{p} \theta_i^{y_i+\alpha_i-1} \end{aligned}$$

Dado que $\sum_{i=1}^{p} \theta_i = 1$, entonces factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una vector aleatorio con distribución *Dirichelt*$(y_1+\alpha_1,\ldots,y_n+\alpha_n)$. ■

Resultado 4.4.2. *La distribución predictiva previa para una observación* $\mathbf{y}$ *está dada por*

$$p(\mathbf{Y}) = \binom{n}{y_1,\ldots,y_p} \frac{\Gamma(\sum_{i=1}^{p}\alpha_i)}{\prod_{i=1}^{p}\Gamma(\alpha_i)} \frac{\prod_{i=1}^{p}\Gamma(y_i+\alpha_i)}{\Gamma(\sum_{i=1}^{p} y_i + \sum_{i=1}^{p}\alpha_i)} \tag{4.4.3}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned}
p(\mathbf{Y}) &= \int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \, d\boldsymbol{\theta} \\
&= \binom{n}{y_1, \ldots, y_p} \frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)} \frac{\Gamma(y_1 + \alpha_1) \cdots \Gamma(y_p + \alpha_p)}{\Gamma(y_1 + \alpha_1 + \cdots + y_p + \alpha_p)} \\
&\times \int_0^1 \cdots \int_0^1 \frac{\Gamma(y_1 + \alpha_1 + \cdots + y_p + \alpha_p)}{\Gamma(y_1 + \alpha_1) \cdots \Gamma(y_p + \alpha_p)} \prod_{i=1}^{p} \theta_i^{y_i + \alpha_i - 1} \, d\theta_1 \cdots d\theta_p \\
&= \binom{n}{y_1, \ldots, y_p} \frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)} \frac{\Gamma(y_1 + \alpha_1) \cdots \Gamma(y_p + \alpha_p)}{\Gamma(y_1 + \alpha_1 + \cdots + y_p + \alpha_p)} \\
&= \binom{n}{y_1, \ldots, y_p} \frac{\Gamma(\sum_{i=1}^p \alpha_i)}{\prod_{i=1}^p \Gamma(\alpha_i)} \frac{\prod_{i=1}^p \Gamma(y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p y_i + \sum_{i=1}^p \alpha_i)}
\end{aligned}$$

■

Resultado 4.4.3. *Después de la recolección de los datos, la distribución predictiva posterior para una nueva observación del vector aleatorio* $\tilde{\mathbf{y}}$ *de tamaño* p*, para* n^* *repeticiones del mismo experimento aleatorio, está dada por*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \binom{n^*}{\tilde{y}_1, \ldots, \tilde{y}_p} \frac{\Gamma(\sum_{i=1}^p (y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(y_i + \alpha_i)} \frac{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))} \tag{4.4.4}$$

Prueba. De la definición de función de distribución predictiva posterior se tiene que

$$\begin{aligned}
p(\tilde{\mathbf{y}} \mid \mathbf{Y}) &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \\
&= \binom{n^*}{\tilde{y}_1, \ldots, \tilde{y}_p} \frac{\Gamma(\sum_{i=1}^p (y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(y_i + \alpha_i)} \frac{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))} \\
&\times \int_0^1 \cdots \int_0^1 \frac{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)} \prod_{i=1}^{p} \theta_i^{\tilde{y}_i + y_i + \alpha_i - 1} \, d\theta_1 \cdots d\theta_p \\
&= \binom{n^*}{\tilde{y}_1, \ldots, \tilde{y}_p} \frac{\Gamma(\sum_{i=1}^p (y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(y_i + \alpha_i)} \frac{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))}
\end{aligned}$$

■

Ejemplo 4.4.1. *Aunque no estoy de acuerdo con la metodología de muestreo de la mayoría de las encuestas electorales, pienso que la acumulación de la información es de alguna forma ilustrativa. En esta entrada se realiza un análisis bayesiano acerca de la intención de voto para las próximas elecciones de la alcaldía de Bogotá, ciudad donde yo resido. El ejercicio es meramente académico y voy a actualizar los resultados de manera sistemática hasta el día de las elecciones.*

El análisis electoral desde el enfoque bayesiano puede parecer sencillo. En una primera instancia, se trata de conocer la probabilidad de éxito de un candidato, que aplicada a una población específica se traduce en la intención de voto hacia

el candidato. Como hay varios candidatos en la disputa, entonces es conveniente suponer que el fenómeno puede ser descrito muy bien mediante el uso de una distribución multinomial. Como el parámetro en este caso es un vector de probabilidades, es adecuado suponer una distribución previa de tipo Dirichlet para este vector. Por lo tanto, haciendo uso del teorema de Bayes, la distribución posterior del parámetro será también de tipo Dirichlet.

En esta primera entrada, desarrollaremos un análisis básico con base en una primera encuesta realizada del 12 al 14, en donde según el portal WEB de la revista Semana se afirma que:

≪Según la encuesta de Ipsos Napoleón Franco, hay un cabeza a cabeza (cada uno con el 22 %) entre los dos candidatos. Mockus es tercero, pero con notable diferencia: 12 %, seguido, muy cerca, por Gina Parody, con 9 %≫.

*Con base en esta información, y teniendo en cuenta que hubo 604 respondientes, se afina la distribución previa que es Dirichlet con parámetros 133 (igual a 604*0.22), 133 (604*0.22), 72 (604*0.12) y 64 (604*0.09), para los candidatos Peñalosa, Petro, Mockus y Parody, respectivamente. En las entradas posteriores se analizarán otras distribuciones previas que pueden ser más convenientes y/o tener ventajas en el análisis.*

Por otro lado, según la última encuesta electoral reportada por un medio de comunicación, correspondiente a la realizada por la firma Centro Nacional de Consultoría, entre el 30 de agosto y el primero de Septiembre, y publicada por el portal WEB de ElTiempo.com afirma que:

≪Peñalosa alcanza el 22 % de preferencia. Segundo aparece Gustavo Petro, con 17 %, en tercer lugar Antanas Mockus, con 12 %. El cuarto lugar es para la candidata Gina Parody, con 11 %≫.

Como se trata de la encuesta más reciente, supondremos que estos datos corresponden a la realización de una distribución multinomial.

Es bien sabido que el análisis conjugado, señala que la distribución posterior del parámetro es de tipo Dirichlet, que en este ejercicio particular, tiene parámetros 353, 302, 192 y 164, para los candidatos Peñalosa, Petro, Mockus y Parody, respectivamente. Después de realizar cien mil simulaciones de Monte Carlo y chequear la convergencia de las cadenas y todo lo otro que se deba chequear, los resultados se presentan a continuación:

Luego, la distribución posterior estima que Peñalosa será el ganador. Nada nuevo hasta acá. La novedad es que realicé un análisis para determinar la probabilidad posterior de que el parámetro de Peñalosa fuese mayor que el parámetro de Petro. Esta probabilidad es del orden de 0.97. Luego, la probabilidad de victoria de Peñalosa sobre Petro al día de hoy y, aunque sea muy difícil, suponiendo que los datos son válidos, es de 0.97.

Código WinBugs

```
model {
y[1:k] ~ dmulti(theta[1:k],n)
theta[1:k] ~ ddirch(alpha[1:k])
```

```
delta <- theta[1]-theta[2]
P <- step(delta)
}

DATA
list(k=4,alpha=c(133,133,72,54),y=c(220,170,120,110),n=620)
```

5 Modelos empíricos y jerárquicos

En las últimas décadas la formulación de modelos estadísticos ha evolucionado demasiado. En un principio, los modelos establecidos obedecían a reglas estándar que se suponían ciertas para toda la población. Sin embargo, el estado de la naturaleza de la mayoría de los problemas práctico no sigue una regla común para todos y cada uno de los elementos de una población aleatoria. De hecho el sentido común establece que para una misma población, pueden existir tendencias comunes entre diferentes miembros de la misma y la estructura de dispersión de los elementos puede obedecer comportamientos disímiles a través de éstos.

Lo anterior ha permitido que el investigador pueda proponer modelos que siguen comportamientos estructurales distintos y en algunos casos que se encuentran anidados en modelos más complejos. En el caso bayesiano, es claro que el momento de coyuntura en el cual el investigador no contempla un punto de retorno está dado en la formulación de la distribución previa para el vector de parámetros de interés $\boldsymbol{\theta}$. Más aún, la influencia de la distribución previa en la resultante distribución posterior está dada por la asignación del vector de hiperparámetros $\boldsymbol{\eta}$ que parametriza la distribución previa. Cuando los valores exactos de los hiperparámetros se desconocen o cuando no se tiene plena certeza del comportamiento estructural de la distribución previa, entonces es necesario estimarlos pues de estos dependen los resultados en cualquier investigación de tipo causal. En otras palabras, una mala asignación de los valores de los hiperparámetros conduce a una distribución previa que no es acorde con la realidad y esto puede conllevar a su vez a que la distribución posterior no concuerde con la realidad, produciendo así resultados engañosos.

Siguiendo los fundamentos filosóficos de la estadística bayesiana, tener que estimar el vector de hiperparámetros envuelve al investigador en una paradoja cuya solución no siempre está dada por métodos bayesianos. En primer lugar, nótese la forma de la distribución previa del vector de parámetros de interés: $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. A simple vista se puede concluir que $\boldsymbol{\eta}$ hace parte de la distribución previa la cual, según la lógica de la filosofía bayesiana, involucra el conocimiento del investigador antes de la recolección de los datos. Por tanto la pregunta directa que surge es ¿Por qué estimar algo que se debería suponer conocido?. En segundo lugar y si se concibe tal estimación, la otra pregunta natural es ¿Se deben utilizar los datos para estimar tales hiperparámetros?. Las posibles respuestas a las anteriores preguntas han creado toda una nueva corriente alterna a la bayesiana pura llamada ≪corriente bayesiana empírica≫[1] la cual utiliza los métodos de estimación pun-

[1] (**?**) menciona que el análisis empírico toma este nombre por dos razones: En primer lugar porque estima el vector de hiper-parámetros $\boldsymbol{\eta}$ con los datos observados, contradiciendo de alguna manera el espíritu y la filosofía de la corriente bayesiana radical. En segundo lugar, porque

tual frecuentista para estimar estos hiperparámetros y por consiguiente definir la distribución previa del vector de parámetros de interés.

LADY TASTING TEA SOBRE EMPIRICAL Y BAYES

Por supuesto, existe la contraparte teórica a la corriente empírica y es la llamada ≪corriente bayesiana jerárquica≫ la cual asume una posición totalmente bayesiana desde su concepción y establece un modelo posterior para los hiperparámetros.

Suponga entonces que la variable de interés sigue un modelo común a toda la población aunque parametrizado por parámetros que toman distintos valores para cada individuo y que está regido por la siguiente expresión

$$Y_i \sim p(Y_i \mid \theta_i)$$

5.1 Análisis empírico

COLOCAR LOS TIPOS DE ANÁLISIS: PARAMÉTRICO Y NO PARAMÉTRICO Y LOS PRINCIPALES RESULTADOS

Este enfoque, criticado por muchos bayesianos radicales, se centra en la escogencia de una estimación $\hat{\boldsymbol{\eta}}$ de $\boldsymbol{\eta}$ obtenida como el valor que hace maxima la verosimilitud marginal previa dada por

$$p(\mathbf{Y} \mid \boldsymbol{\eta}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \; d\boldsymbol{\theta} \tag{5.1.1}$$

Por lo tanto todo el andamiaje inferencial está supeditado a la distribución posterior estimada, $p(\boldsymbol{\theta} \mid Y, \hat{\boldsymbol{\eta}})$. Una vez que ésta esté bien definida, el proceso de estimación puntual, estimación por intervalo y pruebas de hipótesis sigue su curso bayesiano idénticamente como en los capítulos anteriores.

En términos prácticos suponga que se tiene un modelo en dos etapas para cada una de las observaciones. Se asume que existen n observaciones que, si bien no conforman una muestra aleatoria, conservan la característica de intercambiebilidad y están definidas en los siguientes términos

$$Y_i \sim p(Y_i \mid \theta_i) \qquad i = 1, \ldots, n$$

La segunda etapa comienza con la asignación de una distribución[2] previa para los parámetros de interés $theta_i$.

$$\theta_i \sim p(\theta_i \mid \boldsymbol{\eta}) \qquad i = 1, \ldots, n$$

Nótese que detrás de la asignación de la estructura probabilística para cada uno de los θ_i, se supone que éstos últimos determinan una muestra aleatoria de la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. El objetivo de este enfoque es encontrar estimadores que

esta estimación se realiza con métodos frecuentistas ya sean paramétricos o no-paramétricos

[2] En esta etapa la distribución previa no está completamente especificada puesto que se desconocen los hiperparámetros que la indexan.

maximicen la verosimilitud marginal previa la cual, para este caso particular y considerando independencia marginal entre las observaciones y el vector de hiperparámetros, es

$$\begin{aligned} p(Y_i \mid \boldsymbol{\eta}) &= \int p(Y_i, \theta_i \mid \boldsymbol{\eta})\ d\theta_i \\ &= \int p(Y_i \mid \theta_i, \boldsymbol{\eta}) p(\theta_i \mid \boldsymbol{\eta})\ d\theta_i \\ &= \int p(Y_i \mid \theta_i) p(\theta_i \mid \boldsymbol{\eta})\ d\theta_i \end{aligned} \tag{5.1.2}$$

De lo anterior, la verosimilitud marginal previa del vector de observaciones dada por la expresión (4.1.1) queda convertida en

$$\begin{aligned} p(Y \mid \boldsymbol{\eta}) &= \prod_{i=1}^{n} p(Y_i \mid \boldsymbol{\eta}) \\ &= \prod_{i=1}^{n} \int p(Y_i \mid \theta_i) p(\theta_i \mid \boldsymbol{\eta})\ d\theta_i \end{aligned} \tag{5.1.3}$$

5.1.1 Modelo Binomial-Beta

Suponga el siguiente modelo binomial (intercambiable) en una primera etapa

$$Y_i \mid \theta_i \sim Binomial(n_i, \theta_i) \qquad i = 1, \ldots, p$$

Para la segunda etapa, se supone una muestra aleatoria (independientes e idénticamente distribuidos) proveniente de una misma distribución tal que

$$\theta_i \sim Beta(\alpha, \beta) \qquad i = 1, \ldots, p$$

Análisis preliminar

Es bien sabido que la distribución posterior para cada uno de los parámetros de interés involucrados en el anterior contexto está dada por

$$\theta_i \mid Y_i \sim Beta(\alpha + Y_i, \beta + n_i - y_i)$$

Sin embargo, como se desconoce totalmente el valor de los hiperparámetros α y β, entonces se debe encontrar una estimación de estos, $\hat{\alpha}$ y $\hat{\beta}$, respectivamente, para proseguir normalmente con la inferencia bayesiana, pero esta vez enfocados en la estimación de la distribución posterior dada por

$$\theta_i \mid Y_i \sim Beta(\hat{\alpha} + Y_i, \hat{\beta} + n_i - y_i)$$

Para tal fin, nótese que la esperanza y la varianza previa de θ_i están dadas por las siguientes expresiones

$$E(\theta_i) = \frac{\alpha}{\alpha+\beta} \tag{5.1.4}$$

$$Var(\theta_i) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{5.1.5}$$

De (4.1.2) se tiene que

$$\alpha = E(\theta_i)(\alpha+\beta) \tag{5.1.6}$$

y también que

$$1 - E(\theta_i) = \frac{\beta}{\alpha+\beta}$$

por lo tanto

$$\beta = (1 - E(\theta_i))(\alpha+\beta) \tag{5.1.7}$$

y reemplazando (4.1.4) y (4.1.5) en (4.1.3) se concluye que

$$Var(\theta_i) = \frac{E(\theta_i)(1-E(\theta_i))}{(\alpha+\beta+1)}$$

por tanto

$$\alpha+\beta = \frac{E(\theta_i)(1-E(\theta_i))}{Var(\theta_i)} - 1 \tag{5.1.8}$$

Con el anterior razonamiento, es posible encontrar los estimadores basados en el método frecuentista de los momentos los cuales corresponden a

$$\widehat{\alpha+\beta} = \frac{\bar{Y}(1-\bar{Y})}{S^2} - 1 \tag{5.1.9}$$

$$\hat{\alpha} = (\widehat{\alpha+\beta})\bar{Y} \tag{5.1.10}$$

$$\hat{\beta} = (\widehat{\alpha+\beta})(1-\bar{Y}) \tag{5.1.11}$$

Donde $\bar{Y}$ y S^2 es el promedio y la varianza de las cantidades $Y_1/n_1, Y_2/n_2, \ldots, Y_p/n_p$, respectivamente. Con las anteriores estimaciones es posible ahora conectarlas a la distribución posterior de θ_i.

Análisis legítimo

Según **?**, p. 119, el anterior análisis no implica simplemente un punto de partida que da pie a la exploración de la idea de la estimación de los parámetros de

la distribución posterior y, de ninguna manera, constituye un cálculo bayesiano puesto que no está basado en ningún modelo de probabilidad. Sin embargo, el análisis empírico de esta situación, hace uso de la esperanza y varianza condicional a la distribución beta de los parámetros θ_i $(i = 1, \ldots, p)$.

Para realizar este tipo de análisis, vamos a suponer que contamos con una variable Y, distribuida de forma binomial en n ensayos y con probabilidad de éxito θ. De esta manera, se tiene que el primer momento está dado por

$$\begin{aligned} E_{binom}\left(\frac{Y}{n}\right) &= E_{beta}\left(E_{binom}\left(\frac{Y}{n} \mid \theta\right)\right) \\ &= E_{beta}\left(\theta\right) \\ &= \frac{\alpha}{\alpha+\beta} \end{aligned} \tag{5.1.12}$$

Por otro lado, se tiene que la varianza, que es función del primer y segundo momento, está dada por

$$\begin{aligned} Var_{binom}\left(\frac{Y}{n}\right) &= E_{beta}\left(Var_{binom}\left(\frac{Y}{n} \mid \theta\right)\right) + Var_{beta}\left(E_{binom}\left(\frac{Y}{n} \mid \theta\right)\right) \\ &= E_{beta}\left(\frac{1}{n}\theta(1-\theta)\right) + Var_{beta}\left(\theta\right) \\ &= \frac{1}{n}E_{beta}(\theta) - \frac{1}{n}E_{beta}(\theta^2) + Var_{beta} \\ &= \frac{1}{n}E_{beta}(\theta) - \frac{1}{n}Var_{beta}(\theta) - \frac{1}{n}E_{beta}(\theta^2) + Var_{beta} \\ &= \frac{n-1}{n}Var_{beta}(\theta) + \frac{1}{n}E_{beta}(\theta)(1 - E_{beta}(\theta)) \\ &= \frac{n-1}{n}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{1}{n}\frac{\alpha\beta}{(\alpha+beta)^2} \\ &= \frac{1}{n}\frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\left(\frac{n-1}{\alpha+\beta+1}+1\right) \\ &= \frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)\left(\frac{n-1}{\alpha+\beta+1}+1\right) \end{aligned}$$

De esta última expresión, y despejando $\alpha + \beta$, se tiene que

$$\begin{aligned} \alpha+\beta &= \frac{(n-1)E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)}{nVar_{binom}\left(\frac{Y}{n}\right) - E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)} - 1 \\ &= \frac{E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right) - Var_{binom}\left(\frac{Y}{n}\right)}{Var_{binom}\left(\frac{Y}{n}\right) - \frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)} \end{aligned} \tag{5.1.13}$$

Ahora, despejando α de la expresión (4.1.12) se tiene que

$$\alpha = E_{binom}\left(\frac{Y}{n}\right)\frac{E_{binom}\left(\frac{Y}{n}\right)\left(1-E_{binom}\left(\frac{Y}{n}\right)\right)-Var_{binom}\left(\frac{Y}{n}\right)}{Var_{binom}\left(\frac{Y}{n}\right)-\frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1-E_{binom}\left(\frac{Y}{n}\right)\right)} \tag{5.1.14}$$

además, también despejando β de (4.1.12) se tiene que

$$\begin{aligned}\beta &= \frac{\alpha\left(1-E_{binom}\left(\frac{Y}{n}\right)\right)}{E_{binom}\left(\frac{Y}{n}\right)} \\ &= \frac{E_{binom}\left(\frac{Y}{n}\right)\left(1-E_{binom}\left(\frac{Y}{n}\right)\right)-Var_{binom}\left(\frac{Y}{n}\right)}{Var_{binom}\left(\frac{Y}{n}\right)-\frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1-E_{binom}\left(\frac{Y}{n}\right)\right)}\end{aligned} \tag{5.1.15}$$

El anterior enfoque nos ha llevado a poder expresar los parámetros de interés en términos de $E_{binom}\left(\frac{Y}{n}\right)$, $Var_{binom}\left(\frac{Y}{n}\right)$ y n. Una vez que podamos estimar las anteriores cantidades, es posible realizar la inferencia bayesiana empírica de la manera correcta. Para lo anterior, es necesario observar al naturaleza de las observaciones que, aunque no representan una muestra aleatoria, sí son una sucesión de variables aleatorias intercambiables. Por lo anterior, y teniendo en cuenta que la inferencia se realiza con las cantidades $Y_1/n_1, Y_2/n_2, \ldots, Y_p/n_p$, es posible proponer los siguientes estimadores

$$\hat{E}_{binom}\left(\frac{Y}{n}\right) = \bar{Y} \tag{5.1.16}$$

$$\hat{Var}_{binom}\left(\frac{Y}{n}\right) = S^2 \tag{5.1.17}$$

$$\hat{n} = \frac{1}{p}\sum_{i=1}^{p} n_i \tag{5.1.18}$$

Con base en lo anterior, unas estimaciones empíricas de los parámetros α y β son

$$\hat{\alpha} = \bar{Y}\left(\frac{\bar{Y}\left(1-\bar{Y}\right)-S^2}{S^2-\frac{1}{\hat{n}}\bar{Y}\left(1-\bar{Y}\right)}\right) \tag{5.1.19}$$

y

$$\hat{\beta} = \frac{\bar{Y}\left(1-\bar{Y}\right)-S^2}{S^2-\frac{1}{\hat{n}}\bar{Y}\left(1-\bar{Y}\right)} \tag{5.1.20}$$

respectivamente. Cuando la cantidad de ensayos n_i es diferente en cada experimento, existen otras formas de obtener estimaciones para los parámetros α y β (**?**, p. 81).

5.1.2 Modelo Poisson-Gamma

Suponga el siguiente modelo de Poisson intercambiable

$$Y_i \mid \theta_i \sim Poisson(\theta_i)$$

y la distribución del parámetro θ_i (i=1,...,n) es

$$\theta_i \sim Gamma(\alpha, \beta)$$

Donde α y β son hiperparámetros desconocidos. Utilizando el resultado 2.4.1, se tiene que la distribución posterior de cada parámetro θ_i está dada por

$$\theta_i \mid \mathbf{Y} \sim Gamma\left(\sum_{i=1}^{n} Y_i + \alpha, \beta + n\right)$$

Por supuesto, la distribución anterior es inútil a no ser que los hiperparámetros puedan ser estimados. Para realizar esta estimación, el enfoque empírico sugiere utilizar el método de los momentos. Para esto, nótese que el primer momento está dado por

$$\begin{aligned} E_{Poisson}(Y_i) &= E_{Gamma}\left(E_{Poisson}(Y_i \mid \theta_i)\right) \\ &= E_{Gamma}\left(\theta_i\right) \\ &= \frac{\alpha}{\beta} \end{aligned} \tag{5.1.21}$$

Mientras que la varianza, función del primer y segundo momento, está dada por

$$\begin{aligned} Var_{Poisson}(Y_i) &= E_{Gamma}\left(Var_{Poisson}(Y_i \mid \theta_i)\right) + Var_{Gamma}\left(E_{Poisson}(Y_i \mid \theta_i)\right) \\ &= E_{Gamma}\left(\theta_i\right) + Var_{Gamma}\left(\theta_i\right) \\ &= \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2} \\ &= \frac{\alpha}{\beta}(\beta + 1) \end{aligned} \tag{5.1.22}$$

Ahora, siguiendo el enfoque del método de los momentos, es claro que la expresión (4.1.12) puede ser estimada con con la media muestral, $\bar{Y}$; mientras que la expresión (4.1.13) puede ser estimada con la varianza muestral, S^2. Por otro lado, al dividir estas expresiones se tiene que

$$\frac{\frac{\alpha}{\beta}(\beta + 1)}{\frac{\alpha}{\beta}} = 1 + \frac{1}{\beta} \tag{5.1.23}$$

y, siguiendo un razonamiento similar, esta última expresión es estimada por $S^2/\bar{Y}$. Por tanto, un estimador del método de los momentos para β es

$$\hat{\beta} = \frac{1}{\frac{S^2}{\bar{Y}} - 1} = \frac{\bar{Y}}{S^2 - \bar{Y}} \tag{5.1.24}$$

De la expresión (4.1.12), se nota que $\alpha = \beta E_{Poisson}(Y_i)$. Por tanto, un estimador del método de los momentos para α es

$$\hat{\alpha} = \hat{\beta}\bar{Y} = \frac{\bar{Y}^2}{S^2 - \bar{Y}} \tag{5.1.25}$$

De lo anterior, se tiene que, siguiendo el enfoque bayesiano empírico, la distribución posterior para θ_i está dada por

$$\theta_i \mid \mathbf{Y} \sim Gamma\left(\sum_{i=1}^{n} Y_i + \hat{\alpha}, \hat{\beta} + n\right)$$

5.1.3 Modelo Normal-Normal

Uno de los modelos más utilizados en la aplicaciones prácticas se da cuando la verosimilitud de los datos sigue una distribución normal. Considere el siguiente modelo en dos etapas en donde cada una de las observaciones se supone intercambiable y para la primera etapa se tiene que

$$Y_i \mid \theta_i \sim Normal(\theta_i, \sigma^2) \qquad i = 1, \ldots, n$$

en donde el parámetro σ^2 se supone conocido. En la segunda etapa, la distribución previa para los parámetros de interés θ_i es

$$\theta_i \mid \mu \sim Normal(\mu, \tau^2) \qquad i = 1, \ldots, n$$

en donde el parámetro τ^2 se supone conocido.Para poder proseguir con el análisis empírico bayesiano, se necesitan los siguientes resultados.

Resultado 5.1.1. *La verosimilitud marginal previa de una observación condicional al hiperparámetro μ es*

$$Y_i \mid \mu \sim Normal(\mu, \sigma^2 + \tau^2) \qquad i = 1, \ldots, n$$

Prueba. Desarrollando la expresión (4.1.2) se tiene que

$$
\begin{aligned}
p(Y_i \mid \mu) &= \int p(Y_i \mid \theta_i) p(\theta_i \mid \mu)\, d\theta_i \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i-\theta_i)^2}{\sigma^2}\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2}\frac{(\theta_i-\mu)^2}{\tau^2}\right\}\, d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{1}{2}\frac{\tau^2(\theta_i-y_i)^2+\sigma^2(\theta_i-\mu)^2}{\sigma^2\tau^2}\right\}\, d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left[\theta_i^2(\tau^2+\sigma^2)-2\theta_i(y_i\tau^2+\mu\sigma^2)+\tau^2y_i^2+\sigma^2\mu^2\right]\right\}\, d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\theta_i^2-2\theta_i\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}+\frac{\tau^2y_i^2+\sigma^2\mu^2}{\tau^2+\sigma^2}\right]\right\}\, d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[(\theta_i-\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2})^2-\left(\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)^2+\frac{\tau^2y_i^2+\sigma^2\mu^2}{\tau^2+\sigma^2}\right]\right\}\, d\theta_i \\
&= \int \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma^2\tau^2}{\tau^2+\sigma^2}}} \exp\left\{-\frac{1}{2\frac{\sigma^2\tau^2}{\tau^2+\sigma^2}}\left(\theta_i-\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)^2\right\}\, d\theta_i \\
&\times \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2\tau^2}}\sqrt{\frac{\sigma^2\tau^2}{\tau^2+\sigma^2}} \exp\left\{\frac{(y_i\tau^2+\mu\sigma^2)^2}{2\sigma^2\tau^2(\tau^2+\sigma^2)}-\frac{\tau^2y_i^2+\sigma^2\mu^2}{2\sigma^2\tau^2}\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[\frac{(\tau^2y_i^2+\sigma^2\mu^2)(\tau^2+\sigma^2)}{\sigma^2\tau^2}-\frac{(y_i\tau^2+\mu\sigma^2)^2}{\sigma^2\tau^2}\right]\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[\frac{\tau^2\sigma^2y_i^2+\sigma^2\tau^2\mu^2-2y_i\mu\tau^2\sigma^2}{\sigma^2\tau^2}\right]\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[y_i^2+\mu^2-2y_i\mu\right]\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[y_i-\mu\right]^2\right\}
\end{aligned}
$$

la cual corresponde a la función de distribución de una variable aleatoria con densidad $Normal(\mu, \sigma^2+\tau^2)$ ■

Del anterior resultado, y siguiendo la ecuación (4.1.3), se tiene que la verosimilitud marginal previa del vector de observaciones $\mathbf{Y} = (Y_i, \ldots, Y_n)'$ condicionado al hiperparámetro μ es

$$
p(\mathbf{Y} \mid \mu) = \left(\frac{1}{2\pi(\tau^2+\sigma^2)}\right)^{n/2} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\sum_{i=1}^{n}(y_i-\mu)^2\right\}
$$

El objetivo del enfoque empírico bayesiano es encontrar una estadística que maximice la anterior expresión. No es difícil notar que un estimador de máxima verosimilitud para μ está dado por la media muestral $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Con este

estimador para el hiperparámetro se considera que las distribuciones previa y posterior del parámetro de interés quedan totalmente definidas y es posible continuar con el análisis bayesiano común.

5.2 Análisis jerárquico

La verosimilitud en una muestra aleatoria $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ parametrizada por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ está dada por

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(Y_i \mid \theta_i) \tag{5.2.1}$$

Por otro lado, suponga que la distribución previa del parámetro de interés θ_i está parametrizada por un vector de hiperparámetros $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)$ tal que

$$\theta_i \sim p(\theta_i \mid \boldsymbol{\eta})$$

De lo anterior, y suponiendo que existe intercambiabilidad entre cada uno de los parámetros de interés, la distribución previa del vector de parámetros $\boldsymbol{\theta}$, parametrizada por $\boldsymbol{\eta}$ está dada por

$$p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} p(\theta_i \mid \boldsymbol{\eta}) \tag{5.2.2}$$

Por tanto, es posible formular una distribución previa conjunta para $\boldsymbol{\theta}, \boldsymbol{\eta}$ que al igual que en capítulos anteriores, teniendo en cuenta el espíritu jerárquico y dependiente, vendría dada por

$$p(\boldsymbol{\theta}, \boldsymbol{\eta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \tag{5.2.3}$$

Luego, la distribución marginal previa del vector de parámetros de interés viene dada por

$$\begin{aligned} p(\boldsymbol{\theta}) &= \int p(\boldsymbol{\theta}, \boldsymbol{\eta}) \; d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \; d\boldsymbol{\theta} \\ &= \int \cdots \int \prod_{i=1}^{n} p(\theta_j \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \; d\theta_1 \cdots d\theta_n \end{aligned}$$

Con esta formulación, y suponiendo que las observaciones son condicionalmente independientes del vector de hiperparámetros $\boldsymbol{\eta}$[3],la distribución posterior conjunta

[3]Esta suposición tiene como base que las observaciones sólo dependen de $\boldsymbol{\eta}$ a través del vector de parámetros de interés $\boldsymbol{\theta}$.

para $\boldsymbol{\theta}, \boldsymbol{\eta}$ es

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ &= p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \\ &= p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \end{aligned} \tag{5.2.4}$$

Nótese que tanto para la distribución previa como para la distribución posterior de los parámetros, se supone conocido la distribución marginal previa de $\boldsymbol{\eta}$, $p(\boldsymbol{\eta})$, y también la distribución previa de $\boldsymbol{\theta}$ condicional a $\boldsymbol{\eta}$, $p(\boldsymbol{\theta}, \boldsymbol{\eta})$. La anterior formulación es acorde con la filosofía jerárquica pues supone relaciones de dependencia en distintos niveles. Conociendo el comportamiento estructural de $\boldsymbol{\eta}$ se puede conocer el comportamiento estructural de $\boldsymbol{\theta}$. (**?**) afirma que cuando no se tiene certeza acerca del comportamiento de $\boldsymbol{\eta}$ se debe utilizar una distribución previa no informativa aunque siempre se debe tener alguna sospecha acerca del espacio paramétrico al cual sea posible restringirlos.

En términos de estimación, los siguientes pasos son esenciales para realizar un análisis bayesiano propiamente dicho (**?**):

1. Escribir la distribución posterior de $\boldsymbol{\theta}, \boldsymbol{\eta}$ de forma no normalizada como en la expresión (4.2.4).
2. Determinar analíticamente la distribución posterior de $\boldsymbol{\theta}$ condicional a $\boldsymbol{\eta}, \mathbf{Y}$, utilizando la siguiente regla

   $$p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, \mathbf{Y}) \propto p(\boldsymbol{\theta}, \underbrace{\boldsymbol{\eta}}_{fijo} \mid \mathbf{Y})$$

   Es decir, los términos que no dependan de $\boldsymbol{\theta}$ pueden ser introducidos en la constante de proporcionalidad.
3. Determinar la distribución posterior de $\boldsymbol{\eta}$, utilizando alguna de las siguientes expresiones (se debe escoger la más conveniente dependiendo del contexto del problema):

   $$p(\boldsymbol{\eta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\eta}) p(\beta) \tag{5.2.5}$$

   $$p(\boldsymbol{\eta} \mid \mathbf{Y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{Y}) d\boldsymbol{\theta} \tag{5.2.6}$$

   $$p(\boldsymbol{\eta} \mid \mathbf{Y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{Y})}{p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, \mathbf{Y})} \tag{5.2.7}$$

4. Por medio de $p(\boldsymbol{\eta} \mid \mathbf{Y})$ encontrar una estimación para $\boldsymbol{\eta}$.
5. Recurriendo a $p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, \mathbf{Y})$ encontrar una estimación para $\boldsymbol{\theta}$.

5.2.1 Modelo Binomial

Suponga el mismo modelo binomial del Ejemplo 4.1.1 dado por

$$Y_i \sim Binomial(n_i, \theta_i)$$

en donde la distribución de los parámetros de interés es tal que

$$\theta_i \sim Beta(\alpha, \beta)$$

Para esta situación el análisis bayesiano propiamente dicho requiere el planteamiento de un modelo que contemple el comportamiento estructural tanto del vector de parámetros $\boldsymbol{\theta}$ como de los hiperparámetros α, β. Por lo tanto, suponiendo que existe total ignorancia acerca del comportamiento estructural de los hiperparámetros, la distribución previa marginal para los hiperparámetros es no informativa (**?**), ésta está dada por

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

Suponiendo que los parámetros de interés θ_i $(i = 1, \ldots, n)$ conforman una muestra aleatoria, entonces su distribución previa es

$$\begin{aligned} p(\boldsymbol{\theta} \mid \alpha, \beta) &= \prod_{i=1}^{n} p(\theta_i \mid \alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} \theta_i^{\alpha - 1} (1 - \theta_i)^{\beta - 1} \end{aligned}$$

Por último, teniendo en cuenta que la distribución de las observaciones es intercambiable, entonces es posible definir la verosimilitud de la muestra como una productoria tal que

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(Y_i \mid \theta_i) \\ &\propto \prod_{i=1}^{n} \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i} \end{aligned}$$

De esta manera, siguiendo la expresión (4.2.4), la distribución posterior conjunta estaría dada por

$$p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{Y}) \propto (\alpha + \beta)^{-5/2} \prod_{i=1}^{n} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} \theta_i^{\alpha - 1} (1 - \theta_i)^{\beta - 1} \prod_{i=1}^{n} \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i}$$

Utilizando la regla del condicionamiento, la distribución posterior del vector de parámetros de interés condicionado a los hiperparámetros y a los datos observados es

$$\begin{aligned} p(\boldsymbol{\theta} \mid \alpha, \beta, \mathbf{Y}) &\propto p(\boldsymbol{\theta}, \underbrace{\alpha, \beta}_{fijos} \mid \mathbf{Y}) \\ &\propto \prod_{i=1}^{n} \theta_i^{\alpha - 1} (1 - \theta_i)^{\beta - 1} \prod_{i=1}^{n} \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i} \\ &\propto \prod_{i=1}^{n} \theta_i^{\alpha + y_i - 1} (1 - \theta_i)^{\beta + n_i - y_i - 1} \end{aligned}$$

De donde se concluye que la distribución posterior para el vector de parámetros de interés es

$$\theta_i \mid \alpha, \beta, Y_i \sim Beta(\alpha + Y_i, \beta + n_i - y_i)$$

Por supuesto, la anterior distribución es totalmente inútil frente al desconocimiento de los hiperparámetros que deben ser estimados posterior, en este caso particular, utilizando la expresión (4.2.7) la cual da como resultado

$$\begin{aligned} p(\alpha, \beta \mid \mathbf{Y}) &= \frac{p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{Y})}{p(\boldsymbol{\theta} \mid \alpha, \beta, \mathbf{Y})} \\ &\propto \frac{(\alpha+\beta)^{-5/2} \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}}{\prod_{i=1}^{n} \frac{\Gamma(\alpha+y_i+\beta+n_i-y_i)}{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)} \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}} \\ &= (\alpha+\beta)^{-5/2} \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)} \end{aligned}$$

Aunque la anterior distribución no tiene una forma cerrada o conocida, es posible simular valores provenientes de ésta utilizando el método de la grilla descrito con detalle en el apéndice. Una vez que se tienen las observaciones simuladas, entonces se encuentra un estimador para los hiperparámetros y con estos, la distribución posterior del vector de parámetros de interés queda correctamente definida.

En R, una función que calcula la probabilidad posterior para los hiperparámetros está dada por

```
> post<-function(a,b,n,y){
+ P1<- gamma(a+b)
+ P2<- gamma(a)*gamma(b)
+ P3<- gamma(a+y)*gamma(b+n-y)
+ P4<- gamma(a+b+n)
+ (a+b)^(-5/2)*prod((P1/P2)*(P3/P4))
+ }
```

Para implementar el método de la grilla, se debe tener en cuenta que como la distribución es bivariada entonces la grilla debe estar contenida en $\mathbb{N}^2$. En R, una función que devuelve una grilla bivariada está dada por el siguiente código

```
> grilla<-function(a,b){
+ A<-seq(1:length(a))
+ unoA <-rep(1,length(A))
+ B<-seq(1:length(b))
+ unoB <-rep(1,length(B))
+ P1<-kronecker(A,unoB)
+ P2<-kronecker(unoA,B)
+ grid<-cbind(a[P1],b[P2])
+ return(grid)
+ }
```

Para el ejemplo de **?**, p.129, se creó una grilla bivariada contenida entre $\{1, 1.5 \ldots, 49.5, 50\}^2$ (el superíndice denota el producto cartesiano) y para cada punto se calculó la respectiva probabilidad dada por la distribución posterior.

```
> n<-c(20,19,18,20,20,20,23,20,18,18,10,13,48,19,20,18,25,49,48,19,22,
+ 20,17,24,19,50,19,20,20,20,23,46,20,19,20,20,20,20,27,20,22,20,
+ 20,20,20,17,20,46,52,19,20,20,49,20,49,47,19,19,20,47,20,20,46,19,
+ 19,20,20,20,20,24)
> y<-c(0,0,1,2,3,4,6,0,0,1,1,2,10,5,0,0,2,5,9,4,6,
+ 0,0,2,2,10,4,6,0,1,2,5,4,4,6,0,1,2,3,4,5,6,
+ 0,1,2,2,4,11,16,0,1,2,7,4,12,15,0,1,2,7,4,5,15,0,1,2,3,4,5,9)

> a.grid<-seq(1,50,by=0.5)
> b.grid<-seq(1,50,by=0.5)
> ab.grid<-grilla(a.grid,b.grid)
> N.grid<-dim(ab.grid)[1]

> p.ab <- rep(NA, N.grid)
> for(j in 1:N.grid){
+ p.ab[j] <- post(ab.grid[j,1], ab.grid[j,2], n, y)
+ }
```

Luego, se utilizó la función `sample` del ambiente computacional R para generar una muestra aleatoria de tamaño `n=1000` proveniente de la distribución posterior normalizada de los hiperparámetros

```
> p.ab<-as.vector(p.ab/sum(p.ab))
> sum(p.ab)
[1] 1

> r.post<-sample(N.grid,5000,prob=p.ab,replace=T)
> rab.post<-ab.grid[r.post,]
> ra.post<-rab.post[,1]
> rb.post<-rab.post[,2]
```

El objeto `rab.post` es una matriz de dos columnnas y cinco mil filas. Cada fila contiene una observación simulada de la distribución posterior; por tanto, `rab.post` contiene cinco mil duplas simuladas. A continuación, es posible obtener estimaciones puntuales posterior para el vector de hiperparámetros; teniendo en cuenta el criterio de mínima pérdida cuadrática, estas estimaciones son $(\hat{\alpha}, \hat{\beta})' = (2.3675, 14.2862)'$. De la misma manera, también es posible obtener intervalos de credibilidad al 95 %.

```
> mean(ra.post)
[1] 2.3675
> mean(rb.post)
[1] 14.2862
```

```
> quantile(ra.post,c(0.025,0.975))
 2.5% 97.5%
  1.0   4.5
> quantile(rb.post,c(0.025,0.975))
   2.5%   97.5%
 7.0000 27.0125
```

Aunque el objetivo primario del análisis jerárquico es obtener una estimación bayesiana de los hiperparámetros para conectarla directamente a la distribución posterior de cada uno de los parámetros de interés ($theta_i, \quad i = 1, \ldots, n$), es posible preguntarse acerca de la forma estructural de la distribución posterior de los hiperparámetros. De esta manera, un primer acercamiento gráfico se presenta cuando se genera el contorno bivariado para la distribución, se puede notar que la distribución posterior conjunta para $(\alpha, \beta)'$ no tiene una forma conocida y tiene varios picos, justo como se ve en la en la figura XXXX. La forma de la distribución en tercera dimensión, considerada la figura XXXX, comprueba que, en efecto, esta distribución debe ser tratada en conjunto. Por otra parte, se resalta la potencia del método de la grilla que permite simular observaciones bivariadas de una distribución compleja como la desarrollada acá. El código computacional para generar estas gráficas se presenta a continuación.

```
> a<-a.grid
> b<-b.grid

> mat<-matrix(NA, nrow=length(a), ncol=length(b))
> for(i in 1:length(a)){
+ for(j in 1:length(b)){
+ mat[i,j]<-post(a[i],b[j],n,y)
+ }
+ }

> mat<-mat/(sum(mat))
> contour(a,b,mat, xlim=c(1,4), ylim=c(5,22))
> persp(a,b,mat, xlim=c(1,3), ylim=c(5,15),theta=-90,phi=20)
```

5.2.2 Modelo Poisson

Suponga el mismo modelo Poisson de la sección anterior dado por

$$Y_i \mid \theta_i \sim Poisson(\theta_i)$$

donde los Y_i forman una sucesión de variables aleatorias intercambiables y con cada parámetro θ_i $(i = 1, \ldots, n)$ distribuido como

$$\theta_i \mid (\alpha, \beta) \sim Gamma(\alpha, \beta)$$

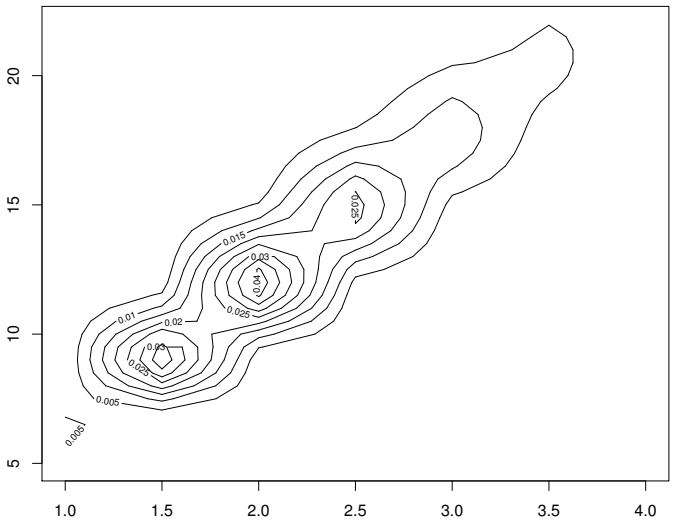

Figura 5.1: *Contorno de la distribución posterior de los hiperparámetros* $(\alpha, \beta)'$

Donde α y β son hiperparámetros desconocidos que vienen de distribuciones Gamma tales que

$$\alpha \sim Gamma(a, b)$$
$$\beta \sim Gamma(c, d)$$

Usualmente los parámetros a, b, c y d son conocidos y tales que las distribuciones de α y β sean planas o no-informativas. De esta manera, el enfoque bayesiano jerárquico plantea que se debe hacer la inferencia conjunta para el vector de parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ y para $(\alpha, beta)'$. Can base en lo anterior, la distribución posterior de los parámetros de interés toma la siguiente forma

$$\begin{aligned} p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{Y}) &\propto \prod_{i=1}^{n} p(Y \mid \theta_i) p(\theta_i \mid \alpha, \beta) p(\alpha) p(\beta) \\ &\propto \prod_{i=1}^{n} \frac{e^{\theta_i} \theta_i^{y_i}}{y_i!} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta_i^{\alpha-1} e^{-\beta\theta_i} e^{-\alpha b} \alpha^{a-1} e^{-\beta d} \beta^{c-1} \end{aligned}$$

Como es usual, y acudiendo a la anterior distribución posterior conjunta, se utilizará la técnica del condicionamiento sucesivo para encontrar las distribuciones posterior marginales de cada uno de los parámetros de interés. En este orden de ideas, se tiene que para cada $theta_i$ $(i = 1, \ldots, n)$ la distribución posterior marginal

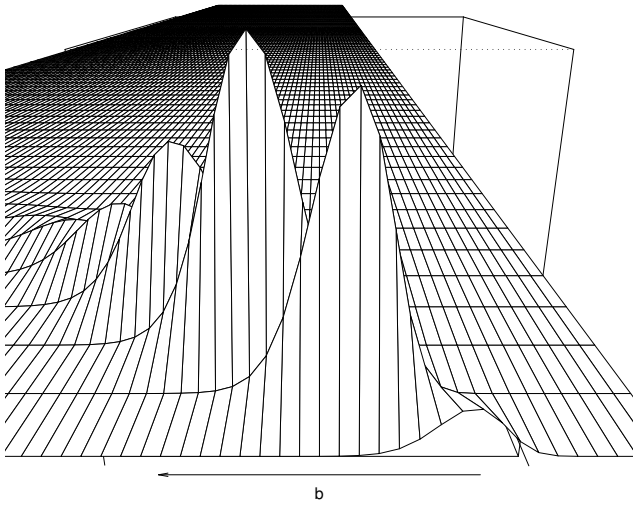

Figura 5.2: *Densidad bivariada de la distribución posterior de los hiperparámetros* $(\alpha, \beta)'$

está dada por

$$\begin{aligned} p(\theta_i \mid \alpha, \beta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \theta_n, \mathbf{Y}) &\propto p(\theta_i, \underbrace{\alpha, \beta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \theta_n}_{fijos} \mid \mathbf{Y}) \\ &\propto e^{\theta_i} \theta_i^{y_i} \theta_i^{\alpha-1} e^{-\beta\theta_i} \\ &= \exp\{-\theta_i(\beta+1)\} \theta^{y_i+\alpha-1} \end{aligned}$$

Con base en lo anterior, se tiene que la distribución posterior para cada parámetro θ_i es

$$\theta_i \mid \alpha, \beta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \theta_n, \mathbf{Y} \sim Gamma(y_i + \alpha, \beta + 1)$$

Para el hiperparámetro α, se tiene que la distribución posterior marginal está dada por

$$\begin{aligned} p(\alpha \mid \boldsymbol{\theta}\beta, \mathbf{Y}) &\propto p(\alpha, \underbrace{\beta, \boldsymbol{\theta}}_{fijos} \mid \mathbf{Y}) \\ &\propto \prod_{i=1}^{n} \frac{\theta_i^{\alpha-1}}{\Gamma(\alpha)} \alpha^{a-1} \exp\{-\alpha b\} \end{aligned}$$

La anterior distribución no tiene una forma conocida y es necesario utilizar métodos numéricos para simular observaciones de provenientes de ésta. Para esto es posible utilizar el método de la grilla descrito con detenimiento en el apéndice de este libro.

Por último, la distribución posterior marginal del hiperparámetro β se encuentra, similarmente mediante el condicionamiento sucesivo, de la siguiente forma

$$\begin{aligned} p(\beta \mid \alpha, \boldsymbol{\theta}, \mathbf{Y}) &\propto p(\beta, \underbrace{\alpha, \boldsymbol{\theta}}_{fijos} \mid \mathbf{Y}) \\ &\propto \prod_{i=1}^{n} \beta^{\alpha} \exp\{-\beta\theta_i\} \beta^{c-1} \exp\{-\beta d\} \\ &= \beta^{\alpha+c-1} \exp\left\{-\beta\left(d + \sum_{i=1}^{n} \theta_i\right)\right\} \end{aligned}$$

Por lo tanto, se concluye que la distribución posterior para el hiperparámetro β es

$$\beta \mid \alpha, \boldsymbol{\theta}, \mathbf{Y} \sim Gamma\left(\alpha + c, d + \sum_{i=1}^{n} \theta_i\right)$$

Para realizar la inferencia bayesiana jerárquica para los parámetros de interés se deben fijar valores iniciales para cada parámetro y mediante simulación renovarlos hasta obtener convergencia. Por ejemplo, un posible camino para obtener convergencia en la simulación se describe a continuación:

- Fijar valores iniciales para α y β.
- Con los anteriores valores simular una observación para cada distribución posterior de los parámetros θ_i $(i = 1, \ldots, n)$.
- Con estos valores de θ_i y el valor inicial de β, simular una observación de la distribución posterior de *alpha*.
- Con los valores de θ_i y la anterior observación de α, simular un nuevo valor para β.
- Repetir el anterior proceso hasta lograr convergencia.

Dado que las distribuciones posterior de los parámetros θ_i $(i = 1, \ldots, n)$ y del hiperparámetro β están ligadas a la distribución Gamma, es posible utilizar la función `dgamma` de R directamente, con los diferentes parámetros de forma y escala, para realizar la simulación de una observación para cada distribución en cada iteración del anterior algoritmo. Sin eambrgo, como la distribución posterior marginal de α no tiene una forma cerrada, es necesario implementar un código propio en R que permita simular un valor proveniente de esta distribución. Es posible utilizar el método de la grilla que, en este caso, es univariado pues se trata de un sólo hiperparámetro. Con base en lo anterior, se tiene la siguiente función que reproduce esta distribución no conocida.

```
> post <- function(theta, alpha, beta, a, b){
```

```
+ N <- length(alpha)
+ res <- rep(NA, N)
+ for(k in 1: N){
+ P1 <- (theta^(alpha[k]-1))/gamma(alpha[k])
+ P2 <- alpha[k]^(a-1)
+ P3 <-  exp(-alpha[k]*beta)
+ res[k] <- prod(P1)*P2*P3
+ }
+ res
+ }
```

Por ejemplo, suponga que $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$ cuyas observaciones, para una iteración en particular, fueron $(2, 2, 3)$ y que la observación para β fue 0.9. De esta manera, se crea una grilla de los posibles valores que puede tomar α y mediante el buen uso de la función sample se simula un valor proveniente de esta distribución rara.

```
> # creación de la grilla para alpha
> alpha.grid <- seq(1, 100, by=0.05)
> be <- 0.9
> a1 <- 2
> b2 <- 3
> t <- c(2,2,3)

> # probabilidad para cada valor en la grilla
> post.alpha <- post(t,alpha.grid,be,a1,b1)
> N.grid <- length(post.alpha)
> post.alpha <- post.alpha/sum(post.alpha)
> sum(post.alpha)
[1] 1

> # simulación de una sola observación
> rpost <- sample(N.grid, 1, prob=post.alpha, replace=TRUE)
> r.alpha <- alpha.grid[rpost]
> r.alpha
[1] 2.05
```

Una vez que se tiene una observación simulada se renuevan los valores para continuar la siguiente iteración del algoritmo. Por otro lado, en términos exploratorios, es posible simular varios valores de la distribución no conocida y determinar qué forma tiene. Para la anterior configuración, y utilizando el siguiente código, se simularon 100 valores de esta distribución. La figura 4.1 da cuenta de la forma de esta distribución. En general, es posible afirmar que su forma es parecida a la de una distribución gamma, esto tiene sentido pues está en función de distribuciones gamma, sesgadas a la derecha y unimodales.

```
# corroborar la estructura de la cadena
```

```
N.sim <- 100

rpost <- sample(N.grid, N.sim, prob=post.alpha, replace=TRUE)
r.alpha <- alpha.grid[rpost]
mean(r.alpha)
plot(r.alpha)

# Los valores que puede tomar alpha
alpha.grid
# Las probabilidades con que toma esos valores
post.alpha
plot(post.alpha)
# tiene forma de gamma
```

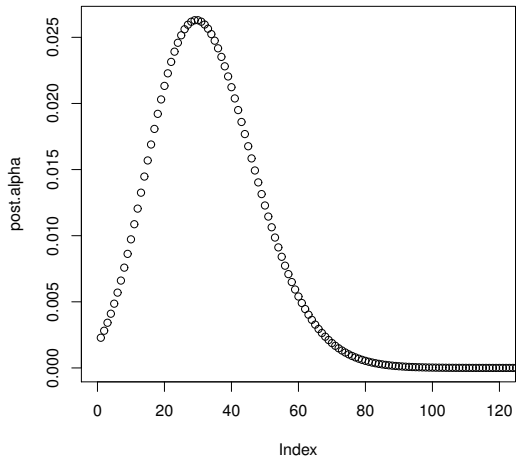

Figura 5.3: *Simulación de observaciones de la distribución posterior marginal del parámetro* α

5.2.3 Modelo Normal

Considere una variación de la estructura jerárquica del Ejemplo 4.1.2, en donde las observaciones siguen el siguiente modelo de probabilidad

$$Y_i \mid \theta_i \sim Normal(\theta_i, \sigma^2) \qquad i = 1, \ldots, n$$

y el parámetro σ^2 se supone conocido. Sin embargo, la distribución previa para los parámetros de interés θ_i es

$$\theta_i \mid \mu \sim Normal(\mu, \tau^2) \qquad i = 1, \ldots, n$$

en donde los parámetros μ y τ^2 son desconocidos.

Hiperparámetros independientes

Otra variante del análisis bayesiano jerárquico al problema de los datos con verosimilitud normal consiste en suponer que los hiperparámetros son independientes previa, es decir que su función de distribución conjunta se puede factorizar como el producto de las distribuciones marginales de cada uno de los hiperparámetros. Más aún, si se supone que las distribuciones previa marginales son no informativas y siguen una estructura probabilística uniforme, entonces se tiene que

$$p(\mu, \tau^2) = p(\mu)p(\tau^2) \propto k$$

Con esta formulación se deduce que la distribución posterior conjunta condicional a una sola observación está dada por

$$\begin{aligned} p(\theta_i, \mu, \tau^2 \mid Y_i) &\propto p(Y_i \mid \theta_i)p(\theta_i \mid \mu, \tau^2)p(\mu, \tau^2) \\ &\propto p(Y_i \mid \theta_i)p(\theta_i \mid \mu, \tau^2) \\ &\propto \exp\left\{\frac{1}{2\sigma^2}(y_i - \theta_i)^2\right\}\frac{1}{\tau}\exp\left\{\frac{1}{2\tau^2}(\theta_i - \mu)^2\right\} \end{aligned} \tag{5.2.8}$$

Y la distribución distribución posterior conjunta condicional a todas las observaciones y a todos los parámetros de interés es

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mu, \tau^2) \\ &\propto \prod_{i=1}^{n} p(Y_i \mid \theta_i) \prod_{i=1}^{n} p(\theta_i \mid \mu, \tau^2) \\ &\propto \exp\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta_i)^2\right\}\frac{1}{\tau^n}\exp\left\{\frac{1}{2\tau^2}\sum_{i=1}^{n}(\theta_i - \mu)^2\right\} \end{aligned}$$

Utilizaremos la técnica del condicionamiento para encontrar la distribución condicional del vector de parámetros de interés $\boldsymbol{\theta}$ y de los hiperparámetros. Por lo tanto se tiene que

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mu, \tau^2, \mathbf{Y}) &\propto p(\boldsymbol{\theta}, \underbrace{\mu, \tau^2}_{fijos}, \mathbf{Y}) \\ p(\mu \mid \boldsymbol{\theta}, \tau^2, \mathbf{Y}) &\propto p(\mu, \underbrace{\boldsymbol{\theta}, \tau^2}_{fijos}, \mathbf{Y}) \\ p(\tau^2 \mid \boldsymbol{\theta}, \mu, \mathbf{Y}) &\propto p(\mu, \underbrace{\boldsymbol{\theta}, \tau^2}_{fijos}, \mathbf{Y}) \end{aligned}$$

Con la anterior formulación se tiene la siguiente serie de resultados que dan cuenta de las distribuciones apropiadas para cada uno de los parámetros.

Resultado 5.2.1. *La distribución posterior del parámetro de interés θ_i es*

$$\theta_i \sim Normal(\mu_i, \tau_1^2)$$

en donde

$$\mu_i = \frac{\frac{1}{\sigma^2}Y_i + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

Prueba. Utilizando la técnica del condicionamiento posterior se tiene que

$$\begin{aligned} p(\theta_i \mid \mu, \tau^2, Y_i) &\propto p(\theta_i, \underbrace{\mu, \tau^2}_{fijos}, Y_i) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta_i)^2 - \frac{1}{2\tau^2}(\theta_i - \mu)^2\right\} \end{aligned}$$

y utilizando el mismo razonamiento que en la demostración del Resultado 2.6.1 se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_i, \tau_1^2)$. ■

Resultado 5.2.2. *La distribución posterior del hiper-parámetro μ es*

$$\mu \sim Normal(\bar{\theta}, \tau^2/n)$$

en donde $\bar{\theta} = \dfrac{1}{n}\sum_{i=1}^{n}\theta_i$.

Prueba. Utilizando la técnica del condicionamiento posterior y teniendo en cuenta que

$$\sum_{i=1}^{n}(\theta_i - \mu)^2 = \sum_{i=1}^{n}(\theta_i - \bar{\theta})^2 + n(\mu - \bar{\theta})^2$$

entonces, se tiene que

$$\begin{aligned} p(\mu \mid \boldsymbol{\theta}, \tau^2, \mathbf{Y}) &\propto p(\mu, \underbrace{\boldsymbol{\theta}, \tau^2}_{fijos}, \mathbf{Y}) \\ &\propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{n}(\theta_i - \mu)^2\right\} \propto \exp\left\{-\frac{n}{2\tau^2}(\mu - \bar{\theta})^2\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\bar{\theta}, \tau^2/n)$. ■

Resultado 5.2.3. *La distribución posterior del hiper-parámetro τ^2 es*

$$\tau^2 \sim Inversa - Gamma(n/2 - 1, nS_\mu^2/2)$$

en donde $nS_\mu^2 = \sum_{i=1}^{n}(\theta_i - \mu)^2$.

Prueba. Utilizando la técnica del condicionamiento posterior se tiene que

$$\begin{aligned} p(\tau^2 \mid \boldsymbol{\theta}, \mu, \mathbf{Y}) &\propto p(\tau^2, \underbrace{\boldsymbol{\theta}, \mu}_{fijos}, \mathbf{Y}) \\ &\propto \frac{1}{\tau^n} \exp\left\{ \frac{1}{2\tau^2} \sum_{i=1}^{n} (\theta_i - \mu)^2 \right\} \\ &\propto \left(\tau^2\right)^{-n/2} \exp\left\{ \frac{nS_\mu^2}{2\tau^2} \right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Inversa - Gamma(n/2 - 1, nS_\mu^2/2)$. ■

Utilizando un algoritmo que genere una cadena de Markov, y utilizando los anteriores resultados se realiza un análisis bayesiano propiamente dicho.

Hiperparámetros dependientes

Siguiendo el algoritmo dado al comienzo de esta sección, en dónde se dan los lineamentos generales para realizar un análisis jerárquico. En primer lugar se debe considerar la distribución posterior de los parámetros, que en este caso depende de la distribución previa de los hiperparámetros.

Suponga entonces, al igual que en capítulos anteriores, que los hiperparámetros son dependientes a una vía. Es decir, que μ depende de τ^2 pero que τ^2 no depende de μ. En estos términos, la distribución previa de los hiperparámetros está dada por

$$p(\mu, \tau^2) = p(\mu \mid \tau^2) p(\tau^2)$$

Luego, siguiendo la regla de bayes y suponiendo que los hiperparámetros son condicionalmente independientes de las observaciones dado el vector de parámetros de interés, la distribución posterior del vector de parámetros de interés $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ y de los hiperparámetros μ, τ^2 es

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mu, \tau^2) p(\mu, \tau^2) \\ &\propto p(\mu, \tau^2) \prod_{i=1}^{n} p(Y_i \mid \theta_i) \prod_{i=1}^{n} p(\theta_i \mid \mu, \tau^2) \\ &\propto p(\mu, \tau^2) \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_i)^2 \right\} \frac{1}{\tau^n} \exp\left\{ \frac{-1}{2\tau^2} \sum_{i=1}^{n} (\theta_i - \mu)^2 \right\} \end{aligned}$$

Con base en lo anterior, se tienen el siguiente resultado para el análisis bayesiano jerárquico de un sólo componente θ_i de $\boldsymbol{\theta}$.

Resultado 5.2.4. *La distribución posterior del componente θ_i perteneciente al vector de parámetros de interés $\boldsymbol{\theta}$ es*

$$\theta_i \sim Normal(\mu_i, \tau_1^2)$$

en donde

$$\mu_i = \frac{\frac{1}{\sigma^2}Y_i + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

Prueba. La prueba del resultado es inmediata al considerar la técnica del condicionamiento posterior como en la demostración del Resultado 4.2.1. puesto que

$$\begin{aligned} p(\theta_i \mid \mu, \tau^2, Y_i) &\propto p(\theta_i, \underbrace{\mu, \tau^2}_{fijos} \mid Y_i) \\ &\propto p(Y_i \mid \theta_i)p(\theta_i \mid \mu, \tau^2)p(\mu, \tau^2) \\ &\propto (Y_i \mid \theta_i)p(\theta_i \mid \mu, \tau^2) \end{aligned}$$

■

siguiendo con el algoritmo del análisis jerárquico, el siguiente paso corresponde a la determinación de la distribución posterior de los hiperparámetros μ, τ^2 la cual, suponiendo que la distribución previa conjunta para ambos hiperparámetros es uniforme y no informativa, está dada por el Resultado 4.1.1.

$$\begin{aligned} p(\mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mu, \tau^2)p(\mathbf{Y} \mid \mu, \tau^2) \\ &\propto \prod_{i=1}^{n} p(Y_i \mid \mu, \tau^2)) \\ &\propto \prod_{i=1}^{n} Normal(\mu, \tau^2 + \sigma^2) \end{aligned}$$

Ahora, por otro lado, el análisis individual de los hiperparámetros está regido por la siguiente expresión

$$p(\mu, \tau^2 \mid \mathbf{Y}) = p(\mu \mid \tau^2, \mathbf{Y})p(\tau^2 \mid \mathbf{Y})$$

En este orden de ideas, se tienen los siguientes resultados acerca de la distribución posterior para μ dada por $p(\mu \mid \tau^2, \mathbf{Y})$ y para τ^2 dada por $p(\tau^2 \mid \mathbf{Y})$

Resultado 5.2.5. *La distribución posterior del hiperparámetro mu condicionada a $\tau^2, \mathbf{Y}$ es*

$$\mu \mid \tau^2, \mathbf{Y} \sim Normal\left(\hat{\mu}, \hat{\tau}^2\right)$$

donde $\hat{\mu} = \bar{Y}$ y $n\hat{\tau}^2 = \sigma^2 + \tau^2$.

Prueba. Utilizando la técnica del condicionamiento posterior, nótese que la distribución posterior de *mu* toma la siguiente forma

$$p(\mu \mid \tau^2, \mathbf{Y}) \propto p(\mu, \underbrace{\tau^2}_{fijo} \mid \mathbf{Y})$$
$$\propto \prod_{i=1}^{n} Normal(\mu, \tau^2 + \sigma^2)$$

Partiendo de este hecho, es fácil confirmar que

$$p(\mu \mid \tau^2, \mathbf{Y}) \propto \qquad \propto \exp\left\{\frac{1}{2(\sigma^2+\tau^2)}\sum_{i=1}^{n}(y_i-\mu)^2\right\}$$
$$= \exp\left\{\frac{1}{2(\sigma^2+\tau^2)}\sum_{i=1}^{n}(y_i^2 - 2\mu Y_i + \mu^2)\right\}$$
$$\propto \exp\left\{\frac{n}{2(\sigma^2+\tau^2)}(\mu^2 - 2\mu\bar{Y})\right\}$$
$$\propto \exp\left\{\frac{n}{2(\sigma^2+\tau^2)}(\mu-\bar{Y})^2\right\}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\hat{\mu}, \hat{\tau}^2)$. ■

Resultado 5.2.6. *La distribución posterior del hiperparámetro τ es*

$$p(\tau^2 \mid \mathbf{Y}) \propto \sqrt{\hat{\tau}} \prod_{i=1}^{n} (\sigma^2+\tau^2)^{-1/2} \exp\left\{-\frac{1}{2(\sigma^2+\tau^2)}(y_i-\hat{\mu})^2\right\}$$

Prueba. En primer lugar, nótese que

$$p(\tau \mid \mathbf{Y}) = \frac{p(\mu, \tau^2 \mid \mathbf{Y})}{p(\mu \mid \tau^2, \mathbf{Y})} \qquad \forall \mu$$
$$\propto \frac{\prod_{i=1}^{n} Normal(\mu, \sigma^2+\tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \qquad \forall \mu$$

La anterior igualdad debe mantenerse para cualquier valor de μ; en particular

se debe mantener para $\mu = \hat{\mu}$ (**?**). Por tanto,

$$\begin{aligned} p(\tau \mid \mathbf{Y}) &\propto \frac{Normal(\hat{\mu}, \sigma^2 + \tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \\ &\propto \frac{\prod_{i=1}^{n} Normal(\hat{\mu}, \sigma^2 + \tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \\ &\propto \sqrt{\hat{\tau}} \prod_{i=1}^{n} (\sigma^2 + \tau^2)^{-1/2} \exp\left\{ -\frac{1}{2(\sigma^2 + \tau^2)} (y_i - \hat{\mu})^2 \right\} \exp\left\{ \frac{1}{2\hat{\tau}^2} (\hat{\mu} - \hat{\mu})^2 \right\} \\ &\propto \sqrt{\hat{\tau}} \prod_{i=1}^{n} (\sigma^2 + \tau^2)^{-1/2} \exp\left\{ -\frac{1}{2(\sigma^2 + \tau^2)} (y_i - \hat{\mu})^2 \right\} \end{aligned}$$

■

En términos de simulación, los anteriores resultados garantizan una estructura formal que permita simular la distribución posterior del hiperparámetro τ^2, y mediante esta encontrar una estimación para reemplazarla en la distribución posterior del hiperparámetro μ y repetir el proceso anterior. Con estos valores bien definidos, entonces utilizar el Resultado 4.2.4 para proseguir con el análisis bayesiano clásico.

6 Modelo lineal Bayesiano

En términos de modelamiento estadístico, la relación entre una variable de interés Y y una matriz de variables auxiliares $\mathbf{X}$, es una de las herramientas estadísticas más utilizadas por los investigadores en los últimos tiempos. Herramientas como la regresión simple, la regresión múltiple y el análisis de varianza forman parte del arsenal de opciones que la ciencia estadística ofrece a los usuarios que van un paso más allá estableciendo relaciones de causalidad en el contexto de propio de la investigación.

Por supuesto, el enfoque bayesiano también ofrece al investigador herramientas poderosas en términos del modelamiento de relaciones causales. Siguiendo el mismo espíritu que en los anteriores capítulos, la inferencia bayesiana proporciona distribuciones posterior para los parámetros de interés y distribuciones predictivas para nuevas observaciones en cada uno de los contextos mencionados anteriormente. Como lo menciona (**?**), es muy útil adoptar la notación matricial para el desarrollo posterior del análisis bayesiano; entonces, se definen

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{y} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nq} \end{pmatrix}$$

y se supone que existe una relación de causalidad de parte de $\mathbf{X}$ reflejada en $\mathbf{Y}$ que puede ser descrita mediante el siguiente modelo probabilístico

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6.0.1}$$

en donde $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ es un vector aleatorio tal que cada una de sus componentes sigue una distribución de probabilidad, que en la mayoría de los casos suele ser normal. Antes de comenzar con la estipulación propia del análisis bayesiano, es necesario aclarar el papel que juegan las variables auxiliares en la inferencia estadística.

En primer lugar, nótese que el interés particular recae en la distribución del vector de n variables aleatorias $\mathbf{Y} = (Y_1 \dots, Y_n)'$ condicional a la matriz de variables auxiliares $\mathbf{X}$ e indexada por el vector de parámetros de interés $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$ dada por $p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X})$.

Basado en lo anterior, es posible y suponiendo que las variables de interés son intercambiables, entonces es posible plantear el siguiente modelo poblacional

$$\begin{aligned} E(Y_i \mid \boldsymbol{\beta}, \mathbf{X}) &= \boldsymbol{\beta}\mathbf{x}_i' = \beta_1 x_{i1} + \cdots + \beta_q x_{iq} \\ Var(Y_i \mid \boldsymbol{\beta}, \mathbf{X}) &= \sigma^2 \qquad\qquad i = 1, \dots, n \end{aligned}$$

en donde generalmente $\boldsymbol{\beta}$ y σ^2 son los parámetros de interés. (**?**) afirma que en la realidad, las observaciones incluyen realizaciones tanto de las variables de interés Y_i como de las variables auxiliares $\mathbf{X}$ y por la anterior razón, el modelamiento bayesiano propiamente dicho debería incluir

a Una distribución para las variables auxiliares $\mathbf{X}$ indexada por un vector de hiperparámetros $\boldsymbol{\phi}$, dada por $p(\mathbf{X} \mid \boldsymbol{\phi})$.

b Una verosimilitud conjunta de los datos observados dada por $p(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\phi}, \boldsymbol{\theta})$, en donde $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)$ es el vector de parámetros de interés.

c Por último, una distribución previa conjunta para los parámetros desconocidos dada por $p(\boldsymbol{\phi}, \boldsymbol{\theta})$.

Sin embargo, en un contexto estándar, los anteriores requerimientos se simplifican al suponer que los parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\phi}$ son independientes previa - es decir $p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\theta})p(\boldsymbol{\phi})$ - y que la distribución de las variables auxiliares es no informativa al igual que la distribución previa del parámetro $\boldsymbol{\phi}$ - es decir que $p(\boldsymbol{\phi} \mid \mathbf{X}) \propto k$. De esta manera, y recurriendo al Resultado 1.1.3, se tiene que

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{Y}, \mathbf{X}) &= p(\boldsymbol{\phi} \mid \mathbf{X}) p(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X}) \\ &\propto p(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{X}) \\ &\propto p(\boldsymbol{\theta}) p(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X}) \end{aligned}$$

Por ende, de aquí en adelante vamos a referirnos a la distribución posterior de $\boldsymbol{\beta}$ comprendiendo que la especificación de esta distribución cubre todo el ámbito probabilístico de las observaciones de las variables auxiliares.

El modelo básico y clásico asume que la verosimilitud para las variables de interés es

$$\mathbf{Y} \mid \boldsymbol{\theta}, \sigma^2, \mathbf{X} \sim Normal_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

en donde $\mathbf{I}_n$ denota la matriz identidad de orden $n \times n$. Por supuesto, el modelo normal no es el único que se puede postular como verosimilitud para los datos. Existen muchos más distribuciones que serán contempladas más adelante.

6.1 Modelo lineal con varianza conocida

En términos generales, la verosimilitud del vector de variables de interés está dada por la siguiente expresión

$$p(\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{X}) \propto \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \tag{6.1.1}$$

en donde $\boldsymbol{\Sigma}$ representa la matriz de varianzas de $\mathbf{Y}$, simétrica y definida positiva. Se tiene que cada una de las entradas de la matriz de varianzas es conocida. Suponga que la distribución previa del vector de parámetros de interés es informativa

y además está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

Nótese que es natural asignarle a $\boldsymbol{\beta}$ una distribución normal multivariante pues cada uno de sus componentes describe una relación numérica de la variable de interés con la correspondiente variable de información auxiliar. Bajo este contexto se tiene el siguiente resultado.

Resultado 6.1.1. *La distribución posterior para el vector de parámetros de interés es*

$$\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\Sigma} \sim Normal_q(\mathbf{b}_q, \mathbf{B}_q)$$

donde

$$\begin{aligned} \mathbf{B}_q &= \left(\mathbf{B}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1} \\ \mathbf{b}_q &= \mathbf{B}_q\left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}\right) \end{aligned}$$

Prueba. De la definición de distribución posterior, y completando cuadrados como en la demostración del Resultado 3.2.1., se tiene que

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\Sigma}) &\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta}-\mathbf{b})\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}+\mathbf{B}^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}+\mathbf{B}^{-1}\mathbf{b})\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\boldsymbol{\beta}'\mathbf{B}_q^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{B}_q^{-1}\mathbf{b}_q\right\} \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\beta}'\mathbf{B}_q^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{B}_q^{-1}\mathbf{b}_q - \frac{1}{2}\mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q\right\} \\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta}-\mathbf{b}_q)\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una vector aleatorio con distribución $Normal_q(\mathbf{b}_q, \mathbf{B}_q)$. ■

Contextualizado en el modelo lineal clásico, en donde la varianza de las observaciones es constante e igual a σ^2 y se supone que no existe correlación entre ellas, entonces la verosimilitud del vector de variables de interés está dada por la siguiente expresión

$$p(\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \tag{6.1.2}$$

en donde el parámetro σ^2 es conocido. La distribución posterior para el vector de parámetros de interés $\boldsymbol{\beta}$ sería

$$\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \sigma^2 \sim Normal_q(\mathbf{b}_q, \mathbf{B}_q)$$

donde

$$\mathbf{B}_q = \left(\mathbf{B}^{-1} + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)^{-1}$$
$$\mathbf{b}_q = \mathbf{B}_q\left(\mathbf{B}^{-1}\mathbf{b} + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{Y}\right)$$

Nótese que si las entradas diagonales de la matriz de covarianzas $\mathbf{B}$ en la distribución previa para $\boldsymbol{\beta}$ es muy grande, entonces $\mathbf{B}^{-1} \longrightarrow \mathbf{0}_{q\times q}$ y por lo tanto, se tiene que

$$\mathbf{B}_q \longrightarrow \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$
$$\mathbf{b}_q \longrightarrow (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

y las anteriores expresiones que coinciden con los estimadores del modelo lineal clásico usando técnicas frecuentistas como la estimación mediante la técnica de los mínimos cuadrados o por él método de máxima verosimilitud.

Ejemplo 6.1.1. *Poner los datos del kidney score de Efron y hacer la regresión contra la edad y ponerle intercepto*

6.2 Modelo lineal con varianza desconocida

En muy raras ocasiones se conoce la estructura de dispersión de las variables de interés y ese desconocimiento de la variabilidad de las observaciones es la regla más que la excepción en una gran cantidad de situaciones de la vida real. En este contexto, es necesario modelar conjuntamente, tanto la matriz de varianzas como el vector de parámetros de interés que moldean la relación de causalidad.

6.2.1 previas no informativas

Para empezar, y siguiendo los supuestos básicos del modelo lineal general, suponga que las variables aleatorias de interés conforman una muestra aleatoria en donde no existe ninguna estructura de correlación y el parámetro de variabilidad es constante a través de todos los individuos de la muestra. de esta manera, la verosimilitud conjunta estará dada por la expresión (5.1.2) en donde, una vez más el parámetro σ^2 es desconocido.

En el caso más sencillo, se supone que la distribución previa conjunta de los parámetros de interés puede ser factorizada de la siguiente manera.

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta} \mid \sigma^2)p(\sigma^2)$$

y asignando a cada uno de los parámetros de interés distribuciones previa no informativas, entonces, al igual que en capítulos anteriores, se concluye que la distribución previa no informativa para $\boldsymbol{\beta} \mid \sigma^2$ puede ser uniforme y constante tal

que $p(\boldsymbol{\beta} \mid \sigma^2) \propto k$ y la de σ^2 tal que $p(\sigma^2) \propto 1/\sigma^2$. Basado en lo anterior, es factible asignar la siguiente distribución previa conjunta

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}) \propto \frac{1}{\sigma^2}$$

En este orden de ideas, es fácil comprobara que la distribución posterior conjunta de los parámetros de interés está dada por

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X}) &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}) \\ &\propto \left(\sigma^2\right)^{-n/2-1} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \end{aligned} \tag{6.2.1}$$

Basados en la anterior distribución conjunta, se tienen los siguientes dos resultados que dan cuenta de las distribuciones marginales posterior para los parámetros de interés.

Resultado 6.2.1. *La distribución posterior del vector de parámetros $\boldsymbol{\beta}$ condicionado a $\sigma^2, \mathbf{Y}, \mathbf{X}$ es*

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X} \sim Normal_q(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

en donde $\hat{\boldsymbol{\beta}}$ denota el vector de estimadores frecuentistas clásicos obtenidos con el método de mínimos cuadrados, dado por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y} \tag{6.2.2}$$

Prueba. Utilizando la técnica del condicionamiento posterior, se tiene que de la expresión

$$\begin{aligned} p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X}) &\propto p(\boldsymbol{\beta}, \underbrace{\sigma^2}_{fijo} \mid \mathbf{Y}, \mathbf{X}) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{X}'\mathbf{X}\mathbf{y}-\mathbf{y}'\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una vector aleatorio con distribución $Normal_q(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. ■

Resultado 6.2.2. *La distribución posterior del parámetro σ^2 es*

$$\sigma^2 \mid \mathbf{Y}, \mathbf{X} \sim Gamma - inversa\left(\frac{n-q}{2}, \frac{S_e^2}{2}\right)$$

en donde $S_e^2 = (\mathbf{Y} - \mathbf{X}'\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}'\hat{\boldsymbol{\beta}})$ *denota la suma de cuadrados de los errores del modelo ajustado.*

Prueba. Para encontrar la distribución posterior de σ^2 se utiliza la siguiente expresión, en virtud del conocimiento de la verosimilitud y la distribución posterior condicional de $\boldsymbol{\beta}$,

$$\begin{aligned} P(\sigma^2 \mid \mathbf{Y}, \mathbf{X}) &= \frac{p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X})}{p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X})} \\ &\propto \left(\sigma^2\right)^{q/2-n/2-1} \exp\left\{-\frac{1}{2\sigma^2}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})]\right\} \\ &\propto \left(\sigma^2\right)^{-(n-q)/2-1} \exp\left\{-\frac{1}{2\sigma^2} S_e^2\right\} \end{aligned}$$

Lo anterior se tiene, puesto que, teniendo en cuenta que

$$\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \tag{6.2.3}$$

después de un simple desarrollo algebraico, se encuentra que

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

que coincide con

$$S_e^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma - inversa\left(\frac{n-q}{2}, \frac{S_e^2}{2}\right)$. ■

Al igual que en los capítulos anteriores, la simulación para este tipo de especificaciones debe tener en cuenta en primer lugar la simulación de la distribución $p(\sigma^2 \mid \mathbf{Y}, \mathbf{X})$ y encontrar un valor estimado para este parámetro. Luego, se debe utilizar este valor para simular la distribución $p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X})$ e igualmente, encontrar un valor estimado para este parámetro.

6.2.2 previas informativas

Esta sección tiene dos acepciones: la primera en donde la distribución de los parámetros de interés puede ser descompuesta como el producto dependiente de dos distribuciones, y la otro, por supuesto, cuando se considera que los parámetros de interés son independientes previa. Sin embargo, para cade uno de los dos casos, es necesario reescribir la verosimilitud de las observaciones para poder obtener resultados conjugados.

En primer lugar, se definen las siguientes cantidades, las cuales se habían utilizado indirectamente en secciones anteriores pero que serán de interés para el

tratamiento riguroso de las distribuciones posterior en este apartado. Luego, se define la *suma de cuadrados del error de estimación de los parámetros de interés* ponderado por las variables de información auxiliar como

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \tag{6.2.4}$$

donde $\hat{\boldsymbol{\beta}}$ está dado por la expresión (5.2.2). Por otro lado, se define la *suma de cuadrados del error de predicción* de las observaciones bajo el modelo lineal general dado por la siguiente expresión

$$S_e^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{6.2.5}$$

Con las anteriores expresiones es posible plantear la siguiente identidad que fue utilizada en la demostración del Resultado 5.2.3.

$$Q(\boldsymbol{\beta}) + S_e^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{6.2.6}$$

Nótese que con esta expresión es posible reescribir la verosimilitud de las observaciones dada por la expresión (5.2.1) de la siguiente manera

$$p(\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(Q(\boldsymbol{\beta}) + S_e^2\right)\right\} \tag{6.2.7}$$

Antes de plantear las distirbuciones previa para los parámetros de interés, nótese que a través de este texto, cuando se trata de modelos con múltiples parámetros de interés, siempre se han planteado dos grandes posibilidades: que los parámetros sean dependientes previa y por supuesto, que los parámetros sean independientes previa. El tratamiento de ambos escenarios es diferente y vamos a estudiar cada caso con detenimiento.

Parámetros dependientes

Los parámetros de interés son $\boldsymbol{\beta}$ y σ^2 y su distribuciones previa conjunta se supone que está dada por

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta} \mid \sigma^2)p(\sigma^2)$$

Específicamente, la distribución previa del parámetro $\boldsymbol{\beta}$ condicionada a σ^2 es informativa y está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \mid \sigma^2 \sim Normal_q(\mathbf{b}, \sigma^2\mathbf{B})$$

en donde $\mathbf{b}$ es un vector de medias y $\mathbf{B}$ es una matriz de varianzas simétrica y definida positiva. Por otro lado, la distribución previa del parámetro σ^2 también se considera informativa y dada por

$$\sigma^2 \sim Inversa - Gamma\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$$

Bajo este marco de referencia se tienen los siguientes resultados.

Resultado 6.2.3. *La distribución posterior conjunta de los parámetros de interés* $\boldsymbol{\beta}, \sigma^2$ *está dada por*

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X}) = (\sigma^2)^{-q/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta} - \mathbf{b}_q)\right\} \times (\sigma^2)^{-n_1/2-1} \exp\left\{-\frac{n_1\sigma_1^2}{2\sigma^2}\right\} \tag{6.2.8}$$

donde

$$\mathbf{B}_q = \left(\mathbf{B}^{-1} + \mathbf{X}'\mathbf{X}\right)^{-1}$$
$$\mathbf{b}_q = \mathbf{B}_q\left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\mathbf{Y}\right)$$

y además

$$n_1 = n_0 + n$$
$$n_1\sigma_1^2 = n_0\sigma_0^2 + (\mathbf{Y} - \mathbf{X}\mathbf{b}_q)'\mathbf{Y} + (\mathbf{b} - \mathbf{b}_q)'\mathbf{B}^{-1}\mathbf{b}$$

Prueba. Antes de empezar con la prueba formal del resultado, es necesario verificar las siguientes identidades

$$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b}) + Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta} - \mathbf{b}_q) + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q \tag{6.2.9}$$

que se tiene puesto que

$$\begin{aligned}
&(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b}) + Q(\boldsymbol{\beta}) \\
&= \boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{B}^{-1}\mathbf{b} + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \boldsymbol{\beta}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{\beta}'(\mathbf{B}^{-1} + \mathbf{X}'\mathbf{X})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{\beta}'(\mathbf{B}_q^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\mathbf{Y}) + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{\beta}'(\mathbf{B}_q^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{B}_q^{-1}b_q + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{\beta}'(\mathbf{B}_q^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{B}_q^{-1}b_q + \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q\mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q \\
&= (\boldsymbol{\beta} - \mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta} - \mathbf{b}_q) + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q
\end{aligned}$$

Por otro lado, debe notarse que

$$n_0\sigma_0^2 + S_e^2 + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q = n_1\sigma_1^2 \tag{6.2.10}$$

La anterior igualdad resulta de

$$\begin{aligned}
&n_0\sigma_0^2 + S_e^2 + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q \\
&= n_0\sigma_0^2 + \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} \\
&\qquad + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{B}_q(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\mathbf{Y}) \\
&= n_0\sigma_0^2 + \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} \\
&\qquad + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{b}_q'(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\mathbf{Y}) \\
&= n_0\sigma_0^2 + \mathbf{Y}'\mathbf{Y} - \mathbf{b}_q'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b} - \mathbf{b}_q'\mathbf{B}^{-1}\mathbf{b} \\
&= n_0\sigma_0^2 + (\mathbf{Y} - \mathbf{X}\mathbf{b}_q)'\mathbf{Y} + (\mathbf{b}' - \mathbf{b}_q'\mathbf{B}^{-1})\mathbf{b} \\
&= n_1\sigma_1^2
\end{aligned}$$

Utilizando las expresiones (5.2.9) y (5.2.10) logra concluirse que

$$\begin{aligned}
Q(\boldsymbol{\beta}) + S_e^2 + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b}) + n_0\sigma_0^2 \\
= (\boldsymbol{\beta} - \mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta} - \mathbf{b}_q) + n_1\sigma_1^2
\end{aligned} \tag{6.2.11}$$

Después de haber probado las anteriores igualdades, el desarrollo de la prueba es más sencillo. Puesto que ahora, de la definición de distribución posterior, al utilizar la identidad dada por la expresión (5.2.11), se tiene que

$$\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X}) &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}) p(\boldsymbol{\beta}, \sigma^2) \\
&\propto \left(\sigma^2\right)^{-\frac{q+n+n_0}{2}-1} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[Q(\boldsymbol{\beta}) + S_e^2 + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b}) + n_0\sigma_0^2\right]\right\} \\
&= \left(\sigma^2\right)^{-\frac{q+n+n_0}{2}-1} \\
&\qquad \times \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta} - \mathbf{b}_q) + n_1\sigma_1^2\right]\right\} \\
&= \left(\sigma^2\right)^{-\frac{q}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta} - \mathbf{b}_q)\right]\right\} \\
&\qquad \times \left(\sigma^2\right)^{-\frac{n_1}{2}-1} \exp\left\{-\frac{n_1}{2\sigma^2}\sigma_1^2\right\}
\end{aligned}$$

con lo que se concluye la prueba del resultado. ■

La distribución posterior conjunta de los parámetros de interés tiene la forma de la distribución Normal-Gamma y es fácil demostrar que ésta es conjugada. Además, acudiendo a las técnicas bien conocidas de condicionamiento posterior e integración analítica se encuentran las distribuciones posterior de cada una de los parámetros de interés enmarcadas en los siguientes resultados.

Resultado 6.2.4. *La distribución posterior del vector de parámetros $\boldsymbol{\beta}$ condicionada a $\sigma^2, \mathbf{Y}, \mathbf{X}$ es*

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X} \sim Normal_q(\mathbf{b}_q, \sigma^2 \mathbf{B}_q)$$

Prueba. Para encontrar la distribución posterior del vector de parámetros *bbeta*, el cual depende previa del parámetro σ^2, es necesario recurrir al condicionamiento posterior notando que

$$p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X}) = p(\boldsymbol{\beta}, \underbrace{\sigma^2}_{fijo} \mid \mathbf{Y}, \mathbf{X})$$

Por lo tanto, de la distribución posterior conjunta dada por (5.2.8) e incorporando los términos que no depende de $\boldsymbol{\beta}$ en la constante de proporcionalidad, se tiene fácilmente que

$$p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X}) \propto \left(\sigma^2\right)^{-\frac{q}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \mathbf{b}_q)' \mathbf{B}_q^{-1} (\boldsymbol{\beta} - \mathbf{b}_q)\right]\right\}$$

De la anterior igualdad, y factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una vector aleatorio con distribución $Normal_q(\mathbf{b}_q, \sigma^2 \mathbf{B}_q)$. ■

Resultado 6.2.5. *La distribución posterior del parámetro σ^2 condicionada es*

$$\sigma^2 \mid \mathbf{Y}, \mathbf{X} \sim Inversa - Gamma\left(\frac{n_1}{2}, \frac{\sigma_1^2}{2}\right)$$

Prueba. En este caso es posible servirse de la técnica de integración analítica de la siguiente manera

$$\begin{aligned} p(\sigma^2 \mid \mathbf{Y}, \mathbf{X}) &= \int p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X})\, d\boldsymbol{\beta} \\ &\propto \left(\sigma^2\right)^{-\frac{n_1}{2}-1} \exp\left\{-\frac{n_1}{2\sigma^2}\sigma_1^2\right\} \\ &\qquad \times \int \left(2\pi\sigma^2\right)^{-\frac{q}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \mathbf{b}_q)' \mathbf{B}_q^{-1} (\boldsymbol{\beta} - \mathbf{b}_q)\right]\right\} d\boldsymbol{\beta} \\ &= \left(\sigma^2\right)^{-\frac{n_1}{2}-1} \exp\left\{-\frac{n_1}{2\sigma^2}\sigma_1^2\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Inversa - Gamma\left(\frac{n_1}{2}, \frac{\sigma_1^2}{2}\right)$. ■

Parámetros dependientes

(**?**) menciona que existe una mayor dificultad para encontrar distribuciones conjugadas conforme el modelo se torna más complejo y la dimensión del espacio de parámetros crece. En esta ocasión se considera que los parámetros son independientes previa; es decir que la distribución previa conjunta está dada por

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2)$$

Como es natural, la distribución previa del vector de parámetros $\boldsymbol{\beta}$ es normal, aunque esta vez la matriz de varianzas no va a depender del otro parámetro σ^2, por lo tanto se tiene que

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

Igualmente, el parámetro σ^2 no depende de $\boldsymbol{\beta}$ y es posible asignarle la siguiente distribución previa

$$\sigma^2 \sim Inversa - Gamma\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$$

Empleando la forma de la verosimilitud dada en la expresión (5.2.7), fácilmente se concluye que la distribución posterior conjunta puede ser escrita como

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{Y}, \mathbf{X}) &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta})p(\sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(Q(\boldsymbol{\beta}) + S_e^2\right)\right\} \\ &\times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b})\right\}(\sigma^2)^{-n_0/2-1} \exp\left\{-\frac{n_0\sigma_0^2}{2\sigma^2}\right\} \\ &= (\sigma^2)^{-\frac{n+n_0}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[Q(\boldsymbol{\beta}) + S_e^2 + n_0\sigma_0^2\right]\right\} \\ &\times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b})\right\} \end{aligned} \tag{6.2.12}$$

(**?**) concluye que no es posible reconocer la anterior distribución posterior como perteneciente a alguna forma analítica conocida y por tanto no es una distribución conjugada (aunque la distribución previa sea un caso particular de esta última distribución).

Una vez más, dado esto, no es posible utilizar la técnica de integración analítica para encontrar las distribuciones posterior marginales de los parámetros de interés. Sin embargo, a pesar de las anteriores razones, sí es posible encontrar fácilmente tales distribuciones marginales utilizando el condicionamiento posterior.

Resultado 6.2.6. *La distribución posterior del parámetro $\boldsymbol{\beta}$ condicionado a $\sigma^2, \mathbf{Y}, \mathbf{X}$ es*

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X} \sim Normal_q(\mathbf{b}_q, \mathbf{B}_q)$$

donde

$$\mathbf{B}_q = \left(\mathbf{B}^{-1} + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)^{-1}$$
$$\mathbf{b}_q = \mathbf{B}_q\left(\mathbf{B}^{-1}\mathbf{b} + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{Y}\right)$$

Prueba. Utilizando el condicionamiento posterior en la expresión (5.2.12), se tiene que

$$\begin{aligned}
p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}, \mathbf{X}) &\propto p(\boldsymbol{\beta}, \underbrace{\sigma^2}_{fijo} \mid \mathbf{Y}, \mathbf{X}) \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}Q(\boldsymbol{\beta}) + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta}-\mathbf{b})\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta}-\mathbf{b})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \frac{2}{\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}'\mathbf{B}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{b}\mathbf{B}^{-1}\right]\right\} \\
&=\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \mathbf{B}^{-1}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{Y} + \mathbf{B}^{-1}\mathbf{b}\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}'\mathbf{B}_q^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{B}_q^{-1}\mathbf{b}_q\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}'\mathbf{B}_q^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{B}_q^{-1}\mathbf{b}_q + \mathbf{b}_q'\mathbf{B}_q^{-1}\mathbf{b}_q\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta}-\mathbf{b}_q)\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de un vector aleatorio con distribución $Normal_q(\mathbf{b}_q, \mathbf{B}_q)$. ■

Resultado 6.2.7. *La distribución posterior del parámetro σ^2 condicionado a $\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}$ es*

$$\sigma^2 \mid \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \sim Inversa - Gamma\left(\frac{n_1}{2}, \frac{n_1\sigma^2_{\boldsymbol{\beta}}}{2}\right)$$

donde $n_1 = n + n_0$

$$n_1\sigma^2_{\boldsymbol{\beta}} = Q(\boldsymbol{\beta}) + S_e^2 + n_0\sigma_0^2$$

Prueba. Utilizando el condicionamiento posterior sobre la expresión (5.2.12), se

tiene que

$$\begin{aligned} p(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}) &\propto p(\sigma^2, \underbrace{\boldsymbol{\beta}}_{fijo} \mid \mathbf{Y}, \mathbf{X}) \\ &\propto (\sigma^2)^{-\frac{n+n_0}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[Q(\boldsymbol{\beta}) + S_e^2 + n_0\sigma_0^2\right]\right\} \\ &\propto (\sigma^2)^{-\frac{n_1}{2}-1} \exp\left\{-\frac{n_1\sigma^2_{\boldsymbol{\beta}}}{2\sigma^2}\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de un vector aleatorio con distribución $Inversa - Gamma\left(\frac{n_1}{2}, \frac{n_1\sigma^2_{\boldsymbol{\beta}}}{2}\right)$. ■

(**?**) afirma que las anteriores especificaciones tienen una altísima relevancia práctica y es concebible fijar un conjunto de valores iniciales e ir actualizando las estimaciones acerca de los parámetros de interés.

Ejemplo 6.2.1. ?, *p. 100 proponen una situación de valorización de predios. Se tienen $n = 389$ observaciones de precios de diferentes predios en Chicago (y sus áreas metropolitanas). Se quiere crear un modelo lineal bayesiano donde las variables predictoras son X_1, la distancia del predio al lago Michigan; X_2, la distancia del predio al aeropuerto de Midway; X_3, la distancia del predio al aeropuerto de O'Hare. Por otro lado la variable respuesta es Y, los valores log-transformados del precio del predio. El modelo es el siguiente:*

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \varepsilon_i \qquad \varepsilon_i \sim Normal(0, \sigma^2)$$

*Donde Z_{ij} son las variables estandarizadas construidas con X_{ij}. Considerando distribuciones previa en el sentido vague para los cuatro coeficientes de regresión. Para la varianza seguiremos la sugerencia de (**?**) consiste en asignar una distribución previa uniforme acotando el límite superior con un valor alto y el límite inferior con un valor positivo cercano a cero. La sintaxis en BUGS para ajustar este modelo está dada por el siguiente comando computacional, que considera la estandarización de los valores de las variables explicativas.*

```
model{for(j in 1:p){
b[j]<-beta[j]/sd(x[,j])
for(i in 1:n){
z[i,j]<-(x[i,j]-mean(x[,j]))/sd(x[,j])
}
}

for(i in 1:n){
y[i]~dnorm(mu[i],tau)
```

```
mu[i]<-beta0+beta[1]*z[i,1]+beta[2]*z[i,2]+beta[3]*z[i,3]
}
beta0~dflat()
for(j in 1:p){beta[j]~dflat()}
sigma~dunif(0.01,100)
tau<-1/(sigma*sigma)
}
```

Para asegurar la convergencia, se corren cuatro distintas cadenas con cinco distintos conjuntos de valores iniciales para los coeficientes de regresión y el parámetro de precisión. Los valores iniciales para los parámetros están dados en la siguiente estructura computacional de BUGS.

```
list(beta0=10, beta=c(0,0,0), sigma=0.1)
list(beta0=-10, beta=c(10,10,10), sigma=10)
list(beta0=0, beta=c(0,0,0), sigma=10)
list(beta0=-10, beta=c(-10,-10,-10), sigma=100)
```

Después de un periodo de calentamiento de las cadenas comprendido por 1000 iteraciones, se realizaron 50 mil iteraciones más para obtener la distribución posterior en cada uno de los cinco casos: 4 coeficientes de regresión y la raíz inversa de la varianza o precisión. Es decir, en los cálculos de simulación de la densidad posterior sólo intervinieron los valores simulados desde la iteración 1001 hasta la iteración 50000. A continuación, se reproducen las salidas para los coeficientes betas y para sigma.

node	mean	sd	MCerror	2.5%	median	97.5%
beta0	5.5800	0.037	1,82E-01	5.506	5.579	5.653
beta[1]	-0.317	0.037	1,70E-01	-0.39	-0.317	-0.24
beta[2]	0.3553	0.043	1,75E-01	0.271	0.3554	0.440
beta[3]	0.1502	0.043	1,86E-01	0.064	0.1505	0.234
sigma	0.7399	0.026	1,19E-01	0.689	0.7390	0.794

La figura 5.1 muestra el comportamiento de las cadenas para los cuatro coeficientes de regresión. Como conclusión, podemos afirmar que la distribución posterior para los parámetros de interés no depende de los valores iniciales en la cadena, por lo tanto la convergencia se garantiza y así mismo las conclusiones obtenidas a partir de este análisis.

Bajo el criterio de mínima perdida cuadrática, los estimadores bayesianos para los coeficientes de regresión y para el parámetro de precisión y para la varianza son

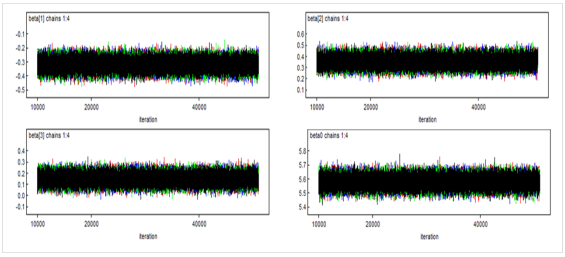

Figura 6.1: *Rápida convergencia de las cadenas.*

$$\begin{aligned}\hat{\beta}_0 &= 5.580\\ \hat{\beta}_1 &= -0.317\\ \hat{\beta}_2 &= 0.355\\ \hat{\beta}_3 &= 0.150\\ \hat{\sigma} &= 0.739^2\end{aligned}$$

Como conclusión, y respondiendo a las preguntas planteadas, el valor promedia (log-transformado) de la tierra dentro y cerca de Chicago está cercano a los 5.58. Sin embargo, este valor aumenta para las propiedades cercanas al lago Michigan y para las propiedades alejadas de los dos aeropuertos. Cada unidad de distancia más cerca al lago Michigan implica un aumento de 0.32 en el valor de la propiedad. Similarmente, cada unidad más lejana del aeropuerto Midway implica un aumento de 0.35 en el valor de la propiedad y una unidad de distancia más lejana al aeropuerto de O'Hare implica un aumento de 0.15 en el valor de la propiedad. Lo anterior tiene sentido pues la zona cercana al lago Michigan está rodeado de lugares culturales dentro de Chicago y está circundada por la zona de mayor actividad económica en chicago: el loop. Por otro lado, los aeropuertos están muy alejados de la zona del lago Michigan y por consiguiente del centro de Chicago. Esto implica que las propiedades cercanas a estos aeropuertos se desvalorizan.

6.3 Modelo lineal con varianzas desiguales

(**?**) señala que las varianzas desiguales y los errores correlacionados pueden ser incluidos en el modelo lineal bayesiano mediante la incorporación de una matriz de varianzas $\mathbf{\Sigma}$ - no es necesariamente proporcional a la matriz identidad - en la verosimilitud de los datos dada por

$$Y \mid \boldsymbol{\beta}, \mathbf{\Sigma} \sim Normal_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma})$$

Al igual que en la sección anterior, se considera que los parámetros son independientes previa y que, la distribución previa del vector de parámetros $\boldsymbol{\beta}$ es normal, la cual no depende de $\mathbf{\Sigma}$ y tiene su propia estructura de varianza, por lo tanto se tiene que

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

Igualmente, la matriz de parámetros de dispersión $\mathbf{\Sigma}$ no depende de $\boldsymbol{\beta}$ y es posible asignarle la siguiente distribución previa

$$\mathbf{\Sigma} \sim Inversa - Wishart_v(\mathbf{\Lambda})$$

Nótese que la cantidad de parámetros individuales que se deben modelar crece a medida que el tamaño de muestra crece. Por otro lado, para encontrar las distribuciones posterior que definan la estructura probabilística posterior, es necesario utilizar la técnica del condicionamiento posterior notando que

$$\begin{aligned} p(\mathbf{Y}, \boldsymbol{\beta}, \mathbf{\Sigma}) &= p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{\Sigma}) p(\boldsymbol{\beta}, \mathbf{\Sigma}) && (6.3.1) \\ &= p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{\Sigma}) p(\boldsymbol{\beta}) p\mathbf{\Sigma}) && (6.3.2) \end{aligned}$$

y para encontrar las distribuciones posterior, se tiene que

$$p(\boldsymbol{\beta} \mid Y, \mathbf{\Sigma}) \propto p(\boldsymbol{\beta}, \mathbf{Y}, \underbrace{\mathbf{\Sigma}}_{fijo})$$

y análogamente,

$$p(\mathbf{\Sigma} \mid Y, \boldsymbol{\beta}) \propto p(\mathbf{\Sigma}, \mathbf{Y}, \underbrace{\boldsymbol{\beta}}_{fijo})$$

Bajo este marco de referencia se tienen los siguientes resultados.

Resultado 6.3.1. *La distribución posterior del parámetro $\boldsymbol{\beta}$ condicionado a $\mathbf{\Sigma}, \mathbf{Y}, \mathbf{X}$ es*

$$\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \mathbf{\Sigma} \sim Normal_q(\mathbf{b}_q, \mathbf{B}_q)$$

donde

$$\begin{aligned} \mathbf{B}_q &= \left(\mathbf{B}^{-1} + \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1} \\ \mathbf{b}_q &= \mathbf{B}_q \left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Y}\right) \end{aligned}$$

Prueba. Utilizando el condicionamiento posterior, y un razonamiento análogo al de la demostración del Resultado 5.1.2, se tiene que

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X}) &\propto p(\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}, \underbrace{\boldsymbol{\Sigma}}_{fijo}) \\ &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) p(\boldsymbol{\beta}) \\ &\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta}-\mathbf{b})\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{b}_q)'\mathbf{B}_q^{-1}(\boldsymbol{\beta}-\mathbf{b}_q)\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de un vector aleatorio con distribución $Normal_q(\mathbf{b}_q, \mathbf{B}_q)$. ■

Resultado 6.3.2. *La distribución posterior del parámetro* $\boldsymbol{\Sigma}$ *condicionado a* $\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}$ *es*

$$\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \sim Inversa - Whishart_{v+q+1}(\mathbf{S}_{\boldsymbol{\beta}})$$

donde $\mathbf{S}_{\boldsymbol{\beta}} = (Y - X\boldsymbol{\beta})(Y - X\boldsymbol{\beta}) + \boldsymbol{\Lambda}$.

Prueba. Utilizando el condicionamiento posterior, y escribiendo la verosimilitud de las observaciones como

$$p(\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{1/2} \exp\left\{-\frac{1}{2} traza[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}]\right\}$$

entonces se tiene que

$$\begin{aligned} p(\boldsymbol{\Sigma} \mid \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}) &\propto p(\boldsymbol{\Sigma}, \mathbf{Y}, \mathbf{X}, \underbrace{\boldsymbol{\beta}}_{fijo}) \\ &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}) \\ &\propto \mid \boldsymbol{\Sigma} \mid^{1/2} \exp\left\{-\frac{1}{2} traza[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}]\right\} \\ &\qquad \times \mid \boldsymbol{\Sigma} \mid^{-\frac{v+q+1}{2}} \exp\left\{-\frac{1}{2} traza[\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}]\right\} \\ &= \mid \boldsymbol{\Sigma} \mid^{-\frac{v+q+1+1}{2}} \\ &\qquad \times \exp\left\{-\frac{1}{2} traza[((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})' + \boldsymbol{\Lambda})\boldsymbol{\Sigma}^{-1}]\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una matriz aleatoria con distribución $Inversa - Whishart_{v+q+1}(\mathbf{S}_{\boldsymbol{\beta}})$. ■

6.4 ANOVA

Uno de los modelos de regresión más utilizados y a la vez más simples es el análisis de varianza ANOVA a una vía, el cual es un caso particular del modelo lineal general. Este enfoque considera la partición de las observaciones en bloques, estratos o subgrupos que el investigador conoce de antemano, antes de la realización del experimento. Uno de los motivos de la realización de la partición es porque se conoce que la estructura probabilística (de localización, de escala, o ambas) cambia significativamente en las observaciones con respecto al grupo de pertenencia. De esta manera, este modelo está dado por la ecuación general $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ toma la siguiente forma

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_J \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_J \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_J \end{pmatrix} \tag{6.4.1}$$

en donde $\mathbf{Y}_j = [Y]_{ij}$ con $j = 1, \ldots, J$ e $i = 1, \ldots, n_j$ denota el vector de observaciones en el subgrupo j, $\mathbf{1}_{n_j}$ es un vector de unos de tamaño n_j y $\boldsymbol{\varepsilon}_j = [\varepsilon]_{ij}$ es el vector de errores en el subgrupo j. El análisis bayesiano de este tipo de modelos está supeditado a los resultados encontrados en este capítulo al reemplazar la matriz de diseño $\mathbf{X}$ por una matriz cuyas columnas sean de unos y de ceros. Una formulación del modelo es la siguiente

$$\begin{aligned} Y_{ij} \mid \boldsymbol{\beta}_j &\sim Normal(\boldsymbol{\beta}_j, \sigma^2) \\ \boldsymbol{\beta}_j \mid \mu, \tau^2 &\sim Normal(\mu, \tau^2) \end{aligned}$$

con σ^2, μ y τ^2 conocidos, pero estos últimos reflejando las suposiciones previa que el investigador considera pertinentes. Tomando la verosimilitud de todo el vector de observaciones, notamos que

$$p(\mathbf{Y} \mid \boldsymbol{\beta}) = \prod_{j=1}^{J} p(\mathbf{Y}_j \mid \beta_j) \tag{6.4.2}$$

Y a su vez, nótese que

$$\begin{aligned} p(\mathbf{Y}_j \mid \beta_j) = \prod_{i=1}^{n_j} p(Y_{ij} \mid \beta_j) &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_j} (Y_{ij} - \beta_j)^2 \right\} \\ &\propto \exp\left\{ -\frac{n_j}{2\sigma^2} (\bar{Y}_j - \beta_j)^2 \right\} \end{aligned}$$

Por lo tanto la expresión XXXXX queda completamente definida como

$$p(\mathbf{Y} \mid \boldsymbol{\beta}) \propto \prod_{j=1}^{J} \exp\left\{ -\frac{1}{2\sigma_j^2} (\bar{Y}_j - \beta_j)^2 \right\} \tag{6.4.3}$$

donde $\sigma_j^2 = \sigma^2/n_j$. Luego, siguiendo la regla de bayes, la distribución posterior del vector de parámetros de interés $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ es

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \mu, \tau^2) \\ &\propto \prod_{j=1}^{J} p(\mathbf{Y}_j \mid \beta_j) \prod_{j=1}^{J} p(\beta_j \mid \mu, \tau^2) \\ &\propto \exp\left\{ \sum_{j=1}^{J} \frac{-1}{2\sigma_j^2} (\bar{y}_j - \beta_j)^2 \right\} \frac{1}{\tau^J} \exp\left\{ \frac{-1}{2\tau^2} \sum_{j=1}^{J} (\theta_j - \mu)^2 \right\} \end{aligned}$$

Bajo este marco de referencia se tienen el siguiente resultado

Resultado 6.4.1. *La distribución posterior del componente β_j perteneciente al vector de parámetros de interés $\boldsymbol{\beta}$ es*

$$\beta_j \sim Normal(\mu_j, \tau_j^2)$$

en donde

$$\mu_j = \frac{\frac{n_j}{\sigma^2}\bar{Y}_j + \frac{1}{\tau^2}\mu}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_j^2 = \left(\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2} \right)^{-1}$$

Prueba. La prueba del resultado es análoga a la demostración del resultado 2.6.4. ■

Nótese que la media de la distribución posterior se puede ver como un promedio ponderado de la media de la distribución previa y la media de las observaciones en la muestra. La ponderación de las anteriores medias son los parámetros de precisión (el inverso de la varianza) de la distribución previa y de la verosimilitud. Por otro lado, el parámetro de precisión de la distribución a posterior está dado por la suma de las precisiones de la verosimilitud y de la distribución previa. Ahora, si el tamaño de muestra es grande, para cualquier escogencia de los parámetros previa, se tiene que la distribución posterior converge a

$$\beta_j \underset{n_j \to \infty}{\sim} Normal(\bar{Y}_j, \sigma_j^2/n)$$

Bajo este contexto, la distribución de los parámetros de interés es idéntica a la que se presenta en el caso frecuentista clásico. Lo mismo acontece si la distribución previa es no informativa o cuando $\sigma_j^2 \to \infty$. Lo anterior se presenta puesto que la distribución previa no aporta mayor información a la distribución posterior, mientras que la verosimilitud recoge toda la información.

6.5 Promediando los modelos bayesiano

XXXXXXXXXXXXXXXXHoeting et. al. (1999)XXXXXXXXXXXXXX afirma que la práctica habitual de la estadística hace caso omiso de la incertidumbre de

los modelos. Los estadísticos suelen seleccionar un modelo de alguna familia de modelos y luego proceden como si el modelo elegido hubiese generado esos datos. Este enfoque hace caso omiso de la incertidumbre en la selección del modelo, dando lugar a inferencias muy confiadas y a la toma de decisiones más riesgosas de lo que uno pensaría.

Un promedio de modelos Bayesianos (BMA, por sus siglas en inglés) proporciona un mecanismo coherente para dar cuenta de la incertidumbre de los modelos. Existen varios métodos de aplicación del BMA que han surgido recientemente y en esta entrada voy a utilizar la información del archivo de datos principal del paquete TeachingSampling para explicar paso a paso la adecuación de esta metodología que arroja coeficientes de regresión que resultan ser un promedio de los coeficientes de cada posible modelo. Más aún, se trata de un promedio ponderado por la respectiva probabilidad a posteriori de cada modelo.

Siguiendo la regla de Bayes, la probabilidad a posteriori de cada modelo (PMP, por sus siglas en inglés) resulta ser proporcional a la verosimilitud marginal del modelo (la probabilidad de los datos dado el modelo) multiplicado por la distribución previa del modelo. En muchas ocasiones, la distribución previa del modelo se asume tipo g-Zelnner, que es una distribución normal con media nula y varianza dependiendo de un hiperparámetro de incertidumbre g. Un valor pequeño de g implica un gran conocimiento previo de que los coeficientes del modelo son nulos, y un valor grande para g implica que el investigador no está muy seguro de que los coeficientes del modelo sean cero.

Con base en el anterior razonamiento, se utilizará la base de datos Lucy para ilustrar el ajuste de un promedio de modelos bayesianos. En primer lugar, cargamos la librería TeachingSampling para poder acceder a los daros y también la librería BMS para realizar el ajuste de los modelos. La base de datos la constituyen 2396 empresas del sector industrial, la variable de interés es el número de empleados de cada empresa y las variables regresoras son el total de impuestos declarados, el total de ingresos, el nivel de industrialización, la zona de ubicación y el tipo de publicidad en el último año fiscal.

```
> library(TeachingSampling)
> library(BMS)
> data(Lucy)
> databma <- data.frame(Emp=Lucy$Employees, Tax=Lucy$Taxes, Inc=Lucy$Income,
+ Lev=as.double(Lucy$Level), Zon=as.double(Lucy$Zone), Spa=as.double(Lucy$SPAM))
```

Para ajustar los modelos, se utiliza la función bms de la librería BMS. Esta función ajusta todos los 2^k posibles modelos (siendo k el número total de variables regresoras), computa todas las PMPs, calcula todos los coeficientes de regresión en cada uno de esos modelos, y al final promedia estos coeficientes utilizando como ponderador las PMPs. Una característica importante en esta función es que la primera columna del archivo de datos debe ser la variable de interés.

```
> Lucybma <- bms(databma, burn=100000, iter=200000, g="BRIC", mprior="uniform", mcmc="bd
```

La función coef arroja las probabilidades de inclusión posteriores (PIP) de cada variable en los modelos, la media posterior de cada coeficiente de regresión (la misma estimación bayesiana) y el error estándar posterior. Cada PIP se calcula como la suma de las PMPs para cada modelo en donde esa covariable fue incluida. Por ejemplo, para Lucy, la variable más importante es Tax, la cual tiene probabilidad de inclusión igual a uno pues fue incluida en todos los posibles modelos. Luego le sigue la variable Inc, con probabilidad de inclusión 0.99, y luego la variable Lev, con probabilidad de inclusión 0.89. Para estas variables, la estimación bayesiana de sus respectivos coeficientes de regresión son 0.66, 0.03 y -5.63, respectivamente.

```
> coef(Lucybma, std.coefs = T, include.constant = T)
                 PIP     Post Mean     Post SD Cond.Pos.Sign Idx
Tax         1.000000  3.486176e-01 0.038496653             1   1
Inc         0.999785  2.437994e-01 0.054688020             1   2
Lev         0.896795 -9.475564e-02 0.043172340             0   3
Spa         0.054590 -1.203361e-03 0.006156046             0   5
Zon         0.020045 -9.232616e-05 0.002313997             0   4
(Intercept) 1.000000  1.747969e+00          NA            NA   0
```

La función topmodels.bma arroja una matriz de unos y ceros, donde las columnas representan el modelo ajustado y las filas las variables regresoras. Las entradas de esta matriz son uno, si la variable regresora fue incluida en el modelo, y cero, en otro caso. En las últimas filas, se presentan las PMP. Para este caso, el mejor modelo, con una probabilidad a posteriori de 0.82, es el que incluye las variables regresoras Tax, Inc y Lev.

```
> topmodels.bma(Lucybma) ## Mejores modelos según la PMP
                   1c         18         1d         1e          19          1a
Tax         1.0000000 1.00000000 1.00000000 1.0000000 1.000000000 1.000000000
Inc         1.0000000 1.00000000 1.00000000 1.0000000 1.000000000 1.000000000
Lev         1.0000000 0.00000000 1.00000000 1.0000000 0.000000000 0.000000000
Zon         0.0000000 0.00000000 0.00000000 1.0000000 0.000000000 1.000000000
Spa         0.0000000 0.00000000 1.00000000 0.0000000 1.000000000 0.000000000
PMP (Exact) 0.8277914 0.09783366 0.04749275 0.0177992 0.005750708 0.002000072
PMP (MCMC)  0.8301100 0.09601000 0.04824500 0.0173450 0.005375000 0.001760000
                   1f           14           1b           15
Tax         1.000000000 1.0000000000 1.0000000000 1.000000e+00
Inc         1.000000000 0.0000000000 1.0000000000 0.000000e+00
Lev         1.000000000 1.0000000000 0.0000000000 1.000000e+00
Zon         1.000000000 0.0000000000 1.0000000000 0.000000e+00
Spa         1.000000000 0.0000000000 1.0000000000 1.000000e+00
PMP (Exact) 0.001015727 0.0001877044 0.0001174979 1.129789e-05
PMP (MCMC)  0.000885000 0.0001850000 0.0000600000 3.000000e-05
```

La función plot.Conv grafica las distribuciones previa y posterior para los tamaños (número de variables incluidas) en el modelo. Para nuestro ejemplo, la

distribución previa daba mayor probabilidad a los modelos que incluían dos o tres variables regresoras, mientras que la distribución posterior da mayor peso a los modelos de tres variables regresoras.

```
> plotConv(Lucybma)
```

La función beta.draws.bma da como resultado los coeficientes de regresión para todos los modelos. Nótese que promediando estos valores, con su respectiva ponderación, se tiene la estimación bayesiana posterior del promedio de modelos dada por la segunda columna de la función coef.

```
> beta.draws.bma(Lucybma[1:5])  ## Los coeficientes de los 5 mejores modelos
             1c         18          1d          1e         19
Tax  0.66206081 0.65436602  0.66466816  0.66017039  0.6570294
Inc  0.02883466 0.04053246  0.02876660  0.02896933  0.0404346
Lev -6.29892336 0.00000000 -6.28361012 -6.34535538  0.0000000
Zon  0.00000000 0.00000000  0.00000000 -0.15891039  0.0000000
Spa  0.00000000 0.00000000 -1.48334674  0.00000000 -1.5043990
```

La función image arroja una gráfica que incluye cada variable. Si para esta variable el color es blanco, significa que no fue incluida en ese modelo, si el color es rojo, implica que el signo del coeficiente de regresión es negativo, y si el color es azul, significa que el signo del coeficiente de regresión es positivo. Nótese que esta figura está basada en probabilidades acumuladas; así que entre más ancha sean los cuadros, implica que el modelo tiene una mayor PMP.

```
> image(Lucybma[1:5])
```

Para tener un acercamiento completo a la distribución posterior de los coeficientes, la función density proyecta una gráfica de la densidad posterior del coeficiente.

```
> density(Lucybma,"Tax")
> density(Lucybma,"Inc")
```

7 Modelos lineales mixtos y multinivel

En los últimos años, los modelos de regresión con estructura probabilística jerárquica han venido en auge. La ventaja de los avances computacionales interviniendo en la práctica estadística ha permitido que el investigador utilice armas teóricas más complejas y novedosas para modelar las observaciones y replicar el estado de la naturaleza. **?** definen este tipo de enfoques como un conjunto de múltiples modelos de regresión en los cuales los parámetros - los coeficientes de regresión - tienen una estructura probabilística enmarcada en etapas superiores.

El ejemplo típico para este tipo de modelos es el siguiente. En muchas escuelas se lleva a cabo un estudio observacional que tiene como objetivo predecir las notas de los estudiantes, en una prueba estatal, utilizando notas anteriores en una prueba anterior. Es posible ajustar un modelo de regresión dentro de cada una de las escuelas y a su vez, los parámetros de cada escuela pueden ser modelados de acuerdo a variables sociales, económicas y demográficas. Como el objetivo principal está enfocado en las notas individuales de cada estudiante, entonces también se pueden ajustar modelos de regresión a cada estudiante. En general, la estructura jerárquica es notoria puesto que se ajustan distintos modelos en distintos niveles (escuelas y estudiantes). Este tipo de modelos recibe el nombre de *modelos de regresión jerárquicos* o *modelos multinivel*.

Aunque es posible proponer modelos que contemplen demasiadas etapas, no es realista ajustar modelos con más de tres niveles. Una formulación general para un modelo con dos niveles es la siguiente

$$\begin{aligned}
\mathbf{Y} \mid \mathbf{X}_1, \boldsymbol{\beta}_1, \boldsymbol{\Sigma}_{\mathbf{Y}} &\sim Normal_n(\mathbf{X}_1\boldsymbol{\beta}_1, \boldsymbol{\Sigma}_{\mathbf{Y}}) \\
\boldsymbol{\beta}_1 \mid \mathbf{X}_2, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1} &\sim Normal_q(\mathbf{X}_2\boldsymbol{\beta}_2, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}) \\
\boldsymbol{\beta}_2 \mid \mathbf{b}, \mathbf{B} &\sim Normal_r(\mathbf{b}, \mathbf{B})
\end{aligned}$$

donde $\mathbf{b}, \mathbf{B}$ son hiperparámetros conocidos. Nótese que el orden de la verosimilitud de las observaciones es n, el tamaño de la muestra, mientras que el orden de la distribución previa para el parámetro $\boldsymbol{\beta}_1$ es q, el número de variables de información auxiliar para el primer nivel, y por último el orden de la distribución previa para el hiperparámetro $\boldsymbol{\beta}_2$ es r, el número de variables de información auxiliar para el segundo nivel.

7.1 Regresión en dos niveles

Esta sección empieza con la advertencia que da **?** acerca de que en muy pocos casos es posible encontrar una solución analítica[1] a este tipo de enfoques jerárquicos. Sin embargo, lo anterior no significa que un análisis bayesiano propiamente dicho no sea posible de realizar. Como se concluirá más adelante, la técnica del condicionamiento posterior será un baluarte importante en la inferencia bayesiana de los modelos lineales jerárquicos.

Considere el siguiente modelo de regresión en dos niveles, más conocido como el modelo de regresión clásico anidado y dado por la siguiente formulación

$$\begin{aligned}\mathbf{Y} \mid \mathbf{X}_1, \boldsymbol{\beta}_1, \sigma^2 &\sim Normal_n(\mathbf{X}_1\boldsymbol{\beta}_1, \sigma^2\mathbf{I}_n)\\ \boldsymbol{\beta}_1 \mid \mathbf{X}_2, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1} &\sim Normal_q(\mathbf{X}_2\boldsymbol{\beta}_2, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1})\\ \boldsymbol{\beta}_2 \mid \mathbf{b}, \mathbf{B} &\sim Normal_r(\mathbf{b}, \mathbf{B})\end{aligned}$$

donde $\mathbf{b}, \mathbf{B}$ son hiperparámetros conocidos. Además la distribución de las estructuras de dispersión es

$$\sigma^2 \sim Inversa - Gamma\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$$

Suponiendo que la matriz de varianzas $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}$ para el vector de parámetros $\boldsymbol{\beta}_1$ tiene una estructura conocida, entonces la distribución conjunta de las observaciones y todos los parámetros de interés[2] estaría dada por

$$\begin{aligned}p(\mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2) &= p(\mathbf{Y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2)p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \sigma^2)p(\boldsymbol{\beta}_2, \sigma^2)\\ &= p(\mathbf{Y} \mid \boldsymbol{\beta}_1, \sigma^2)p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2)p(\boldsymbol{\beta}_2)p(\sigma^2) \qquad (7.1.1)\end{aligned}$$

suponiendo que $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}$ es conocida y basados en la anterior expresión se tiene el siguiente conjunto de resultados.

Resultado 7.1.1. *La distribución a posterior del vector de parámetros de interés* $\boldsymbol{\beta}_1$ *condicionado a* $\mathbf{Y}, \boldsymbol{\beta}_2, \sigma^2$ *es*

$$\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}_2, \sigma^2 \sim Normal_q(\mathbf{b}_q, \mathbf{B}_q)$$

donde

$$\begin{aligned}\mathbf{B}_q &= \left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1} + \frac{1}{\sigma^2}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\\ \mathbf{b}_q &= \mathbf{B}_q\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1}\mathbf{X}_2\boldsymbol{\beta}_2 + \frac{1}{\sigma^2}\mathbf{X}_1'\mathbf{Y}\right)\end{aligned}$$

[1] Referente a la utilización de técnicas de integración.

[2] Suponiendo que el vector de parámetros $\boldsymbol{\beta}_2$ es condicionalmente independiente de las observaciones y es independiente de σ^2.

Prueba. Del condicionamiento posterior se tiene que

$$\begin{aligned} p(\boldsymbol{\beta}_1 \mid \mathbf{Y}, \boldsymbol{\beta}_2, \sigma^2) &\propto p(\boldsymbol{\beta}_1, \underbrace{\mathbf{Y}, \boldsymbol{\beta}_2, \sigma^2}_{fijos}) \\ &\propto p(\mathbf{Y} \mid \boldsymbol{\beta}_1, \sigma^2) p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2) \end{aligned}$$

por tanto la prueba del resultado es inmediata acudiendo a un razonamiento idéntico al usado en la demostración del Resultado 5.1.2. ■

Resultado 7.1.2. *La distribución a posterior del vector de parámetros de interés* $\boldsymbol{\beta}_2$ *condicionado a* $\mathbf{Y}, \boldsymbol{\beta}_1, \sigma^2$ *es*

$$\boldsymbol{\beta}_2 \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}_1, \boldsymbol{\Sigma} \sim Normal_r(\mathbf{b}_r, \mathbf{B}_r)$$

donde

$$\begin{aligned} \mathbf{B}_r &= \left(\mathbf{X}_2' \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1} \mathbf{X}_2 + \mathbf{B}^{-1} \right)^{-1} \\ \mathbf{b}_r &= \mathbf{B}_r \left(\mathbf{X}_2' \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1} \boldsymbol{\beta}_1 + \mathbf{B}^{-1} \mathbf{b} \right) \end{aligned}$$

Prueba. Del condicionamiento posterior se tiene que

$$\begin{aligned} p(\boldsymbol{\beta}_2 \mid \mathbf{Y}, \boldsymbol{\beta}_1, \sigma^2) &\propto p(\boldsymbol{\beta}_2, \underbrace{\mathbf{Y}, \boldsymbol{\beta}_1, \sigma^2}_{fijos}) \\ &\propto p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2) p(\boldsymbol{\beta}_2) \end{aligned}$$

Por tanto el desarrollo algebraico conduce a que

$$\begin{aligned} p(\boldsymbol{\beta}_2 \mid \mathbf{Y}, \boldsymbol{\beta}_1, \sigma^2) &\propto \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\beta}_1 - \mathbf{X}_2 \boldsymbol{\beta}_2)' \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1} (\boldsymbol{\beta}_1 - \mathbf{X}_2 \boldsymbol{\beta}_2) + (\boldsymbol{\beta}_2 - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta}_2 - \mathbf{b}) \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}_2' (\mathbf{X}_2' \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1} \mathbf{X}_2 + \mathbf{B}^{-1}) \boldsymbol{\beta}_2 - 2 \boldsymbol{\beta}_2' (\mathbf{X}_2' \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}^{-1} \boldsymbol{\beta}_1 + \mathbf{B}^{-1} \mathbf{b}) \right] \right\} \\ &= \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}_2' \mathbf{B}_r^{-1} \boldsymbol{\beta}_2 + \boldsymbol{\beta}_2' \mathbf{B}_r^{-1} \mathbf{b}_r \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}_2' \mathbf{B}_r^{-1} \boldsymbol{\beta}_2 + \boldsymbol{\beta}_2' \mathbf{B}_r^{-1} \mathbf{b}_r - \frac{1}{2} \mathbf{b}_r' \mathbf{B}_r^{-1} \mathbf{b}_r \right\} \\ &= \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta}_2 - \mathbf{b}_r)' \mathbf{B}_r^{-1} (\boldsymbol{\beta}_2 - \mathbf{b}_r) \right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una vector aleatorio con distribución $Normal_r(\mathbf{b}_r, \mathbf{B}_r)$. ■

? afirma que es sorprendente que la distribución posterior de $\boldsymbol{\beta}_2$ no dependa de las observaciones. Este hecho se debe al carácter jerárquico del modelo que pasa, a través de $\boldsymbol{\beta}_1$, toda la información contenida en $\mathbf{y}$ a $\boldsymbol{\beta}_2$.

Resultado 7.1.3. *La distribución a posterior del parámetro de interés σ^2 condicionado a $\mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_1$ es*

$$\sigma^2 \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \sim Inversa - Gamma\left(\frac{n_1}{2}, \frac{n_1 S^2_{\boldsymbol{\beta}_1}}{2}\right)$$

donde $n_1 = n_0 + n$ *y* $n_1 S^2_{\boldsymbol{\beta}_1} = n_0\sigma_0^2 + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_1)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_1)$.

Prueba. Del condicionamiento posterior se tiene que

$$\begin{aligned} p(\sigma^2 \mid \mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &\propto p(\sigma^2, \underbrace{\mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2}_{fijos}) \\ &\propto p(Y \mid \boldsymbol{\beta}_1, \sigma^2) p(\sigma^2) \end{aligned}$$

Y desarrollando el anterior producto, se concluye que

$$\begin{aligned} p(\sigma^2 \mid \mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \\ &\quad \times (\sigma^2)^{-n_0/2-1} \exp\left\{-\frac{n_0\sigma_0^2}{2\sigma^2}\right\} \\ &\propto (\sigma^2)^{-\frac{n_1}{2}-1} \exp\left\{-\frac{n_1 S_1^2}{2\sigma^2}\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Inversa - Gamma(n_1/2, n_1 S^2_{\boldsymbol{\beta}_1}/2)$. ■

Con los anteriores resultados, se tiene la distribución posterior condicional de todos y cada uno de los parámetros de interés y con esto se llega a una análisis bayesiano propiamente dicho.

Por otro lado, en algunas ocasiones es más realista asumir que no se conoce la matriz $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}$ aunque si se tiene alguna idea del comportamiento probabilístico de esta matriz que refleja estructura de variación del vector de parámetros de interés. En esta situación, un razonamiento parecido al de las últimas páginas nos transporta directamente, asumiendo que $\boldsymbol{\Sigma}$ es independiente de σ^2, a la siguiente situación

$$p(\mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}) = p(\mathbf{Y} \mid \boldsymbol{\beta}_1, \sigma^2) p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}) p(\boldsymbol{\beta}_2) p(\sigma^2) p(\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}) \quad (7.1.2)$$

Suponiendo que la distribución previa de $\boldsymbol{\Sigma}$ está dada por

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1} \sim Inversa - Wishart_{n_w}(n_w \mathbf{S}_w)$$

entonces se tiene que, utilizando el condicionamiento posterior, las distribuciones condicionales de los parámetros de interés $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, y σ^2 se mantienen idénticas debido a la independencia de estos con respecto a $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}$.

Resultado 7.1.4. *La distribución posterior de la matriz de varianzas* $\mathbf{\Sigma}_{\boldsymbol{\beta}}$ *condicionada a* $\mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2$ *es*

$$\mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 \sim Inversa - Wishart_{n_w^*}\left(n_w^* \mathbf{S}_w^*\right)$$

en donde $n_w^* \mathbf{S}_w^* = (\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2)(\boldsymbol{\beta}_1 - \mathbf{X}_\mathbf{2}\boldsymbol{\beta}_2)' + n_w \mathbf{S}_w$

Prueba. Del condicionamiento posterior se tiene que

$$\begin{aligned} p(\mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid \mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2) &\propto p(\mathbf{\Sigma}_{\boldsymbol{\beta}_1}, \underbrace{\mathbf{Y}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2}_{fijos}) \\ &\propto p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \mathbf{\Sigma}_{\boldsymbol{\beta}_1}) p(\mathbf{\Sigma}_{\boldsymbol{\beta}_1}) \end{aligned}$$

y teniendo en cuenta que la distribución previa de $\boldsymbol{\beta}_1$ puede ser reescrita como

$$\begin{aligned} p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \mathbf{\Sigma}_{\boldsymbol{\beta}_1}) \propto & \mid \mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_1 - \mathbf{X}_\mathbf{2}\boldsymbol{\beta}_2)' \mathbf{\Sigma}_{\boldsymbol{\beta}_1}^{-1} (\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2)\right\} \\ = & \mid \mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid^{-1/2} \exp\left\{-\frac{1}{2} traza\left[(\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2)(\boldsymbol{\beta}_1 - \mathbf{X}_\mathbf{2}\boldsymbol{\beta}_2)' \mathbf{\Sigma}_{\boldsymbol{\beta}_1}^{-1}\right]\right\} \end{aligned}$$

Por tanto la distribución posterior de $\mathbf{\Sigma}_{\boldsymbol{\beta}}$ toma la siguiente forma

$$\begin{aligned} p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \mathbf{\Sigma}_{\boldsymbol{\beta}_1}) \propto & \mid \mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid^{-1/2} \exp\left\{-\frac{1}{2} traza\left[(\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2)(\boldsymbol{\beta}_1 - \mathbf{X}_\mathbf{2}\boldsymbol{\beta}_2)' \mathbf{\Sigma}_{\boldsymbol{\beta}_1}^{-1}\right]\right\} \\ & \times \mid \mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid^{-\frac{n_w + r + 1}{2}} \exp\left\{-\frac{1}{2} traza\left[n_w \mathbf{S}_w \mathbf{\Sigma}_{\boldsymbol{\beta}_1}^{-1}\right]\right\} \\ = & \mid \mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid^{-\frac{(n_w - 1) + r + 1}{2}} \\ & \times \exp\left\{-\frac{1}{2} traza\left[(\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2)(\boldsymbol{\beta}_1 - \mathbf{X}_\mathbf{2}\boldsymbol{\beta}_2)' \mathbf{\Sigma}_{\boldsymbol{\beta}_1}^{-1} + n_w \mathbf{S}_w \mathbf{\Sigma}_{\boldsymbol{\beta}_1}^{-1}\right]\right\} \\ = & \mid \mathbf{\Sigma}_{\boldsymbol{\beta}_1} \mid^{-\frac{n_w^* + r + 1}{2}} \exp\left\{-\frac{1}{2} traza\left[n_w^* \mathbf{S}_w^* \mathbf{\Sigma}_{\boldsymbol{\beta}_1}^{-1}\right]\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una matriz aleatoria con distribución $Inversa - Wishart_{n_w^*}\left(n_w^* \mathbf{S}_w^*\right)$. ■

7.2 ANOVA jerarquica

Uno de los modelos de regresión más utilizados y a la vez más simples es el análisis de varianza ANOVA a una vía, el cual es un caso particular del modelo lineal general. Este enfoque considera la partición de las observaciones en bloques, estratos o subgrupos que el investigador conoce de antemano, antes de la realización del

experimento. Uno de los motivos de la realización de la partición es porque se conoce que la estructura probabilística (de localización, de escala, o ambas) cambia significativamente en las observaciones con respecto al grupo de pertenencia. De esta manera, este modelo está dado por la ecuación general $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ toma la siguiente forma

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_J \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_J \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_J \end{pmatrix} \tag{7.2.1}$$

en donde $\mathbf{Y}_j = [Y]_{ij}$ con $j = 1, \ldots, J$ e $i = 1, \ldots, n_j$ denota el vector de observaciones en el subgrupo j, $\mathbf{1}_{n_j}$ es un vector de unos de tamaño n_j y $\boldsymbol{\varepsilon}_j = [\varepsilon]_{ij}$ es el vector de errores en el subgrupo j. El análisis bayesiano de este tipo de modelos está supeditado a los resultados encontrados en capítulos anteriores al reemplazar la matriz de diseño $\mathbf{X}$ por una matriz cuyas columnas sean de unos y de ceros.

Sin embargo si la estructura probabilística del vector de parámetros de interés es desconocida, entonces este modelo se considera jerárquico y debe ser tratado como tal. Suponiendo las anteriores condiciones, una formulación del modelo, asumiendo que los hiperparámetros son no informativos, es la siguiente

$$\begin{aligned} Y_{ij} \mid \boldsymbol{\beta}_j &\sim Normal(\boldsymbol{\beta}_j, \sigma^2) \\ \boldsymbol{\beta}_j \mid \mu, \tau^2 &\sim Normal(\mu, \tau^2) \end{aligned}$$

con σ^2 conocido, pero μ y τ^2 hiperparámetros desconocidos. Fijándonos en la verosimilitud de todo el vector de observaciones, notamos que

$$p(\mathbf{Y} \mid \boldsymbol{\beta}) = \prod_{j=1}^{J} p(\mathbf{Y}_j \mid \beta_j) \tag{7.2.2}$$

Y a su vez, nótese que

$$\begin{aligned} p(\mathbf{Y}_j \mid \beta_j) = \prod_{i=1}^{n_j} p(Y_{ij} \mid \beta_j) &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_j} (Y_{ij} - \beta_j)^2 \right\} \\ &\propto \exp\left\{ -\frac{n_j}{2\sigma^2} (\bar{Y}_j - \beta_j)^2 \right\} \end{aligned}$$

Por lo tanto la expresión 6.1.4 queda completamente definida como

$$p(\mathbf{Y} \mid \boldsymbol{\beta}) \propto \prod_{j=1}^{J} \exp\left\{ -\frac{1}{2\sigma_j^2} (\bar{Y}_j - \beta_j)^2 \right\} \tag{7.2.3}$$

donde $\sigma_j^2 = \sigma^2/n_j$. Por otro lado, suponga que los hiperparámetros son dependientes aunque su distribución previa conjunta es no informativa. Es decir, la distribución previa de los hiperparámetros está dada por

$$p(\mu, \tau^2) = p(\mu \mid \tau^2)p(\tau) \propto 1$$

Luego, siguiendo la regla de bayes y suponiendo que los hiperparámetros son condicionalmente independientes de las observaciones dado el vector de parámetros de interés, la distribución posterior del vector de parámetros de interés $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ y de los hiperparámetros μ, τ^2 es

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\beta})p(\boldsymbol{\beta} \mid \mu, \tau^2)p(\mu, \tau^2) \\ &\propto \prod_{j=1}^{J} p(\mathbf{Y}_j \mid \beta_j) \prod_{j=1}^{J} p(\beta_j \mid \mu, \tau^2) \\ &\propto \exp\left\{ \sum_{j=1}^{J} \frac{-1}{2\sigma_j^2}(\bar{y}_j - \beta_j)^2 \right\} \frac{1}{\tau^J} \exp\left\{ \frac{-1}{2\tau^2} \sum_{j=1}^{J} (\theta_j - \mu)^2 \right\} \end{aligned}$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 7.2.1. *La distribución posterior del componente β_j perteneciente al vector de parámetros de interés $\boldsymbol{\beta}$ es*

$$\beta_j \sim Normal(\mu_j, \tau_j^2)$$

en donde

$$\mu_j = \frac{\frac{1}{\sigma^2}\bar{Y}_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_j^2 = \left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2} \right)^{-1}$$

Prueba. La prueba del resultado es inmediata al considerar la técnica del condicionamiento posterior como en la demostración del Resultado 4.2.1. puesto que

$$\begin{aligned} p(\beta_j \mid \mu, \tau^2, \mathbf{Y}_j) &\propto p(\beta_j, \underbrace{\mu, \tau^2}_{fijos} \mid \mathbf{Y}_j) \\ &\propto p(\mathbf{Y}_j \mid \beta_j)p(\beta_j \mid \mu, \tau^2)p(\mu, \tau^2) \\ &\propto p(\mathbf{Y}_j \mid \beta_j)p(\beta_j \mid \mu, \tau^2) \end{aligned}$$

■

El siguiente paso corresponde a la determinación de la distribución posterior de los hiperparámetros μ, τ^2 la cual, suponiendo que la distribución previa conjunta

para ambos hiperparámetros es uniforme y no informativa, está dada por

$$\begin{aligned} p(\mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mu, \tau^2) p(\mathbf{Y} \mid \mu, \tau^2) \\ &\propto \prod_{j=1}^{J} p(\mathbf{Y}_j \mid \mu, \tau^2) \\ &\propto \prod_{j=1}^{J} Normal(\mu, \tau^2 + \sigma_j^2) \end{aligned}$$

Lo anterior se puede deducir directamente del Resultado 4.1.1. haciendo unos pequeños cambios en el álgebra. Por otro lado, el análisis individual de los hiperparámetros está regido por la siguiente expresión

$$p(\mu, \tau^2 \mid \mathbf{Y}) = p(\mu \mid \tau^2, \mathbf{Y}) p(\tau^2 \mid \mathbf{Y})$$

En este orden de ideas, se tienen los siguientes resultados acerca de la distribución posterior para μ dada por $p(\mu \mid \tau^2, \mathbf{Y})$ y para τ^2 dada por $p(\tau^2 \mid \mathbf{Y})$

Resultado 7.2.2. *La distribución posterior del hiperparámetro mu condicionada a* $\tau^2, \mathbf{Y}$ *es*

$$\mu \mid \tau^2, \mathbf{Y} \sim Normal\left(\hat{\mu}, \hat{\tau}^2\right)$$

En donde

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma^2+\tau^2} \bar{Y}_j}{\sum_{j=1}^{J} \frac{1}{\sigma^2+\tau^2}} \qquad y \qquad \hat{\tau}^2 = \left(\sum_{j=1}^{J} \frac{1}{\sigma^2 + \tau^2} \right)^{-1} \tag{7.2.4}$$

Prueba. Utilizando la técnica del condicionamiento posterior, nótese que la distribución posterior de *mu* toma la siguiente forma

$$\begin{aligned} p(\mu \mid \tau^2, \mathbf{Y}) &\propto p(\mu, \underbrace{\tau^2}_{fijo} \mid \mathbf{Y}) \\ &\propto \prod_{j=1}^{J} Normal(\mu, \tau^2 + \sigma_j^2) \end{aligned}$$

Partiendo de este hecho, es fácil confirmar que

$$\begin{aligned}
p(\mu \mid \tau^2, \mathbf{Y}) &\propto \exp\left\{\frac{-1}{2}\sum_{j=1}^{J}\frac{1}{\tau^2+\sigma_j^2}(\bar{y}_j-\mu)^2\right\} \\
&= \exp\left\{\frac{-1}{2}\left[\sum_{j=1}^{J}\frac{\bar{y}_j^2}{\tau^2+\sigma_j^2} - 2\mu\sum_{j=1}^{J}\frac{y_j}{\tau^2+\sigma_j^2} + \mu^2\sum_{j=1}^{J}\frac{1}{\tau^2+\sigma_j^2}\right]\right\} \\
&\propto \exp\left\{\frac{-1}{2}\left[\frac{\mu^2}{\hat{\tau}^2} - 2\frac{\mu\hat{\mu}}{\hat{\tau}^2}\right]\right\} \\
&\propto \exp\left\{\frac{-1}{2}\left[\frac{\mu^2}{\hat{\tau}^2} - 2\frac{\mu\hat{\mu}}{\hat{\tau}^2} + \frac{\hat{\mu}^2}{\hat{\tau}^2}\right]\right\} \\
&= \exp\left\{\frac{-1}{2\hat{\tau}^2}(\mu-\hat{\mu})^2\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\hat{\mu}, \hat{\tau}^2)$. ■

Resultado 7.2.3. *La distribución posterior del hiperparámetro τ es*

$$p(\tau^2 \mid \mathbf{Y}) \propto \sqrt{\hat{\tau}}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left\{-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_j-\hat{\mu})^2\right\}$$

Prueba. En primer lugar, nótese que

$$\begin{aligned}
p(\tau \mid \mathbf{Y}) &= \frac{p(\mu, \tau^2 \mid \mathbf{Y})}{p(\mu \mid \tau^2, \mathbf{Y})} \qquad \forall\mu \\
&\propto \frac{\prod_{j=1}^{J} Normal(\mu, \sigma^2+\tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \qquad \forall\mu
\end{aligned}$$

La anterior igualdad debe mantenerse para cualquier valor de μ; en particular se debe mantener para $\mu = \hat{\mu}$. Por tanto,

$$\begin{aligned}
p(\tau \mid \mathbf{Y}) &\propto \frac{Normal(\hat{\mu}, \sigma^2+\tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \\
&\propto \frac{\prod_{i=1}^{n} Normal(\hat{\mu}, \sigma^2+\tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \\
&\propto \sqrt{\hat{\tau}}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left\{-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_j-\hat{\mu})^2\right\}\exp\left\{\frac{1}{2\hat{\tau}^2}(\hat{\mu}-\hat{\mu})^2\right\} \\
&\propto \sqrt{\hat{\tau}}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left\{-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_j-\hat{\mu})^2\right\}
\end{aligned}$$

■

Como ya se debe tener en claro, en términos de simulación, los anteriores resultados garantizan una estructura formal que permita simular la distribución posterior del hiperparámetro τ^2, y mediante esta encontrar una estimación para reemplazarla en la distribución posterior del hiperparámetro μ y repetir el proceso anterior y mediante la definición de estos valores, entonces proseguir con el análisis bayesiano clásico.

7.3 Modelos multinivel

Gelman cap 12, 13 , 16

Es lo mismo que decir un modelo lineal mixto

7.3.1 Intercepto aleatorio

La siguiente programación permite simular[3] una población con estructura multinivel en el intercepto dado por la siguiente expresión

$$Y_{ij} \sim Normal(\alpha_j + \beta X_{ij}, \sigma_y^2)$$
$$\alpha_j \sim Normal(\gamma_0 + \gamma_1 U_j, \sigma_\alpha^2)$$

para $i = 1, \ldots, n_j$ y $j = 1, \ldots, J$, X_{ij} es una característica de información auxiliar a nivel del individuo y U_j es una característica de información auxiliar an nivel de bloque o estructura conteniendo al individuo. Por ejemplo, en estudios de calidad educativa, Y_{ij} puede denotar las calificaciones del alumno i-ésimo perteneciente a la escuela j-ésima, X_{ij} puede ser cualquier característica conocida del alumno; por ejemplo, nivel socioeconómico, rendimiento académico anterior al estudio, etc. Por último, U_j, denotará un característica concerniente a la escuela j-ésima; por ejemplo, el puesto que ocupa en el ranking de escuelas, el presupuesto que la escuela destina a perfeccionamiento docente, la cantidad de profesores que tiene, etc. Este modelo establece que para cada escuela el intercepto es diferente.

En primer lugar, suponemos que existen tres bloques o estructuras jerárquicas (en el caso del anterior ejemplo, existirán tres escuelas). En la primera escuela se realizó una muestra aleatoria simple de 10 estudiante, en la seguda se observaron a 30 estudiantes, al igual que en la última. Luego, se especifican los parámetros del modelo; en este caso, β, γ_0, γ_1, σ_y^2 y σ_{alpha}^2 .

```
library(mvtnorm)

## Simulating the population
```

[3] ? afirman que este tipo de ejercicios, en donde se simulan datos de un modelo, es una buena forma de entender el proceso de ajuste del modelo.

```
b <- -0.5
g0 <- 5
g1 <- 2
sig.y <- 0.8
sig.a <- 0.2
```

La siguiente instrucción permite simular la variable de información auxiliar para las escuelas y también los interceptos aleatorios del modelo.

```
J=3
u <- runif(J, 100, 250)
a <- rnorm(J, g0 + g1*u, sig.a)
a
```

El siguiente paso es concerniente a la simulación de los valores de los individuos

```
n=70
escuela <- c(rep(1,10), rep(2,30), rep(3,30))
u.e <- cbind(1,c(rep(u[1],10), rep(u[2],30), rep(u[3],30)))
x <- runif(n, 10, 40)
y <- rnorm(n, a[escuela] + b*x, sig.y)
y
```

Como se observó en la teoría expuesta anteriormente, para obtener inferencias posterior de los parámetros de interés β, γ_0 y γ_1, es necesario utilizar la técnica del condicionamiento sucesivo junto con el método de Gibbs. Para esto se especifican los valores de la distribución posterior de los parámetros.

```
## Verosimilitud
Sigma.y <- diag(rep(sig.y^2, n))

## previas
b.pri <- 0
B.pri <- 1000
g.pri <- c(0,0)
G.pri <- diag(c(100, 100))
```

Luego, se crean dos funciones que permiten la simulación de valores provenientes de las distribución posterior condicional de β y de $(\gamma_0, \gamma_1)'$. Estas funciones, siguiendo el espíritu de la técnica del condicionamiento sucesivo, depende valores iniciales de los parámetros.

```
## posterior beta
pos.beta<- function(g0.ini, g1.ini){
a.now <- g0.ini + g1.ini*u
y.beta <- y - a.now[escuela]
```

```
B.pos <- solve((1/B.pri)+t(x)%*%solve(Sigma.y)%*%x)
b.pos <- B.pos*((b.pri/B.pri)+t(x)%*%solve(Sigma.y)%*%y.beta)
rnorm(1, b.pos, sqrt(B.pos))
}

## posterior gama
pos.gama<- function(b.ini){
y.gama <- y - x*b.ini
G.pos <- solve(solve(G.pri)+t(u.c)%*%solve(Sigma.y)%*%u.c)
g.pos <- G.pos%*%(solve(G.pri)%*%g.pri+t(u.c)%*%solve(Sigma.y)%*%y.gama)
rmvnorm(1,g.pos,G.pos)
}
```

Por último, se fijan los valores iniciales y se programa el sencillo método de Gibbs cuya convergencia es bastante rápida. EL número de simulaciones es de 1000 y el periodo de calentamiento es la mitad de las cadenas. Es decir, se supone que las cadenas convergen después de la iteración 500.

```
# Gibbs sampler
nsim=1000
b.ini <- -50
g.ini <- c(50, 20)
b.mcmc <- matrix(NA, nrow=nsim, 1)
g.mcmc <- matrix(NA, nrow=nsim, 2)

for(k in 1:nsim){
b.mcmc[k,] <- b.ini
g.mcmc[k,] <- g.ini
b.ini <- pos.beta(g.ini[1], g.ini[2])
g.ini <- pos.gama(b.ini)
}

> mean(b.mcmc[500:1000])
[1] -0.5129415
> colMeans(g.mcmc[500:1000,])
[1] 5.773111 1.999123

plot(b.mcmc)
plot(g.mcmc[,1])
plot(g.mcmc[,2])
```

Las estimaciones bayesianas están dadas por $\hat{\beta} = -0.51$, $\hat{\gamma}_0 = 5.77$ y $\hat{\gamma}_1 = 1.99$. La figura XXXXX muestra el comportamiento de las cadenas y su rápida convergencia, aún cuando los valores iniciales están bastante alejados de los valores reales.

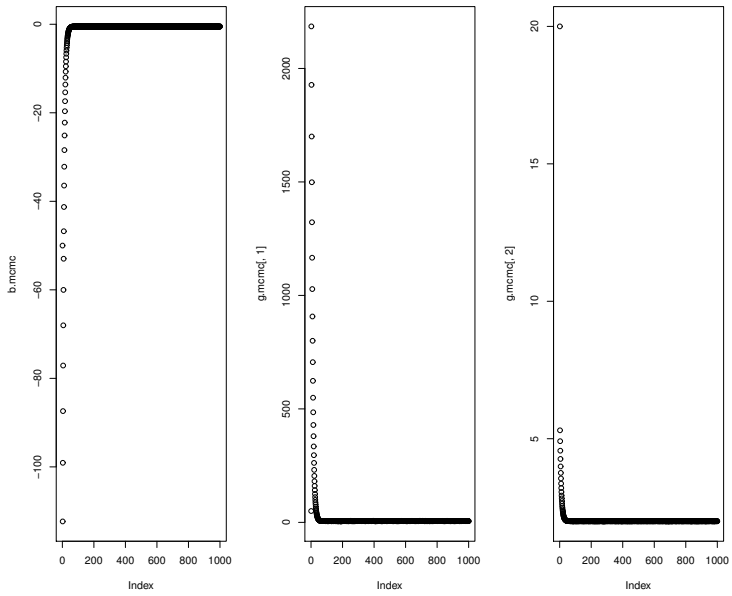

Figura 7.1: *Comportamiento y rápida convergencia de las cadenas para el ejemplo de las escuelas*

7.3.2 Simulación de un modelo multinivel

7.4 Modelos lineal mixto

8 Modelo lineal generalizado

8.1 Familia exponencial

En esta parte, se estudia la familia exponencial que es útil en algunos temas tratados en este libro, estos son, teoría de estimación puntual y prueba de hipótesis; y además resulta ser útil en la teoría bayesiana.

8.1.1 Familia exponencial uniparamétrica

DefiniciÃ³n8.1.1. *Una distribución de probabilidad con parámetro θ pertenece a la familia exponencial uniparamétrica si la función de densidad se puede escribir de la forma*

$$f_Y(y,\theta) = \exp\{d(\theta)T(y) - c(\theta)\}h(y) \tag{8.1.1}$$

donde $T(y)$ y $h(y)$ son funciones que depende de y únicamente, y $d(\theta)$ y $c(\theta)$ son funciones que depende de θ únicamente.

Ejemplo 8.1.1. *La distribución Poisson con parámetro θ pertenece a la familia exponencial uniparamétrica puesto que*

$$\begin{aligned} f(y,\theta) &= \frac{e^{-\theta}\theta^y}{y!} I_{\{0,1,\cdots\}}(y) \\ &= \exp\{y\ln\theta - \theta\}\frac{I_{0,1,\cdots}(y)}{y!}, \end{aligned}$$

*en conclusión $f(y,\theta)$ es de la forma (**??**) con $d(\theta) = \ln\theta$, $T(y) = y$, $c(\theta) = \theta$ y $h(y) = \frac{I_{\{0,1,\cdots\}}(y)}{y!}$.*

Ejemplo 8.1.2. *La distribución Gamma con parámetro de forma k conocida pertenece a la familia exponencial uniparamétrica puesto que*

$$\begin{aligned} f(y,\theta) &= \frac{y^{k-1}e^{-y/\theta}}{\theta^k\Gamma(k)} I_{(0,\infty)}(y) \\ &= \exp\left\{-\frac{y}{\theta} - k\ln\theta\right\}\frac{y^{k-1}I_{(0,\infty)}(y)}{\Gamma(k)}, \end{aligned}$$

*el cual es de la forma (**??**) con $d(\theta) = -1/\theta$, $T(y) = y$, $c(\theta) = k\ln\theta$ y $h(y) = \frac{y^{k-1}I_{(0,\infty)}(y)}{\Gamma(k)}$.*

Resultado 8.1.1. *Si $Y_1, \cdots, Y_n$ son variables aleatorias independientes e idénticamente distribuidas con función de densidad común perteneciente a la familia exponencial uniparamétrica, entonces la función de densidad conjunta $f(y_1, \cdots, y_n)$ también pertenece a la familia exponencial uniparamétrica.*

Prueba.

$$\begin{aligned} f(y_1, \cdots, y_n, \theta) &= \prod_{i=1}^{n} f(y_i, \theta) \\ &= \prod_{i=1}^{n} \exp\{d(\theta)T(y_i) - c(\theta)\}h(y_i) \\ &= \exp\left\{ d(\theta) \sum_{i=1}^{n} T(y_i) - nc(\theta) \right\} \prod_{i=1}^{n} h(y_i) \end{aligned}$$

el cual es de la forma (**??**). ■

8.1.2 Familia exponencial multi-paramétrica

DefiniciÃ³n8.1.2. *Una distribución de probabilidad pertenece a la familia exponencial multi-paramétrica si la función de densidad se puede escribir de la forma*

$$f_Y(y, \Theta) = \exp\{d(\Theta)'T(y) - c(\Theta)\}h(y) \tag{8.1.2}$$

donde $T(y)$ y $d(\Theta)$ son funciones vectoriales, $h(y)$ y $c(\Theta)$ son funciones reales.

Ejemplo 8.1.3. *La distribución Gamma con parámetro de forma k y parámetro de escala θ pertenece a la familia exponencial multi-paramétrica pues*

$$\begin{aligned} f_Y(y) &= \frac{y^{k-1}e^{-y/\theta}}{\theta^k \Gamma(k)} I_{(0,\infty)}(y) \\ &= \exp\{-\frac{y}{\theta} + (k-1)\ln y - k\ln\theta - \ln\Gamma(k)\} I_{(0,\infty)}(y) \\ &= \exp\left\{ \left(-\frac{1}{\theta}, k-1\right) \begin{pmatrix} y \\ \ln y \end{pmatrix} - (k\ln\theta + \ln\Gamma(k)) \right\} I_{(0,\infty)}(y) \end{aligned}$$

*el cual es de la forma (**??**) con $\Theta = (\theta, k)'$, $d(\Theta) = (-\frac{1}{\theta}, k-1)'$, $c(\Theta) = k\ln\theta + \ln\Gamma(k)$, $T(y) = (y, \ln y)'$ y $h(y) = I_{(0,\infty)}(y)$.*

La representación en forma de la familia exponencial no es única, pues nótese que en el ejemplo anterior también se puede tomar $\eta(\theta) = (\frac{1}{\theta}, k-1)'$ y $T(y) = (-y, \ln y)'$.

Las distribuciones normal, gamma, exponencial, chi-cuadrado, beta, Bernoulli, binomial, binomial negativa, multinomial, Poisson y geométrica todas pertenecen

a la familia exponencial. También la distribución Weibull pertenecen a la familia exponencial cuando el parámetro de forma es conocida. Por el otro lado, las distribuciones Cauchy, Laplace, uniforme y Weibull cuando el parámetro de forma es desconocida no pertenecen a la familia exponencial.

8.2 Parametrización y funciones de vínculo

Para la formulación de cualquier modelo lineal generalizado se parte de las siguientes igualdades:

$$\eta_i = \mathbf{X}_i'\boldsymbol{\beta} \tag{8.2.1}$$
$$\theta_i = E(Y_i) \tag{8.2.2}$$
$$\eta_i = g(\theta) \tag{8.2.3}$$

- El primer paso en el modelamiento bayesiano de este tipo de modelos es calcular la inversa de la función $g(\cdot)$, puesto que la verosimilitud de los datos debe estar en términos de $\theta_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{X}_i'\boldsymbol{\beta})$. De esta forma, se garantiza que la verosimilitud esté en términos de los parámetros de interés $\boldsymbol{\beta}$.
- En segundo lugar, se debe proponer una distribución previa para $\boldsymbol{\beta}$; en este libro siempre trabajaremos con distribuciones previa normales, aunque se recalca que no necesariamente deben especificarse distribuciones previa normales.
- Por último, se escribe la distribución posterior, que por lo general no tiene una forma cerrada, y se utilizan métodos numéricos y de simulación para hacer inferencias posterior sobre $\boldsymbol{\beta}$.

8.2.1 A posterior generalizada

8.3 Modelo Normal

El único caso en donde la distribución posterior tiene una forma conocida se presenta cuando la función de vínculo es la función identidad. Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es un conjunto de variables aleatorias intercambiables cada una con distribución normal estándar de media θ_i y varianza σ^2. Luego, cuando la función de vinculo toma la siguiente forma

$$\eta_i = g(\theta_i) = \theta_i = \mathbf{X}_i'\boldsymbol{\beta}. \tag{8.3.1}$$

En este caso, cuando la distribución previa de los parámetros de interés es normal multivariante, entonces se llega fácilmente al caso del modelo lineal bayesiano. Como se puede comprender en el capítulo XXXXX, en la verosimilitud de los

datos no necesariamente se conoce la matriz de covarianzas; de esta forma, se tienen varios acercamientos para que la distribución posterior final del vector de parámetros de interés $\boldsymbol{\beta}$ sea conjugada.

8.4 Modelo Bernoulli con vínculo logístico

Este caso es típico en donde la variable respuesta sólo toma dos valores, uno en caso de un evento exitoso y cero, cuando se presenta un fracaso. Se supone que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es un conjunto de variables aleatorias intercambiables cada una con distribución bernoulli de parámetro θ_i. En este apartado, se considera que la función de vínculo[1] es logística, de tal manera que

$$\eta_i = g(\theta_i) = logit(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) \tag{8.4.1}$$

Fácilmente se encuentra que la función inversa para $g(\cdot)$, está dada por

$$\theta_i = g^{-1}(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$$

Notando que $\eta_i = \mathbf{X}_i'\boldsymbol{\beta}$ y siguiendo con el modelamiento, se tiene que la verosimilitud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i} \\ p(\mathbf{Y} \mid \boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})}\right)^{y_i} \left(1-\left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})}\right)\right)^{1-y_i} \end{aligned} \tag{8.4.2}$$

Suponga que la distribución previa para $\boldsymbol{\beta}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

En la vida práctica, cuando se desconoce el comportamiento de los parámetros de interés, es común proponer que todas las entradas del vector de medias previa $\mathbf{b}$ sean nulas y que la diagonal de la matriz de covarianzas previa $\mathbf{B}$ sea guarismos grandes para reflejar mediante la dispersión, el desconocimiento del problema que se trate; si se quiere, también es posible que las entradas afuera de la diagonal sean nulas, lo cual induce que los parámetros se consideren independientes previa.

De esta manera, la distribución posterior toma la siguiente forma.

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) \propto \prod_{i=1}^{n} &\left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})}\right)^{y_i} \left(1-\left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})}\right)\right)^{1-y_i} \\ &\times \exp\left\{\frac{-1}{2}(\boldsymbol{\beta}-\mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta}-\mathbf{b})\right\} \end{aligned}$$

[1] **?** afirma que es inconveniente fijar la función de vínculo como $\eta_i = \theta_i$, puesto que se debe garantizar que $0 \leq \theta_i \leq 1$.

La anterior expresión no tiene una forma cerrada y no es sencillo, en primera instancia, simular observaciones u obtener inferencias posterior. Sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$.

Algoritmo de Gibbs

- Fijar valores iniciales para el vector de parámetros de interés; por ejemplo, estos valores iniciales pueden estar dados por $\boldsymbol{\beta}^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_q^0)'$.
- Para simular un valor de la distribución posterior condicional para β_0 se tiene en cuenta que

$$p(\beta_1 \mid \beta_2, \ldots, \beta_q, \mathbf{D}) \propto (\beta_1, \underbrace{\beta_2 \ldots, \beta_q}_{fijos} \mid \mathbf{D})$$

  Por tanto, utilizando los valores iniciales $\beta_2^0, \ldots, \beta_q^0$ en la distribución conjunta y dejando a β_1 como una variable, entonces es posible utilizar el método de la grilla para simular una nueva observación, β_1^1, de esta distribución univariada. Este nuevo valor β_1^1 reemplaza al valor β_1^0.
- Realizar el anterior procedimiento para simular una nueva observación β_2^1 de

$$p(\beta_2 \mid \beta_1, \beta_3 \ldots, \beta_q, \mathbf{D}) \propto (\beta_2, \underbrace{\beta_1, \beta_3 \ldots, \beta_q}_{fijos} \mid \mathbf{D})$$

- Repetir este proceso hasta que todos los componentes del vector de valores iniciales $\boldsymbol{\beta}^0$ sean reemplazados en su totalidad por un nuevo vector de valores dado por $\boldsymbol{\beta}^1 = (\beta_1^1, \beta_2^1, \ldots, \beta_q^1)'$.
- ...
- Simular un número grande de vectores $\boldsymbol{\beta}$ hasta obtener convergencia. Al final, todos los vectores simulados son considerados como realizaciones de la distribución multivariada posterior conjunta del vector de parámetros de interés dada por el anterior resultado.

Una vez que se tengan las observaciones simuladas de la distribución posterior del vector $\boldsymbol{\beta}$, es posible realizar todo tipo de inferencias posterior para $\boldsymbol{\beta}$.

Ejemplo 8.4.1. ?, *p. 220 consideran una aplicación ecológica en donde la variable predictora X es la distancia hasta el borde del bosque y la variable respuesta Y es la variable binaria indicando la copresencia de dos especies en un sector forestal en donde se seleccionó una muestra de $n = 603$ locaciones. Suponga que se quiere ajustar el siguiente modelo de regresión logística.*

$$logit(p_i) = \beta_0 + \beta_1 X_i \qquad i = 1, \ldots, n.$$

Considerando distribuciones previa en el sentido vague para los dos parámetros del modelo, entonces la sintaxis en BUGS para ajustar este modelo está dada por el siguiente comando computacional

```
model{
for(i in 1:n){
z[i]~dbern(p[i])
logit(p[i])<-beta0+beta1*(x[i]-mean(x[]))
}
beta0~dflat()
beta1~dflat()
}
```

En primer lugar se establecieron tres conjuntos de valores iniciales para los dos coeficientes de tal forma que fueran dispersos para asegurar que los resultados sean validos puesto que con cada una de las especificaciones de valores iniciales, se debe llegar a la misma distribución posterior. Los valores iniciales para los dos parámetros están dados en la siguiente estructura computacional de BUGS

```
list(beta0=1, beta1=1)
list(beta0=0, beta1=10)
list(beta0=-100, beta1=200)
```

Después de un periodo de calentamiento de las cadenas, comprendido por 1000 iteraciones, se realizaron 30 mil iteraciones más para obtener la distribución posterior en cada uno de los tres casos anteriores. Es decir, en los cálculos de simulación de la densidad posterior sólo intervinieron los valores simulados desde la iteración 1001 hasta la iteración 30001. A continuación, se reproducen las salidas para el coeficiente β_0.

```
Caso    node    mean      sd MC error     2.5%  median   97.5%
1       beta0 -0.4047 0.09085 5,14E-01 -0.5845 -0.4037 -0.2279
2       beta0 -0.4043 0.09038 5,50E-01 -0.5808 -0.4044 -0.2263
3       beta0 -0.4047 0.0906 5,57E-01 -0.5819 -0.4039 -0.2284
```

Para cada una de las tres cadenas, el comportamiento posterior es similar. Lo propio sucede también con el parámetro β_1*, como se ilustra a continuación*

```
Caso  node mean     sd MC error    2.5% median  97.5%
1       beta1 -1.72 0.1966 0.001190    -2.109 -1.716 -1.344
2       beta1 -1.72 0.1971 0.001057 -2.111 -1.717 -1.342
3       beta1 -1.72 0.1981 0.001137 -2.121 -1.717 -1.340
```

A continuación se muestra el comportamiento de las cadenas para ambos parámetros. Como conclusión, podemos afirmar que la distribución posterior para los parámetros de interés no depende de los valores iniciales en la cadena, por lo tanto la convergencia se garantiza y así mismo las conclusiones obtenidas a partir de este análisis.

Como estamos interesados en conocer si la proximidad al eje del bosque es un predictor significativo de la copresencia de las especies. En otras palabras, estamos interesados en conocer si el parámetro β_1 *es significativo. A primera vista,*

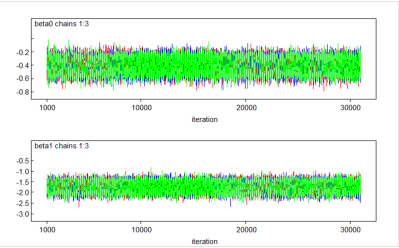

Figura 8.1: *Convergencia de las cadenas para los coeficientes de la regresión*

fijándonos en los valores de estimación puntual del parámetro, estaríamos tentados a responder afirmativamente. Lo anterior se corrobora mediante los siguientes criterios:

- *El intervalo de credibilidad al 95 % no contiene al valor cero: la interpretación de las regiones de credibilidad bayesianas difiere de la interpretación de las regiones de confianza frecuentista. La primera se refiere a la probabilidad de que el verdadero valor de β_1 esté contenido en el intervalo. Por tanto la probabilidad de que β_1 difiera de cero es mayor que 0.95.*
- *Probabilidad de inclusión: Siguiendo el razonamiento de* **?**, *p. 89, se crea un nuevo parámetro π_1, que denota la probabilidad de incluir a la variable X en el modelo. De esta manera, el nuevo modelo que redefine la estructura de la media de la variable $logit(p_i)$ está dado por*

$$logit(p_i) = \beta_0 + \pi_1\beta_1 X_i \qquad i = 1, \ldots, n.$$

La distribución previa para π_1 seguirá el patrón de una distribución no informativa; por lo tanto será $Bernoulli(0.5)$. En términos de la inclusión de X en el modelo, estamos interesados en conocer el comportamiento estructural de π_1 a posterior, el cual se muestra a continuación.

Node	mean	sd	MC error	2.5%	median	97.5%
pi	1.0	0.0	5.68E-13	1.0	1.0	1.0

De lo anterior se concluye que el parámetro siempre tomó el mismo valor igual a uno para todas las iteraciones de la cadena, mostrando una fuerte y rotunda evidencia a favor de la inclusión de la variable X. A continuación se muestra la sintaxis empleada para el ajuste de este modelo en BUGS.

```
model{
```

```
for(i in 1:n){
y[i]~dbern(p[i])
logit(p[i])<-beta0+pi*beta1*(x[i])
}
beta0~dflat()
beta1~dflat()
pi~dbern(0.5)
}
```

- *DIC: Este es un criterio de información bayesiana, que se puede considerar como una generalización del criterio de información de AKAIKE. Para el modelo ajustado con la variable proximidad al eje del bosque, este criterio da un valor de 728.4. Por otro lado, el DIC para el modelo que no incluye esta variable dado por*

$$logit(p_i) = \beta_0$$

es de 820.9. Este último valor muestra que el valor del DIC para el modelo que sí contiene a X es mucho menor con respecto al valor del DIC del modelo que no tiene en cuenta a esta variable. Como conclusión, el modelo ajusta mucho mejor cuando X es considerada como la variable predictora.

Por otro lado, este mismo modelo se puede implementar con el algoritmo de Metropolis-Hastings en R. En general, no existe un paquete que realice este tipo de ajustes (según el conocimiento de los autores), por tanto es necesario crear una función en R.

Para conseguir similares resultados que los obtenidos usando BUGS, es necesario replicar el ejercicio bajo las mismas condiciones; es decir, utilizando distribuciones previa en el sentido vague, que serán traducidas como distribuciones uniformes. De esta manera se tienen los siguientes comentarios:

- *La distribución posterior para el vector de coeficientes de regresión es proporcional a la verosimilitud de los datos.*
- *La función generadora de valores candidatos en la iteración t para el algoritmo (comúnmente conocida como Jumping distribution) corresponde a una densidad normal bivariada para β_0 y para β_1 con vector de medias $(\beta_0(t-1), \beta_1(t-1))'$ y matriz de varianzas diagonal $(0.01, 0.01)$. Esto implica que β_0 y β_1 se consideran condicionalmente independientes dados los valores de los coeficientes en la iteración anterior. Como es bien sabido, bajo estas condiciones la distribución normal bivariada se convierte en el producto de las dos distribuciones marginales.*
- *Como la base de datos consta de un gran volumen de información, se procede a reescribir la verosimilitud, aplicando logaritmos, puesto que de lo contrario surgen problemas numéricos debido a que la verosimilitud para cada y_i toma valores menores que uno y al multiplicarlos más de 600 veces (el tamaño de la muestra) esta función se convierte en cero bajo el ambiente computacional de R.*

La función construida es la siguiente:

```
MHlogit <- function(beta0, beta1, x, y, Nsim){

# Creación de la distribución posterior.
# Toma la forma de la verosimilitud pues las apriori son vagues
post <- function(beta0, beta1, x, y) {
lp <- beta0 + beta1 * x
p <- exp(lp)/(1 + exp(lp))
#modificación para evitar problemas numéricos
sum(y*log(p)+(1-y)*log(1-p))
}

ind <-rep(0,Nsim)

betas <- matrix(NA,ncol=2,nrow=Nsim)
for(k in 1:Nsim){
betas[k,1] <- beta0
betas[k,2] <- beta1

#Genera un valor candidato
beta0.can <- rnorm(1,beta0,0.1)
beta1.can <- rnorm(1,beta1,0.1)

#Jumping distribution
q1 <- dnorm(beta0,beta0.can,0.1)*dnorm(beta1,beta1.can,0.1)
q2 <- dnorm(beta0.can,beta0,0.1)*dnorm(beta1.can,beta1,0.1)

#posterior
p1 <- post(beta0.can,beta1.can, x, y)
p2 <- post(beta0,beta1, x, y)

#Aceptación beta0
#modificación para evitar problemas numéricos
T.val <- min(1,(exp(p1-p2)*(q1/q2)))
u <- runif(1)
if (u<= T.val){
beta0=beta0.can
beta1=beta1.can
ind[k]<-1
}  }
return(list(ind=ind,betas=betas))
}
```

La implementación para este modelo en particular está dada por el siguiente código, que describe que el procedimiento tuvo 50 mil iteraciones y para efectos de estimación se conservan las últimas 30 mil.

```
x.new<-x-mean(x)
# Valores iniciales
beta0=1
beta1=1

res <- MHlogit(beta0,beta1,x.new,y,Nsim=50000)

mean(res$betas[-(1:20000),1])
[1] -0.4035946

mean(res$betas[-(1:20000),2])
[1] -1.71853
```

Como se puede observar, los resultados del algoritmo Metropolis-Hastings en R coinciden plenamente con los obtenidos usando el algoritmo de Gibbs en BUGS. La siguiente gráfica muestra la evolución de las cadenas para los dos coeficientes de regresión.

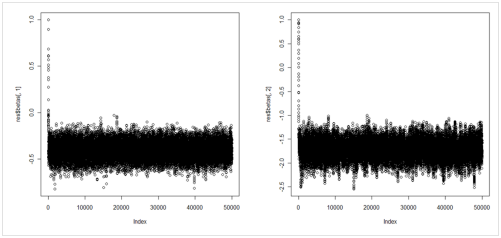

Figura 8.2: *Convergencia de las cadenas para los coeficientes de la regresión usando el algoritmo de Metropolis-Hastings*

Por ultimo, la razón de aceptación fue del 61.5 % así como lo demuestra la siguiente salida

```
sum(res$ind)/50000
[1] 0.61528
```

8.5 Modelo Binomial con vínculo logístico

En este caso la variable respuesta representa conteos de éxitos que se tuvieron en un conjunto de distintos experimentos. Se supone que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es un conjunto de variables aleatorias intercambiables cada una con distribución binomial de parámetro θ_i. Con la función de vínculo logística, considerando las expresiones XXXXX, XXXXX, y notando que $\eta_i = \mathbf{X}_i'\boldsymbol{\beta}$, se tiene que la verosimilitud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} \binom{n}{y_i} \theta_i^{y_i} (1-\theta_i)^{n-y_i} \\ &= \prod_{i=1}^{n} \binom{n}{y_i} \left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})} \right)^{y_i} \left(1 - \left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})} \right) \right)^{n-y_i} \end{aligned} \tag{8.5.1}$$

Suponga que la distribución previa para $\boldsymbol{\beta}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

De esta manera, la distribución posterior toma la siguiente forma.

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})} \right)^{y_i} \left(1 - \left(\frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})} \right) \right)^{n-y_i} \\ \times \exp\left\{ \frac{-1}{2} (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b}) \right\} \end{aligned}$$

Una vez más, la anterior expresión no tiene una forma cerrada; sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$.

Ejemplo 8.5.1. *Gelman raticas y laboratorios*

8.6 Modelo Binomial con vínculo Probit

Para la distribución binomial, también es posible trabajar con la función de vínculo probit, de tal manera que

$$\eta_i = g(\theta_i) = probit(\theta_i) = \Phi^{-1}(\theta_i) \tag{8.6.1}$$

Fácilmente se encuentra que la función inversa para $g(\cdot)$, está dada por

$$\theta_i = g^{-1}(\eta_i) = \Phi(\eta_i) \tag{8.6.2}$$

Notando que $\eta_i = \mathbf{X}_i'\boldsymbol{\beta}$ y siguiendo con el modelamiento, se tiene que la verosimilitud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{n} \binom{n}{y_i} \theta_i^{y_i} (1-\theta_i)^{n-y_i} \\ &= \prod_{i=1}^{n} \binom{n}{y_i} (\Phi(\mathbf{X}_i'\boldsymbol{\beta}))^{y_i} (1-\Phi(\mathbf{X}_i'\boldsymbol{\beta}))^{n-y_i} \end{aligned} \tag{8.6.3}$$

Suponiendo que la distribución previa para $\boldsymbol{\beta}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

De esta manera, la distribución posterior toma la siguiente forma.

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) \propto \prod_{i=1}^{n} &(\Phi(\mathbf{X}_i'\boldsymbol{\beta}))^{y_i} (1-\Phi(\mathbf{X}_i'\boldsymbol{\beta}))^{n-y_i} \\ &\times \exp\left\{ \frac{-1}{2} (\boldsymbol{\beta}-\mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta}-\mathbf{b}) \right\} \end{aligned}$$

Una vez más, la anterior expresión no tiene una forma cerrada; sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$.

Ejemplo 8.6.1. *Ejemplo econométrico del libro de R*

8.7 Modelo Poisson con vínculo logaritmico

Este caso es típico en donde la variable respuesta toma valores enteros positivos, denotando el número de ocurrencias de un evento. Se supone que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es un conjunto de variables aleatorias intercambiables cada una con distribución Poisson de parámetro θ_i. Considerando la función de vínculo logaritmica, de tal manera que

$$\eta_i = g(\theta_i) = \log(\theta_i) \tag{8.7.1}$$

Teniendo en cuenta que la función inversa para $g(\cdot)$, está dada por

$$\theta_i = g^{-1}(\eta_i) = \exp(\eta_i) \tag{8.7.2}$$

Notando que $\eta_i = \mathbf{X}_i'\boldsymbol{\beta}$ y siguiendo con el modelamiento, se tiene que la verosi-

militud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{n} \frac{\exp(-\theta_i)\theta_i^{y_i}}{y_i!} \\ &= \prod_{i=1}^{n} \frac{\exp(-\exp(\mathbf{X}_i'\boldsymbol{\beta}))(\exp(\mathbf{X}_i'\boldsymbol{\beta}))^{y_i}}{y_i!} \\ &= \frac{1}{\prod_{i=1}^{n} y_i!} \exp\left\{\sum_{i=1}^{n} (y_i \mathbf{X}_i'\boldsymbol{\beta} - \exp(\mathbf{X}_i'\boldsymbol{\beta}))\right\} \end{aligned} \tag{8.7.3}$$

Suponga que la distribución previa para $\boldsymbol{\beta}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

De esta manera, la distribución posterior toma la siguiente forma.

$$p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) \propto \exp\left\{\sum_{i=1}^{n} (y_i \mathbf{X}_i'\boldsymbol{\beta} - \exp(\mathbf{X}_i'\boldsymbol{\beta})) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b})\right\}$$

Una vez más, la anterior expresión no tiene una forma cerrada; sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$.

Ejemplo 8.7.1. *Ejemplo*

8.8 Modelos de sobredispersión

Cuando se tienen datos provenientes de conteos, se pueden tomar la distribución binomial o la distribución Poisson para modelarlos. Sin embargo, si la varianza de los datos es mucho mayor que su media, sería difícil argumentar que puedan provenir de la distribución binomial - puesto que para esta distribución siempre la media es más grande que su varianza - o que provengan de una distribución Poisson - puesto que para esta distribución siempre su media es igual a su varianza. El anterior fenómeno se conoce con el nombre de sobredispersión.

8.8.1 Modelo Binomial negativa con vínculo log-complementario

Se supone que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es un conjunto de variables aleatorias intercambiables cada una con distribución Binomial negativa de parámetro θ_i. Considerando la función de vínculo log-complementario, de tal manera que

$$\eta_i = g(\theta_i) = \log(1 - \theta_i) \tag{8.8.1}$$

Teniendo en cuenta que la función inversa para $g(\cdot)$, está dada por

$$\theta_i = g^{-1}(\eta_i) = 1 - \exp(\eta_i) \tag{8.8.2}$$

Notando que $\eta_i = \mathbf{X}_i'\boldsymbol{\beta}$ y siguiendo con el modelamiento, se tiene que la verosimilitud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta, r) &= \prod_{i=1}^{n} \frac{\Gamma(r+y_i)}{y_i!\Gamma(r)} \theta^r (1-\theta)^{1-y_i} \\ &= \prod_{i=1}^{n} \frac{\Gamma(r+y_i)}{y_i!\Gamma(r)} (1-\exp(\mathbf{X}_i'\boldsymbol{\beta}))^r (\exp(\mathbf{X}_i'\boldsymbol{\beta}))^{1-y_i} \end{aligned} \tag{8.8.3}$$

Suponga que la distribución previa para $\boldsymbol{\beta}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

De esta manera, la distribución posterior toma la siguiente forma.

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, r) \propto \prod_{i=1}^{n} (1-\exp(\mathbf{X}_i'\boldsymbol{\beta}))^r \\ \times \exp\left\{ \sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\beta}(1-y_i) - \frac{1}{2}(\boldsymbol{\beta}-\mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta}-\mathbf{b}) \right\} \end{aligned}$$

Una vez más, la anterior expresión no tiene una forma cerrada; sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$.

Ejemplo 8.8.1. *Ejemplo*

8.8.2 Regresión Poisson jerárquica

mirar en gelman

8.9 Modelo Gamma con vínculo recíproco

En este apartado se considera que la variable respuesta toma valores positivos y reales; lo anterior es muy conveniente a la hora de modelar los distintos tipos de variables económicas y financieras. Asuma que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es un conjunto de variables aleatorias intercambiables cada una con distribución Gamma de parámetros (α, β). Antes de continuar, y siguiendo las ideas de **?**, es conveniente reparametrizar la distribución haciendo $\beta = \alpha/\theta$; luego, la forma funcional de la distribución de una variable Y, está dada por:

$$p(y \mid \theta, \alpha) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\theta}\right)^{\alpha} y^{\alpha-1} \exp\left\{\frac{y\alpha}{\theta}\right\} \tag{8.9.1}$$

Con esta transformación, $E(Y) = \theta$ y $Var(Y) = \theta^2/\alpha$. Por otro lado, considerando la función de vínculo recíproco, de tal manera que

$$\eta_i = g(\theta_i) = \theta^{-1} \tag{8.9.2}$$

Teniendo en cuenta que la función inversa para $g(\cdot)$, está dada por

$$\theta_i = g^{-1}(\eta_i) = \eta_i^{-1} \tag{8.9.3}$$

Notando que $\eta_i = \mathbf{X}_i'\boldsymbol{\beta}$ y siguiendo con el modelamiento, se tiene que la verosimilitud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\beta}, \alpha) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\eta_i^{-1}}\right)^{\alpha} y_i^{\alpha-1} \exp\left\{\frac{y_i\alpha}{\eta_i^{-1}}\right\} \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \left(\mathbf{X}_i'\boldsymbol{\beta}\alpha\right)^{\alpha} y_i^{\alpha-1} \exp\left\{\mathbf{X}_i'\boldsymbol{\beta} y_i \alpha\right\} \end{aligned} \tag{8.9.4}$$

Suponiendo que la distribución previa para $\boldsymbol{\beta}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

De esta manera, la distribución posterior toma la siguiente forma.

$$p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \alpha) \propto \prod_{i=1}^{n} \left(\mathbf{X}_i'\boldsymbol{\beta}\alpha\right)^{\alpha} \exp\left\{\mathbf{X}_i'\boldsymbol{\beta} y_i \alpha - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b})\right\}$$

Una vez más, la anterior expresión no tiene una forma cerrada; sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$.

? anota que es muy importante tener claro que utilizando la función canónica de vínculo recíproco $g(\theta) = \theta^{-1}$, pueden aparecer problemas con los métodos de Monte Carlo puesto que el parámetro θ debe ser positivo; de esta manera, y dependiendo de las covariables, puede suceder que las estimaciones posterior para $\boldsymbol{\beta}$ induzcan valores negativos para θ_i. Para evitar este problema, es recomendable trabajar con otras funciones de vínculo, como la logarítmica, o incluso la función lineal (**?**).

Ejemplo 8.9.1. *Ejemplo variable ingreso Marco y Lucy*

8.10 Regresión Beta con vínculo log-log-complementario

Este tipo de distribuciones es ideal cuando la variable respuesta representa porcentajes, proporciones que toma valores entre cero y uno. Se supone que $\mathbf{Y} =$

$\{Y_1, \ldots, Y_n\}$ es un conjunto de variables aleatorias intercambiables cada una con distribución Beta de parámetros (α, β). Al igual de en la anterior sección, es conveniente reparametrizar siguiendo la propuesta de **?** al definir $\theta = \alpha/(\alpha + \beta)$ y $\phi = \alpha + \beta$; entonces, la nueva estructura de la distribución de una variable Y, está dada por:

$$p(y \mid \theta, \phi) = \frac{\Gamma(\phi)}{\Gamma(\theta\phi)\Gamma(\phi - \theta\phi)} y^{\theta\phi - 1}(1-y)^{(\phi - \phi\theta) - 1} \tag{8.10.1}$$

Con esto, $E(Y) = \theta$ y $Var(Y) = \theta(1-\theta)/(1+\phi)$. Por otro lado, se considera la función de vínculo log-log-complementario, de tal manera que

$$\eta_i = g(\theta_i) = \log(-\log(1-\theta)) \tag{8.10.2}$$

Teniendo en cuenta que la función inversa para $g(\cdot)$, está dada por

$$\theta_i = g^{-1}(\eta_i) = 1 - \exp(-\exp(\eta_i)) \tag{8.10.3}$$

Notando que $\eta_i = \mathbf{X}_i'\boldsymbol{\beta}$, se tiene que la verosimilitud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi) &= \prod_{i=1}^{n} \frac{\Gamma(\phi)}{\Gamma(\theta\phi)\Gamma(\phi - \theta\phi)} y_i^{\theta\phi - 1}(1-y_i)^{(\phi - \phi\theta) - 1} \\ &= \prod_{i=1}^{n} \frac{\Gamma(\phi)}{\Gamma(\theta\phi)\Gamma(\phi - \theta\phi)} y_i^{(1 - \exp(-\exp(\mathbf{X}_i'\boldsymbol{\beta})))\phi - 1} \\ &\qquad \times (1-y_i)^{(\phi - \phi(1 - \exp(-\exp(\mathbf{X}_i'\boldsymbol{\beta})))) - 1} \end{aligned} \tag{8.10.4}$$

Suponiendo que la distribución previa para $\boldsymbol{\beta}$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

De esta manera, la distribución posterior toma la siguiente forma.

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \alpha) \propto \prod_{i=1}^{n} & y_i^{(1 - \exp(-\exp(\mathbf{X}_i'\boldsymbol{\beta})))\phi - 1}(1-y_i)^{(\phi - \phi(1 - \exp(-\exp(\mathbf{X}_i'\boldsymbol{\beta})))) - 1} \\ & \times \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b}) \right\} \end{aligned}$$

Una vez más, la anterior expresión no tiene una forma cerrada; sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$.

? anota que es muy importante tener claro que utilizando la función canónica de vínculo recíproco $g(\theta) = \theta^{-1}$, pueden aparecer problemas con los métodos de Monte Carlo puesto que el parámetro θ debe ser positivo; de esta manera, y dependiendo de las covariables, puede suceder que las estimaciones posterior para $\boldsymbol{\beta}$ induzcan valores negativos para θ_i. Para evitar este problema, es recomendable trabajar con otras funciones de vínculo, como la logarítmica, o incluso la función lineal (**?**).

Ejemplo 8.10.1. *Ejemplo porcentajes*

8.11 Regresión multinomial con vínculo logit

La distribución multinomial es una extensión de la distribución binomial; de esta forma se está interesa en conocer el número de éxitos en cada una de las categorías de la variable. **?** considera un ejemplo en donde la variable respuesta tiene tres categorías: grupo de edad, nivel educativo y nivel socioeconómico. La respuesta está dada en términos del número de personas que se identifican como pertenecientes al partido político republicano. Cuando se considera esta situación en diferentes regiones de los Estados Unidos, entonces se tiene un conjunto de vectores aleatorias con distribución multinomial.

Generalizando lo anterior, se considera $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ un conjunto de vectores aleatorios intercambiables cada una con distribución multinomial de parámetros $(n_i, \boldsymbol{\theta}_i)$. Note que todos lo vectores tienen K categorias; en particular, el i-ésimo vector del conjunto se define como $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})'$, el vector de parámetros de interés es $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})$, $\sum_{k=1}^{K} \theta_{ik} = 1$, cual debe cumplir con la restricción $n_i = \sum_{k=1}^{K} Y_{ik}$, y por ende se tiene que la distribución, condicional a n_i, de $\mathbf{Y}_i$ sigue una distribución multinomial tal que:

$$p(\mathbf{Y}_i \mid n_i, \boldsymbol{\theta}_i) = \binom{n_i}{y_{i1}, \ldots, y_{iK}} \prod_{k=1}^{K} \theta_{ik}^{y_{ik}} \tag{8.11.1}$$

Con base en lo anterior, la verosimilitud de los datos se tiene mediante la siguiente expresión

$$p(\mathbf{Y} \mid n, \boldsymbol{\theta}) = \prod_{i=1}^{n} \binom{n_i}{y_{i1}, \ldots, y_{iK}} \prod_{k=1}^{K} \theta_{ik}^{y_{ik}} \tag{8.11.2}$$

Antes de proseguir con el modelamiento bayesiano, es útil notar que en este caso el vínculo no es un vector sino una matriz que responde a una relación lineal entre las covariables y una matriz de coeficientes de regresión, como se puede ver a continuación.

$$\begin{bmatrix} \eta_{11} & \eta_{12} & \cdots & \eta_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{n1} & \eta_{n2} & \cdots & \eta_{nK} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qK} \end{bmatrix}$$

Es decir,

$$\boldsymbol{\eta} = X'\boldsymbol{\beta} \tag{8.11.3}$$

Lo anterior conlleva a que $\eta_{ik} = \mathbf{X}_i'\boldsymbol{\beta}_k$, donde $\mathbf{X}_i$ es la i-ésima fila de la matriz X y $\boldsymbol{\beta}_k$ es la k-ésima columna de la matriz $\boldsymbol{\beta}$. Ahora, tomando como línea de base la primera columnna de la matriz $\boldsymbol{\eta}$, y utilizando la función de vinculo logístico, se tiene que para los elementos en las restantes columnnas de $\boldsymbol{\eta}$,

$$\eta ik = g(\theta_{ik}) = logit(\theta_{ik}) = \log\left(\frac{\theta_{ik}}{1 - \sum_{k=2}^{K} \theta_{ik}}\right) \tag{8.11.4}$$

Con un poco de álgebra se comprueba que la función inversa para $g(\cdot)$ está dada por la siguiente expresión

$$\theta_{ik} = g^{-1}(\eta_{ik}) = \frac{\exp(\eta_{ik})}{1 + \sum_{k=2}^{K} \exp(\eta_{ik})} \qquad \forall k = 2, \ldots, K$$

y para la primera columna, se tiene que la función inversa para $g(\cdot)$ es

$$\theta_{ik} = g^{-1}(\eta_{ik}) = \frac{1}{1 + \sum_{k=2}^{K} \exp(\eta_{ik})}$$

Luego, para todas las columnas, se se tiene que la función inversa para $g(\cdot)$ es

$$\theta_{ik} = g^{-1}(\eta_{ik}) = \frac{\exp(\eta_{ik})}{1 + \sum_{k=2}^{K} \exp(\eta_{ik})} \quad \text{con } \eta_{i1} = 0 \;\; \forall k = 1, \ldots, K \tag{8.11.5}$$

Notando que $\eta_{ik} = \mathbf{X}_i' \boldsymbol{\beta}_k$, se tiene que la verosimilitud de los datos toma la siguiente forma

$$\begin{aligned} p(\mathbf{Y} \mid n, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \binom{n_i}{y_{i1}, \ldots, y_{iK}} \prod_{k=1}^{K} \theta_{ik}^{y_{ik}} \\ p(\mathbf{Y} \mid n, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \binom{n_i}{y_{i1}, \ldots, y_{iK}} \prod_{k=1}^{K} \left(\frac{\exp(\mathbf{X}_i' \boldsymbol{\beta}_k)}{1 + \sum_{k=2}^{K} \exp(\mathbf{X}_i' \boldsymbol{\beta}_k)} \right)^{y_{ik}} \end{aligned}$$

Suponiendo que la distribución previa para $\boldsymbol{\beta}_k$ está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta}_k \sim Normal_q(\mathbf{b}_k, \mathbf{B}_k)$$

De esta manera, la distribución posterior para el k ésimo vector $\boldsymbol{\beta}_k$ toma la siguiente forma.

$$\begin{aligned} p(\boldsymbol{\beta}_k \mid \mathbf{Y}, \mathbf{X}, n) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i' \boldsymbol{\beta}_k)}{1 + \sum_{k=2}^{K} \exp(\mathbf{X}_i' \boldsymbol{\beta}_k)} \right)^{y_{ik}} \\ \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b}) \right\} \end{aligned}$$

Una vez más, la anterior expresión no tiene una forma cerrada; sin embargo, con ayuda de la técnica del condicionamiento sucesivo, el algoritmo de Gibbs y el método de simulación de la grilla es posible obtener fácilmente observaciones multivariantes provenientes de la distribución posterior del vector $\boldsymbol{\beta}$. Nótese que por la restricción $\eta_{i1} = 0$, es necesario fijar el primer valor de $\boldsymbol{\beta}_k$ igual a cero para todo k.

Ejemplo 8.11.1. *Ejemplo aligators*

8.12 Técnicas de aproximación y algoritmos iterativos

8.12.1 Técnica de la aproximación de la log-verosimilitud a la normal

Gelman 422 y 423

8.12.2 Técnica de la aproximación de la verosimilitud a la normal

Gammerman pg 217 Algoritmo IRLS

9 Modelos longitudinales

pg 335 de Carlin 2009

10 Modelo lineal dinámico

10.1 Representación de estado-espacio y filtro de Kalman

10.1.1 Modelos autorregresivos

10.1.2 Promedios móviles

10.1.3 Modelos ARIMA

10.1.4 Modelos autorregresivos

10.2 Modelos bayesianos dinámicos

Los modelos lineales dinámicos son especificados por las dos siguientes ecuaciones,

$$y_t = \mathbf{F}_t'\boldsymbol{\beta}_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_t^2) \tag{10.2.1}$$

$$\boldsymbol{\beta}_t = \mathbf{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \omega_t \sim N_p(\mathbf{0}, \mathbf{W}_t) \tag{10.2.2}$$

donde $\{y_t\}$ es una serie de tiempo observada univariada, condicionalmente independiente dado $\boldsymbol{\beta}_t$ y σ_t^2, $\mathbf{F}_t$ es un vector de variables aleatorias de dimensión $p \times 1$, y $\boldsymbol{\beta}_t$ es el vector de coeficientes de regresión en el tiempo t, también llamado el vector de estado, $\mathbf{G}_t$ es de dimensión $p \times p$. Los errores ϵ_t y $\boldsymbol{\omega}_t$ son independientes, con varianza σ_t^2 y $\mathbf{W}_t$, respectivamente. Adicionalmente se supone que $\boldsymbol{\beta}_1 \sim N_p(\mathbf{a}_1, \mathbf{R}_1)$. La primera ecuación se denomina la ecuación de observación, y la segunda, ecuación de sistema.

Algunos caso especiales de los modelos lineales dinámicos son los modelos de regresión dinámico que se dan cuando $\mathbf{G}_t = \mathbf{I}_p$ para todo t. Y los modelos de regresión clásicos o estáticos también es un caso particular de los modelos lineales dinámicos con $\mathbf{G}_t = \mathbf{I}_p$ y $\mathbf{W}_t = 0$ para todo t.

Ejemplo 10.2.1. *En general, los modelos autoregresivos $AR(p)$, se puede escribir como un modelo lineal dinámico. Considere un modelo $AR(2)$ dado por $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$, al definir $\boldsymbol{\beta}_t = \begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix}$, se tiene que*

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix},$$

$$\begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \end{pmatrix} + \begin{pmatrix} e_t \\ 0 \end{pmatrix}$$

Ejemplo 10.2.2. *Gamerman (2003) afirma que el modelo de crecimiento lineal dado por*

$$y_t = \beta_{1t} + \epsilon_t \quad \epsilon_t \sim N(0, \sigma_t^2)$$
$$\beta_{1t} = \beta 1, t-1 + \beta_{2t} + \omega_{1t}$$
$$\beta_{2t} = \beta_{2,t-1} + \omega_{2t}$$

es un modelo lineal dinámico con $\mathbf{F}_t = (1, 0)$, $\mathbf{G}_t = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ *y* $\boldsymbol{\omega}_t = (\omega_{1t}, \omega_{2t})' \sim N_2(\mathbf{0}, \mathbf{W}_t)$.

El objetivo del análisis es encontrar la distribución posterior de los coeficientes de regresión $\boldsymbol{\beta}_t$ para todo $t = 1, \ldots, n$. Esto se lleva a cabo usando procedimientos recursivos que constan de los procesos del filtro del suavizamiento que se describen a continuación. Se supone que σ_t^2, $\mathbf{W}_t$, $\mathbf{F}_t$ y $\mathbf{G}_t$ son conocidas.

10.3 Fase de filtro

El proceso del filtro tiene como objetivo encontrar la distribución de β_t dada toda la información $y_1, \ldots, y_t$ para todo $t = 1, \ldots, n$.

Notación: y^t denota el conjunto de información disponible hasta el tiempo t.

Suponga que en el tiempo $t-1$ la distribución de $\boldsymbol{\beta}_{t-1} \mid y^{t-1}$ es $N_p(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$ y de la ecuación del sistema se tiene que $\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1} \sim N_p(\mathbf{G}_t\boldsymbol{\beta}_{t-1}, \mathbf{W}_t)$. Ahora usando el **resultado de la esperanza condicional de la normal** (PONER EL RESULTADO DE RECONSTRUCCIÓN DE LA PÁGINA 22 DE GAMERMAN), con $\boldsymbol{\beta}_t \mid y^{t-1}$ como x_1 y $\boldsymbol{\beta}_{t-1} \mid y^{t-1}$ como x_2, se tiene que la distribución de $\boldsymbol{\beta}_t \mid y^{t-1}$ es $N_p(\mu_1, \Sigma_{11})$, para encontrar μ_1, nótese que $x_1 \mid x_2 = \boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}$ cuya esperanza es $\mathbf{G}_t\boldsymbol{\beta}_{t-1} = \mathbf{G}_t\mathbf{m}_{t-1} + \mathbf{G}_t(\boldsymbol{\beta}_{t-1} - \mathbf{m}_{t-1})$, de donde se tiene que $\mu_1 = \mathbf{G}_t\mathbf{m}_{t-1}$. Ahora según el resultado, $\Sigma_{11} = B_2 + B_1\Sigma_{22}B_1' = \mathbf{W_t} + \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t'$. En conclusión

$$\boldsymbol{\beta}_t \mid y^{t-1} \sim N_p(\mathbf{a}_t, \mathbf{R}_t) \tag{10.3.1}$$

con $\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}$ y $\mathbf{R}_t = \mathbf{W_t} + \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t'$.

Cuando se observa la observación t, se puede actualizar la distribución de $\boldsymbol{\beta}_t$

calculando la distribución $p(\boldsymbol{\beta}_t \mid y^t)$, tenemos que

$$\begin{aligned} p(\boldsymbol{\beta}_t \mid y^t) &= p(\boldsymbol{\beta}_t \mid y_t, y^{t-1}) \\ &\propto p(y_t, y^{t-1}, \boldsymbol{\beta}_t) \\ &\propto p(y_t, y^{t-1}, \boldsymbol{\beta}_t)\frac{1}{p(y^{t-1})} \\ &= \frac{p(y_t, y^{t-1}, \boldsymbol{\beta}_t)}{p(\boldsymbol{\beta}_t, y^{t-1})}\frac{p(\boldsymbol{\beta}_t, y^{t-1})}{p(y^{t-1})} \\ &= p(y_t \mid \boldsymbol{\beta}_t, y^{t-1})p(\boldsymbol{\beta}_t \mid y^{t-1}) \\ &= p(y_t \mid \boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t \mid y^{t-1}) \end{aligned}$$

puesto que dado $\boldsymbol{\beta}_t$, la distribución de y_t es independiente de informaciones previas. Ahora usando $y_t \mid \boldsymbol{\beta}_t \sim N(\mathbf{F}_t'\boldsymbol{\beta}_t, \sigma_t^2)$ y $\boldsymbol{\beta}_t \mid y^{t-1} \sim N_p(\mathbf{a}_t, \mathbf{R}_t)$, se tiene que

$$\begin{aligned} p(\boldsymbol{\beta}_t \mid y^t) &\propto \exp\left\{-\frac{1}{2\sigma_t^2}(y_t - \boldsymbol{\beta}_t'\mathbf{F}_t)^2\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t - \mathbf{a}_t)'\mathbf{R}_t^{-1}(\boldsymbol{\beta}_t - \mathbf{a}_t)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_t'(\sigma_t^{-2}\mathbf{F}_t\mathbf{F}_t' + \mathbf{R}_t^{-1})\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{R}_t^{-1}\mathbf{a}_t)\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_t'\mathbf{C}_t^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\mathbf{C}_t^{-1}\mathbf{C}_t(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{R}_t^{-1}\mathbf{a}_t)\right]\right\} \end{aligned}$$

con $\mathbf{C}_t^{-1} = \sigma_t^{-2}\mathbf{F}_t\mathbf{F}_t' + \mathbf{R}_t^{-1}$. Gamerman (2003) afirma que $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{A}_t'\mathbf{Q}_t$ con $\mathbf{Q}_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + \sigma_t^2$ y $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t/\mathbf{Q}_t$. Entonces se tiene que

$$\begin{aligned} \mathbf{m}_t := \mathbf{C}_t(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{R}_t^{-1}\mathbf{a}_t) &= (\mathbf{R}_t - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t}{\mathbf{Q}_t})(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{R}_t^{-1}\mathbf{a}_t) \\ &= \mathbf{a}_t + \sigma_t^{-2}\mathbf{R}_t\mathbf{F}_t y_t - \frac{\sigma_t^{-2}\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t y_t}{\mathbf{Q}_t} - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{a}_t}{\mathbf{Q}_t} \\ &= \mathbf{a}_t + \frac{\mathbf{R}_t\mathbf{F}_t}{\mathbf{Q}_t}(\sigma_t^{-2}y_t\mathbf{Q}_t - \sigma_t^{-2}\mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t y_t - \mathbf{F}_t'\mathbf{a}_t) \\ &= \mathbf{a}_t + \mathbf{A}_t(y_t - \mathbf{F}_t'\mathbf{a}_t). \end{aligned}$$

De esta forma se tiene que

$$p(\boldsymbol{\beta}_t \mid y^t) \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_t'\mathbf{C}_t^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\mathbf{C}_t^{-1}\mathbf{m_t}\right]\right\},$$

esto es,

$$\boldsymbol{\beta}_t \mid y^t \sim N_p(\mathbf{m_t}, \mathbf{C}_t). \tag{10.3.2}$$

En conclusión, el proceso del filtro consta de

1. Fijar una distribución previa $\boldsymbol{\beta}_1 \mid y^0 \sim N_p(\mathbf{a}_1, \mathbf{R}_1)$
2. Obtener la distribución actualizada $\boldsymbol{\beta}_1 \mid y^1$ usando (**??**)
3. Calcular la distribución de $\boldsymbol{\beta}_2 \mid y^1$ usando (**??**)

4. Calcular la distribución de $\boldsymbol{\beta}_2 \mid y^2$ usando (**??**)
5. ...
6. Obtener la distribución de $\boldsymbol{\beta}_n \mid y^n$ usando (**??**).

10.4 Fase de suavizamiento

Una vez concluido el algoritmo del filtro, se procede con el suavizamiento que tiene como fin encontrar la distribución de β_t dada toda la información $y_1, \ldots, y_n$ para todo $t = 1, \ldots, n$.

En primer lugar, se encuentra la distribución conjunta de las observaciones y_t y los coeficientes de regresión $\boldsymbol{\beta}_t$ para $t = 1, \ldots, n$.

Resultado 10.4.1. *La función de densidad de la distribución conjunta de* $\mathbf{y}^n = (y_1, \ldots, y_n)'$ *y* $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)'$ *está dada por*

$$p(\mathbf{y}^n, \boldsymbol{\beta}) = \prod_{t=1}^{n} p(\mathbf{y}_t \mid \boldsymbol{\beta}_t) \prod_{t=2}^{n} p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) p(\boldsymbol{\beta}_1). \tag{10.4.1}$$

Prueba. Tenemos que

$$p(y^n, \boldsymbol{\beta}) = p(y_1, \ldots, y_n \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n).$$

Ahora

$$\begin{aligned} & p(y_1, \ldots, y_n \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \\ &= p(y_1 \mid y_2, \ldots, y_n, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) p(y_2 \mid y_3, \ldots, y_n, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \ldots p(y_n \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \\ &= p(y_1 \mid \boldsymbol{\beta}_1) p(y_2 \mid \boldsymbol{\beta}_2) \ldots p(y_n \mid \boldsymbol{\beta}_n), \end{aligned}$$

puesto que para todo $i = 1, \ldots, n$, dado $\boldsymbol{\beta}_i$, y_i es independiente de y_j para todo $j > i$.

Por otro lado

$$\begin{aligned} & p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \\ &= p(\boldsymbol{\beta}_n \mid \boldsymbol{\beta}_{n-1}, \ldots, \boldsymbol{\beta}_1) p(\boldsymbol{\beta}_{n-1} \mid \boldsymbol{\beta}_{n-2}, \ldots, \boldsymbol{\beta}_1) \ldots p(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) p(\boldsymbol{\beta}_1) \\ &= p(\boldsymbol{\beta}_n \mid \boldsymbol{\beta}_{n-1}) p(\boldsymbol{\beta}_{n-1} \mid \boldsymbol{\beta}_{n-2}) \ldots p(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) p(\boldsymbol{\beta}_1), \end{aligned}$$

puesto que para todo $i = 1, \ldots, n-1$, dado $\boldsymbol{\beta}_i$, $\boldsymbol{\beta}_{i+1}$ es independiente de $\boldsymbol{\beta}_j$ para todo $j < i + 1$. ■

Usando la anterior función de densidad conjunto, se puede encontrar la distribución posterior de $\boldsymbol{\beta}_t$ dada toda la información $y_1, \ldots, y_n$ para todo $t = 1, \ldots, n$, como lo ilustra el siguiente resultado.

Resultado 10.4.2. *Dado* $y_1, \ldots, y_n$, $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_{t-1}$, $\boldsymbol{\beta}_{t+1}$, ..., $\boldsymbol{\beta}_n$, *la distribución posterior de* $\boldsymbol{\beta}_t$ *es* $N_p(\mathbf{b}_t, \mathbf{B}_t)$ *con*

$$\mathbf{b}_t = \begin{cases} \mathbf{B}_1(\sigma_1^{-2}\mathbf{F}_1 y_1 + \mathbf{G}_2'\mathbf{W}_2^{-1}\boldsymbol{\beta}_2 + \mathbf{R}_1^{-1}\mathbf{a}_1) & t = 1 \\ \mathbf{B}_t(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{G}_{t+1}'\mathbf{W}_{t+1}^{-1}\boldsymbol{\beta}_{t+1} + \mathbf{W}_t^{-1}\mathbf{G}_t\boldsymbol{\beta}_{t-1}) & t = 2, \ldots, n-1 \\ \mathbf{B}_n(\sigma_n^{-2}\mathbf{F}_n y_n + \mathbf{W}_n^{-1}\mathbf{G}_n\boldsymbol{\beta}_{n-1}) & t = n \end{cases}$$

y

$$\mathbf{B}_t = \begin{cases} (\sigma_1^{-2}\mathbf{F}_1\mathbf{F}_1' + \mathbf{G}_2'\mathbf{W}_2^{-1}\mathbf{G}_2 + \mathbf{R}_1^{-1})^{-1} & t = 1 \\ (\sigma_t^{-2}\mathbf{F}_t\mathbf{F}_t' + \mathbf{G}_{t+1}'\mathbf{W}_{t+1}^{-1}\mathbf{G}_{t+1} + \mathbf{W}_t^{-1})^{-1} & t = 2, \ldots, n-1 \\ (\sigma_n^{-2}\mathbf{F}_n\mathbf{F}_n' + \mathbf{W}_n^{-1})^{-1} & t = n \end{cases}$$

Prueba.

$$\begin{aligned} & p(\boldsymbol{\beta}_t \mid y_1, \ldots, y_n, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_{t+1}, \ldots, \boldsymbol{\beta}_n) \\ \propto\ & p(y_1, \ldots, y_n, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \\ =\ & p(y_1 \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, y_2, \ldots, y_n)p(y_2 \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, y_3, \ldots, y_n) \ldots p(y_n \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \\ =\ & p(y_1 \mid \boldsymbol{\beta}_1) \ldots p(y_n \mid \boldsymbol{\beta}_n)p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \\ \propto\ & p(y_t \mid \boldsymbol{\beta}_t)p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) \\ =\ & p(y_t \mid \boldsymbol{\beta}_t)p(\boldsymbol{\beta}_n \mid \boldsymbol{\beta}_{n-1}, \ldots, \boldsymbol{\beta}_1)p(\boldsymbol{\beta}_{n-1} \mid \boldsymbol{\beta}_{n-2}, \ldots, \boldsymbol{\beta}_1) \ldots p(\boldsymbol{\beta}_1) \\ =\ & p(y_t \mid \boldsymbol{\beta}_t)p(\boldsymbol{\beta}_n \mid \boldsymbol{\beta}_{n-1})p(\boldsymbol{\beta}_{n-1} \mid \boldsymbol{\beta}_{n-2}) \ldots p(\boldsymbol{\beta}_1) \\ \propto\ & p(y_t \mid \boldsymbol{\beta}_t)p(\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_{t-1})p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) \end{aligned}$$

para $t = 2, \cdots, n-2$. Mientras que para $t = 1$

$$p(\boldsymbol{\beta}_1 \mid y_1, \ldots, y_n, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n) \propto p(y_1 \mid \boldsymbol{\beta}_1)p(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)p(\boldsymbol{\beta}_1), \tag{10.4.2}$$

y para $t = n$

$$p(\boldsymbol{\beta}_n \mid y_1, \ldots, y_n, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{n-1}) \propto p(y_n \mid \boldsymbol{\beta}_n)p(\boldsymbol{\beta}_n \mid \boldsymbol{\beta}_{n-1}). \tag{10.4.3}$$

En primer lugar, consideramos $t = 2, \ldots, n-2$. Recurriendo a la ecuación de observación y la distribución normal de ϵ_t, se tiene que $y_t \mid \boldsymbol{\beta}_t \sim N(\mathbf{F}_t'\boldsymbol{\beta}_t, \sigma_t^2)$; por otro lado, la ecuación de sistema y la distribución normal de $\boldsymbol{\omega}_t$ conducen a $\boldsymbol{\beta}_{t+1} \mid \boldsymbol{\beta}_t \sim N_p(\mathbf{G}_{t+1}\boldsymbol{\beta}_t, \mathbf{W}_{t+1})$ y $\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1} \sim N_p(\mathbf{G}_t\boldsymbol{\beta}_{t-1}, \mathbf{W}_t)$. Usando estas distribuciones, se tiene que

$$\begin{aligned} & p(\boldsymbol{\beta}_t \mid y_1, \ldots, y_n) \\ \propto\ & \exp\left\{-\frac{1}{2\sigma_t^2}(y_t - \mathbf{F}_t'\boldsymbol{\beta}_t)^2\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_{t+1} - \mathbf{G}_{t+1}\boldsymbol{\beta}_t)'\mathbf{W}_{t+1}^{-1}(\boldsymbol{\beta}_{t+1} - \mathbf{G}_{t+1}\boldsymbol{\beta}_t)\right\} \\ & \qquad \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})'\mathbf{W}_t^{-1}(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})\right\} \\ \propto\ & \exp\left\{\frac{1}{2}\left[\boldsymbol{\beta}_t'(\sigma_t^{-2}\mathbf{F}_t\mathbf{F}_t' + \mathbf{G}_{t+1}'\mathbf{W}_{t+1}^{-1}\mathbf{G}_{t+1} + \mathbf{W}_t^{-1})\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{G}_{t+1}'\mathbf{W}_{t+1}^{-1}\boldsymbol{\beta}_{t+1} + \mathbf{W}_t^{-1}\mathbf{G}_t\boldsymbol{\beta}_{t-1})\right]\right. \end{aligned}$$

Al definir $\mathbf{B}_t = (\sigma_t^{-2}\mathbf{F}_t\mathbf{F}_t' + \mathbf{G}_{t+1}'\mathbf{W}_{t+1}^{-1}\mathbf{G}_{t+1} + \mathbf{W}_t^{-1})^{-1}$, se tiene que

$$\begin{aligned} p(\boldsymbol{\beta}_t \mid y_1, \ldots, y_n) &\propto \exp\left\{\frac{1}{2}\left[\boldsymbol{\beta}_t'\mathbf{B}_t^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\mathbf{B}_t^{-1}\mathbf{B}_t(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{G}_{t+1}'\mathbf{W}_{t+1}^{-1}\boldsymbol{\beta}_{t+1} + \mathbf{W}_t^{-1}\mathbf{G}_t\boldsymbol{\beta}_{t-1})\right]\right\} \\ &= \exp\left\{\frac{1}{2}\left[\boldsymbol{\beta}_t'\mathbf{B}_t^{-1}\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\mathbf{B}_t^{-1}\mathbf{b}_t\right]\right\} \end{aligned}$$

con $\mathbf{b}_t = \mathbf{B}_t(\sigma_t^{-2}\mathbf{F}_t y_t + \mathbf{G}_{t+1}'\mathbf{W}_{t+1}^{-1}\boldsymbol{\beta}_{t+1} + \mathbf{W}_t^{-1}\mathbf{G}_t\boldsymbol{\beta}_{t-1})$, de donde se concluye que $\boldsymbol{\beta}_t \mid y_1, \ldots, y_n \sim N_p(\mathbf{b}_t, \mathbf{B}_t)$ para $t = 2, \ldots, n-2$.

Ahora, considera $t = 1$, en cuyo caso la distribución posterior de $\boldsymbol{\beta}_1$ está dada en (**??**). Tenemos que $y_1 \mid \boldsymbol{\beta}_1 \sim N(\mathbf{F}_1'\boldsymbol{\beta}_1, \sigma_1^2)$, $\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1 \sim N_p(\mathbf{G}_2\boldsymbol{\beta}_1, \mathbf{W}_2)$ y $\boldsymbol{\beta}_1 \sim N_p(\mathbf{a}_1, \mathbf{R}_1)$. Entonces tenemos que

$$\begin{aligned} & p(\boldsymbol{\beta}_1 \mid y_1, \ldots, y_n) \\ &\propto p(y_1 \mid \boldsymbol{\beta}_1)p(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)p(\boldsymbol{\beta}_1) \\ &\propto \exp\left\{-\frac{1}{2\sigma_1^2}(y_1 - \mathbf{F}_1'\boldsymbol{\beta}_1)^2\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_2 - \mathbf{G}_2\boldsymbol{\beta}_1)'\mathbf{W}_2^{-1}(\boldsymbol{\beta}_2 - \mathbf{G}_2\boldsymbol{\beta}_1)\right\} \\ &\qquad\qquad \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_1 - \mathbf{a}_1)'\mathbf{R}_1^{-1}(\boldsymbol{\beta}_1 - \mathbf{a}_1)\right\} \\ &\propto \exp\left\{\frac{1}{2}\left[\boldsymbol{\beta}_1'(\sigma_1^{-2}\mathbf{F}_1\mathbf{F}_1' + \mathbf{G}_2'\mathbf{W}_2^{-1}\mathbf{G}_2 + \mathbf{R}_1^{-1})\boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1'(\sigma_1^{-2}\mathbf{F}_1 y_1 + \mathbf{G}_2'\mathbf{W}_2^{-1}\boldsymbol{\beta}_2 + \mathbf{R}_1^{-1}\mathbf{a}_1)\right]\right\} \\ &= \exp\left\{\frac{1}{2}\left[\boldsymbol{\beta}_1'\mathbf{B}_1^{-1}\boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1'\mathbf{B}_1^{-1}\mathbf{B}_1(\sigma_1^{-2}\mathbf{F}_1 y_1 + \mathbf{G}_2'\mathbf{W}_2^{-1}\boldsymbol{\beta}_2 + \mathbf{R}_1^{-1}\mathbf{a}_1)\right]\right\} \\ &= \exp\left\{\frac{1}{2}\left[\boldsymbol{\beta}_1'\mathbf{B}_1^{-1}\boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_1'\mathbf{B}_1^{-1}\mathbf{b}_1\right]\right\} \end{aligned}$$

con $\mathbf{B}_1 = (\sigma_1^{-2}\mathbf{F}_1\mathbf{F}_1' + \mathbf{G}_2'\mathbf{W}_2^{-1}\mathbf{G}_2 + \mathbf{R}_1^{-1})^{-1}$ y $\mathbf{b}_1 = \mathbf{B}_1(\sigma_1^{-2}\mathbf{F}_1 y_1 + \mathbf{G}_2'\mathbf{W}_2^{-1}\boldsymbol{\beta}_2 + \mathbf{R}_1^{-1}\mathbf{a}_1)$. De donde se concluye que $\boldsymbol{\beta}_1 \mid y_1, \ldots, y_n \sim N_p(\mathbf{b}_1, \mathbf{B}_1)$.

El procedimiento para $t = n$ es análogo, y queda demostrado el resultado. ■

Las distribuciones del anterior resultado permite llevar a cabo un algoritmo de muestreador de Gibbs. (Ver Gaberman MCMC, p. 173)

Para encontrar la distribución suavizada $\boldsymbol{\beta}_t \mid y^n$ para $t = 1, \ldots, n$, se encuentra

en primer lugar la distribución suavizada conjunta de $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_2$. Tenemos que

$$\begin{aligned}
p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n \mid y^n) &= \frac{p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n, y^n)}{p(y^n)} \\
&= \frac{p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n \mid y^n) p(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_3, \ldots, \boldsymbol{\beta}_n, y^n) \ldots p(\boldsymbol{\beta}_{n-1} \mid \boldsymbol{\beta}_n, y^n) p(\boldsymbol{\beta}_n, y^n)}{p(y^n)} \\
&= p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n \mid y^n) p(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_3, \ldots, \boldsymbol{\beta}_n, y^n) \ldots p(\boldsymbol{\beta}_{n-1} \mid \boldsymbol{\beta}_n, y^n) p(\boldsymbol{\beta}_n \mid y^n) \\
&= p(\boldsymbol{\beta}_n \mid y^n) \prod_{i=1}^{n-1} p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{i+1}, \ldots, \boldsymbol{\beta}_n, y^n) \\
&= p(\boldsymbol{\beta}_n \mid y^n) \prod_{i=1}^{n-1} p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{i+1}, y^i) \qquad (10.4.4)
\end{aligned}$$

la última igualdad se tiene por el hecho de que dado $\boldsymbol{\beta}_{i+1}$, $\boldsymbol{\beta}_i$ es independiente de y_j y $\boldsymbol{\beta}_j$ con $j > i$.

Integrando con respecto a $\boldsymbol{\beta}_1$ la ecuación (**??**), se tiene que

$$\begin{aligned}
p(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n \mid y^n) &= p(\boldsymbol{\beta}_n \mid y^n) \prod_{i=2}^{n-1} p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{i+1}, y^t) \int p(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n, y^t) d\boldsymbol{\beta}_1 \\
&= p(\boldsymbol{\beta}_n \mid y^n) \prod_{i=2}^{n-1} p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{i+1}, y^i),
\end{aligned}$$

puesto que la integral de una función de densidad condicional vale 1. Análogamente integrando con respecto a $\boldsymbol{\beta}_2$, ..., $\boldsymbol{\beta}_{t-1}$, se tiene que

$$p(\boldsymbol{\beta}_t, \ldots, \boldsymbol{\beta}_n \mid y^n) = p(\boldsymbol{\beta}_n \mid y^n) \prod_{i=t}^{n-1} p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{i+1}, y^i),$$

y además

$$\begin{aligned}
p(\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t+1} \mid y^n) &= \frac{p(\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t+1}, y^n)}{p(y^n)} \\
&= \frac{p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, y^n) p(\boldsymbol{\beta}_{t+1}, y^n)}{p(y^n)} \\
&= p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, y^n) p(\boldsymbol{\beta}_{t+1} \mid y^n) \\
&= p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t+1}, y^t) p(\boldsymbol{\beta}_{t+1} \mid y^n)
\end{aligned}$$

Resultado 10.4.3.

$$\boldsymbol{\beta}_t \mid y^n \sim N(\mathbf{m}_t^n, \mathbf{C}_t^n) \qquad (10.4.5)$$

con $\mathbf{m}_t^n = \mathbf{m}_t + \mathbf{C}_t \mathbf{G}_{t+1}' \mathbf{R}_{t+1}^{-1} (\mathbf{m}_{t+1}^n - \mathbf{a}_{t+1})$ *y* $\mathbf{C}_t^n = \mathbf{C}_t - \mathbf{C}_t \mathbf{G}_{t+1}' \mathbf{R}_{t+1}^{-1} (\mathbf{R}_{t+1} - \mathbf{C}_{t+1}^n) \mathbf{R}_{t+1}^{-1} \mathbf{G}_{t+1} \mathbf{C}_t$

En síntesis, el algoritmo del suavizamiento consta de los siguientes pasos:

1. Obtener las distribuciones filtrados $\boldsymbol{\beta}_1 \mid y^1$, $\boldsymbol{\beta}_2 \mid y^2$, ..., $\boldsymbol{\beta}_n \mid y^n$.
2. Obtener las distribuciones suavizados $\boldsymbol{\beta}_1 \mid y^n$, $\boldsymbol{\beta}_2 \mid y^n$, ..., $\boldsymbol{\beta}_n \mid y^n$. Esto se lleva a cabo de la siguiente forma,
   (a) usando la distribución de $\boldsymbol{\beta}_n \mid y^n \sim N(\mathbf{m}_n, \mathbf{C}_n)$ y notando que $\mathbf{m}_n = \mathbf{m}_n^n$ y $\mathbf{C}_n = \mathbf{C}_n^n$, calcular $\mathbf{m}_{n-1}^n$ y $\mathbf{C}_{n-1}^n$, y así obtener la distribución de $\boldsymbol{\beta}_{n-1} \mid y^n$.
   (b) usar la distribución de $\boldsymbol{\beta}_{n-1} \mid y^n$ para encontrar la de $\boldsymbol{\beta}_{n-2} \mid y^n$ similarmente.
   (c) Repetir el proceso hasta obtener la distribución de $\boldsymbol{\beta}_1 \mid y^n$.

10.4.1 σ_t^2 y $\mathbf{W}_t$ desconocidos

Cuando σ_t^2 y $\mathbf{W}_t$, las varianzas de los errores ϵ_t y $\boldsymbol{\omega}_t$ son desconocidas, las distribuciones encontradas anteriormente siguen siendo válidas condicionando sobre σ_t^2 y $\mathbf{W}_t$.

Para encontrar las distribuciones concernientes a σ_t^2 y $\mathbf{W}_t$, suponga que éstas son constantes a través del tiempo, esto es, $\sigma_t^2 = \sigma^2$ y $\mathbf{W}_t = \mathbf{W}$ para $t = 1, \ldots, n$. Tenemos las siguientes distribuciones posterior de σ^2 y $\mathbf{W}$.

$$
\begin{aligned}
& p(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{W}, y_1, \ldots, y_n) \\
&= \frac{p(\sigma^2, \boldsymbol{\beta}, \mathbf{W}, y_1, \ldots, y_n)}{p(\boldsymbol{\beta}, \mathbf{W}, y_1, \ldots, y_n)} \\
&\propto p(y_1 \mid \sigma^2, \boldsymbol{\beta}, \mathbf{W}, y_2, \ldots, y_n) p(y_2 \mid \sigma^2, \boldsymbol{\beta}, \mathbf{W}, y_3, \ldots, y_n) \ldots p(y_n \mid \sigma^2, \boldsymbol{\beta}, \mathbf{W}) p(\sigma^2, \boldsymbol{\beta}, \mathbf{W}) \\
&= p(y_1 \mid \sigma^2, \boldsymbol{\beta}_1) p(y_2 \mid \sigma^2, \boldsymbol{\beta}_2) \ldots p(y_n \mid \sigma^2, \boldsymbol{\beta}_n) p(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{W}) p(\boldsymbol{\beta}, \mathbf{W}) \\
&\propto \prod_{i=1}^{n} p(y_i \mid \sigma^2, \boldsymbol{\beta}_i) p(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{W})
\end{aligned}
$$

y de forma análogo, se tiene que

$$
p(\mathbf{W} \mid \boldsymbol{\beta}, \sigma^2, y_1, \ldots, y_n) \propto \prod_{i=2}^{n} p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_{i-1}, \mathbf{W}) p(\mathbf{W} \mid \boldsymbol{\beta}, \sigma^2)
$$

Suponiendo que las distribuciones previa son $\sigma^2 \sim IG(n_\sigma/2, n_\sigma S_\sigma/2)$ y $\mathbf{W} \sim IW(n_W/2, n_W S_W/2)$, y además son independientes, entonces

$$
\begin{aligned}
& p(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{W}, y_1, \ldots, y_n) \\
&\propto \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \mathbf{F}_i' \boldsymbol{\beta}_i)^2 \right\} \frac{(n_\sigma S_\sigma/2)^{n_\sigma/2}}{\Gamma(n_\sigma/2)} (\sigma^{-2})^{\frac{n_\sigma}{2}-1} \exp\left\{ -\frac{n_\sigma S_\sigma}{2\sigma^2} \right\} \\
&\propto (\sigma^{-2})^{\frac{n+n_\sigma}{2}-1} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (y_i - \mathbf{F}_i' \boldsymbol{\beta}_i)^2 + n_\sigma S_\sigma \right] \right\} \\
&= (\sigma^{-2})^{\frac{n_\sigma^*}{2}-1} \exp\left\{ -\frac{1}{2\sigma^2} n_\sigma^* S_\sigma^* \right\}
\end{aligned}
$$

con $n_\sigma^* = n + n_\sigma$ y $n_\sigma^* S_\sigma^* = \sum_{i=1}^n (y_i - \mathbf{F}_i' \boldsymbol{\beta}_i)^2 + n_\sigma S_\sigma$. De lo anterior se concluye que $\sigma^2 \mid \boldsymbol{\beta}, \mathbf{W}, y_1, \ldots, y_n \sim IG(n_\sigma^*/2, n_\sigma^* S_\sigma^*/2)$.

Por otro lado,

$$
\begin{aligned}
& p(\mathbf{W} \mid \boldsymbol{\beta}, \sigma^2, y_1, \ldots, y_n) \\
\propto & \prod_{i=2}^{n} \mid \mathbf{W} \mid^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})'\mathbf{W}^{-1}(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})\right\} \mid \mathbf{W} \mid^{-\frac{n_W}{2}+\frac{p+1}{2}} \exp\left\{-tr(\frac{n_W S_W}{2}\mathbf{W}^{-1})\right\} \\
\propto & \mid \mathbf{W} \mid^{-\frac{n-1}{2}-\frac{n_W}{2}+\frac{p+1}{2}} \exp\left\{-tr(\frac{1}{2}\sum_{i=2}^{n}(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})'\mathbf{W}^{-1})\right\} \exp\left\{-tr(\frac{n_W S_W}{2}\mathbf{W}^{-1})\right\} \\
= & \mid \mathbf{W} \mid^{-\frac{n+n_W-1}{2}+\frac{p+1}{2}} \exp\left\{-tr\left[\frac{1}{2}(\sum_{i=2}^{n}(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})' + n_W S_W)\mathbf{W}^{-1}\right]\right\} \\
= & \mid \mathbf{W} \mid^{-\frac{n_W^*}{2}+\frac{p+1}{2}} \exp\left\{-tr\left[\frac{1}{2}n_W^* S_W^* \mathbf{W}^{-1}\right]\right\}
\end{aligned}
$$

con $n_W^* = n_W + n - 1$ y $n_W^* S_W^* = \sum_{i=2}^n (\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1})' + n_W S_W$.
De lo anterior se concluye que $\mathbf{W} \mid \boldsymbol{\beta}, \sigma^2, y_1, \ldots, y_n \sim IW(n_W^*/2, n_W^* S_W^*/2)$.

11 Modelos en poblaciones finitas

11.1 Diseños estadísticos

? afirma que se debe ser un estadístico ingenuo si se afirma que toda inferencia debería ser condicional a los datos, sin importar de dónde o cómo fueron seleccionados. Esta es una concepción errada del principio de verosimilitud. La noción de que el método de selección de la muestra es irrelevante en el análisis inferencial puede ser contradicha con un argumento muy simple: Suponga que se tienen a disposición diez datos provenientes del lanzamiento de diez dados; todos ellos correspondieron al número seis. La actitud del estadístico acerca de la naturaleza de los datos sería diferente si (1) sólo se hicieron diez lanzamientos, (2) se hicieron sesenta lanzamientos pero se decidió reportar sólo los que resultaron ser seis, (3) aparecieron diez seis en quinientos lanzamientos y se decidió reportar ´honestamente estas realizaciones. En tales situaciones es claro que la distribución de los datos observados sigue un patrón completamente distinto que no debe ser ignorado.

Una **población finita** es un conjunto de N elementos $\{e_1, e_2, ..., e_N\}$. Cada unidad puede ser identificada sin ambigüedad por un conjunto de rótulos. Sea $U = \{1, 2, ..., N\}$ el conjunto de rótulos de la población finita[1]. En términos generales, es el conjunto de unidades que conforman el universo de estudio. N es comúnmente llamado el tamaño poblacional. Se utilizará el subíndice i para denotar la existencia física del i-ésimo elemento. En algunas ocasiones el objetivo de la investigación es poder estimar el tamaño de la población.

Para resolver los objetivos del estudio se recurre al planteamiento de una estrategia de muestreo con el fin de seleccionar una **muestra aleatoria** que representa un subconjunto de la población que ha sido extraído mediante un mecanismo estadístico de selección. Notaremos con una letra mayúscula S a la muestra aleatoria[2] y con una letra minúscula s a una realización de la misma. De tal forma que, sin ambigüedad, una muestra seleccionada (realizada) es el conjunto de unidades pertenecientes a

$$s = \{1, ..., k, ..., n\}.$$

El número de componentes de s es llamado el **tamaño de muestra** y no siempre es fijo. Es decir, en algunos casos n es una cantidad aleatoria. El conjunto de todas las posibles muestras se conoce como **soporte**. Haciendo una analogía con

[1] El tamaño de la población no es necesariamente conocido.
[2] Nótese que S es una variable aleatoria.

la inferencia estadística clásica, el soporte generado por una muestra aleatoria corresponde al espacio muestral generado por una variable aleatoria.

En términos generales, un diseño de muestreo no es sino una distribución de probabilidad multivariante definida sobre un conjunto de muestras que pertenecen a un soporte. Pero, una distribución de probabilidad no es más sino un modelo que se asume; en este caso, es un modelo que permite la selección de muestras probabilísticas. Una muestra s induce un vector de inclusión $\mathbf{I}(s) = (I_1(s), \ldots, I_k(s), \ldots, I_N(s))'$, en donde $I_k(s)$ está definida por (2.1.8). Dado el anterior esquema, otra forma de denotar el diseño de muestreo es $p(\mathbf{I})$ el cual se conoce para todos los posibles valores de $\mathbf{I}$ en todas las posibles muestras s. Por otro lado, si se asume que la medición de la característica de interés y_k en los individuos de la población está sujeta a un error, entonces éstas deben ser vistas como realizaciones de variables aleatorias Y_k. De esta forma, es necesario definir un modelo para los valores poblacionales que puede depender de cierto parámetro. En este caso, si $Y = (Y_1, \ldots, Y_k, \ldots, Y_N)'$ es el vector poblacional de la característica de interés, entonces $p(\mathbf{Y} \mid \theta)$ definirá tal modelo.

Para realizar cualquier tipo de inferencias acerca del parámetro θ es necesario trabajar con una distribución de probabilidad conjunta de $(\mathbf{I}, \mathbf{Y})$ que permita unificar todo el esquema anterior en un sólo proceso. La pregunta que atañe al estadístico es la siguiente: ¿cómo se puede expresar esa distribución conjunta en términos de $p(\mathbf{I})$ y de $p(\mathbf{I}, \theta)$? **?** dan la respuesta a esta pregunta motivando la suposición de que $\mathbf{Y}$ sea independiente de $\mathbf{I}$. En algunos caso como en **?**, capítulo 8 el diseño de muestreo depende de los valores de la característica de interés; por ejemplo, en un estudio en de casos y controles, la respuesta y_k es de tipo binario, indicando si la k-ésima unidad corresponde a un caso o a un control. A su vez, los casos y controles inducen estratos cuyas muestras son seleccionadas independientemente. En este caso, el diseño de muestreo depende directamente de los valores de la característica de interés. Por lo tanto, la relación entre $\mathbf{I}, \mathbf{Y}$ debe ser expresada como

$$p(\mathbf{I}, \mathbf{Y} \mid \theta) = p(\mathbf{I} \mid \mathbf{Y})p(\mathbf{Y} \mid \theta)$$

En este caso, se dice que el diseño de muestreo es *informativo* y no puede ser ignorado en términos de inferencia para θ. Por otro lado, si el diseño de muestreo es *no informativo*, la relación entre $\mathbf{I}, \mathbf{Y}$ debe ser expresada como

$$p(\mathbf{I}, \mathbf{Y} \mid \theta) = p(\mathbf{I})p(\mathbf{Y} \mid \theta)$$

y claramente, el diseño de muestreo puede ser ignorado. **?** afirman que los diseños de muestreo que dependen directamente de la variable de interés no son raros en la práctica. Sin emabargo, los diseños de muestreo implementados cuando el marco de muestreo es muy deficiente como el muestreo en dos fases, en donde se selecciona una primera muestra y con base en los resultados de esta se diseña la estrategia para una segunda submuestra, no puede ser catalogado como no informativo y por tanto no puede ser ignorado. Por otro lado, es más común encontrar que el diseño de muestreo dependa de otras variables de información auxiliar, como en el diseño estratificado o el diseño proporcional al tamaño. A

continuación se presenta el marco general dado por **?** para modelar conjuntamente el diseño de muestreo y el mecanismo probabilístico que origina a la variable de interés. Además, **?** afirman que diseños de muestreo como el aleatorio simple, aleatorio estratificado, proporcional al tamaño, el muestreo a conveniencia o el muestreo balanceado corresponde a casos en donde es posible ignorar el mecanismo de selección. También concluyen que aunque algunas veces los diseños de muestreo pueden ser ignorados en términos de inferencia para $\boldsymbol{\beta}$, es equivocado pensar que siempre pueden ser ignorados en términos de inferencia predictiva para el total poblacional T_y.

Aunque en este capítulo sólo abordaremos la inferencia Bayesiana en donde el diseño de muestreo se considera no informativo, pero los autores recalcan que en la práctica los diseños estadísticos son mayormente complejos y no pueden ser ignorados tan fácilmente. Por ejemplo, el siguiente diseño de muestreo, conocido como diseño aleatorio simple, es muy utilizado en la práctica en las últimas etapas de diseños de muestreo complejos.

$$p(\mathbf{I}) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } \sum_{k\in U} I_k = n \\ 0 & \text{en otro caso} \end{cases} \tag{11.1.1}$$

Claramente este diseño de muestreo no depende de la característica de interés y puede catalogarse como no informativo. Otro diseño de muestreo con estas características es el diseño aleatorio estratificado dado por

$$p(\mathbf{I}) = \begin{cases} \prod_{h=1}^{H} \frac{1}{\binom{N_h}{n_h}}, & \text{si } \sum_{h=1}^{H} n_h = n \\ 0, & \text{en otro caso} \end{cases} \tag{11.1.2}$$

Los anteriores diseños de muestreo cumplen la siguiente propiedad

$$p(\mathbf{I}, \mathbf{Y} \mid \theta) = p(\mathbf{I}) p(\mathbf{Y} \mid \theta)$$

y por lo tanto en términos de inferencia pueden ser ignorados. Este capítulo está enfocado en la inferencia de parámetros para poblaciones finitas cuando el diseño de muestreo es ignorado. La inferencia bayesiana para poblaciones finitas supone el uso de información auxiliar que relaciona a la característica de interés con las variables auxiliares mediante un modelo de superpoblación, ξ. Bajo esta perspectiva la observación de la característica de interés en las unidades poblacionales y_k se define como la realización de una variable aleatoria Y_k. Partiendo de que el total poblacional se puede escribir como

$$T_y = \sum_{k\in s} Y_k + \sum_{k\notin s} Y_k, \tag{11.1.3}$$

la tarea es estimar por medio del modelo bayesiano, las respectivas observaciones y_k de los elementos que no fueron seleccionados en la muestra. Denotando esta

estimación como $E(Y_k)$, un predictor para el total estaría dado por:

$$\hat{T}_y = \sum_{k \in s} Y_k + \sum_{k \notin s} E_\xi(Y_k) \tag{11.1.4}$$

y por tanto la realización de $\hat{T}_y$ con los datos específicos de la muestra seleccionada s estaría definida como

$$\hat{t}_y = \sum_{k \in s} y_k + \sum_{k \notin s} \hat{E}_\xi(Y_k) \tag{11.1.5}$$

donde $\hat{E}_\xi(Y_k)$ es una estimación de $E_\xi(Y_k)$ realizada con los datos obtenidos de la muestra seleccionada s. Suponga que el vector de las variables de interés es $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ y para cada elemento de la población la realización de estas variables aleatorias es $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$. Suponga que el objetivo es estimar una combinación lineal[3] $\mathbf{l}'\mathbf{y}$. Para tal fin, se selecciona una muestra s de tamaño n. Nótese que tanto $\mathbf{y}$ como $\mathbf{l}$ se pueden particionar de la siguiente manera: $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ y $\mathbf{l} = (\mathbf{l}_s', \mathbf{l}_r')'$; en donde el subíndice s se refiere a que el vector contiene los n elementos de la muestra seleccionada y el subíndice r se refiere a que el vector contiene los $N - n$ elementos que no fueron seleccionados en la muestra. En este orden de ideas, **?** desarrollaron el siguiente marco de referencia con los siguientes resultados que dan cuenta de la estimación de parámetros poblacionales de interés en inferencia de poblaciones finitas como lo son totales o medias.

Resultado 11.1.1. *Para cualquier cantidad lineal poblacional* $\mathbf{l}'\mathbf{Y}$*, el predictor Bayesiano que minimiza la pérdida del error cuadrático bajo cualquier modelo, para el cual* $Var_\xi(\mathbf{Y}_r \mid \mathbf{Y}_s)$ *existe, está dado por*

$$\mathbf{l}_s'\mathbf{Y}_s + \mathbf{l}_r' E_\xi(\mathbf{Y}_r \mid \mathbf{Y}_s) \tag{11.1.6}$$

Existe una infinidad de modelos bayesianos que se pueden proponer, por tanto la lógica que sigue este capítulo es la revisión de los modelos más simples que van a ser generalizados en secciones posteriores para dar así una visión amplia, más no exhaustiva, del espectro que siguen los modelos bayesianos en inferencia de poblaciones finitas.

11.2 Modelo simple

Suponga que $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ y además asuma que el modelo de superpoblación ξ que rige el comportamiento de la estructura probabilística de la variable de interés en la población finita está dado por

$$Y_i = \theta + e_i \qquad i = 1, \ldots, N \tag{11.2.1}$$

[3] Si el objetivo es estimar el total poblacional, entonces $\mathbf{l}' = (1, 1, \ldots, 1)$. Si el objetivo es estimar la media poblacional, entonces $\mathbf{l}' = (1/N, 1/N, \ldots, 1/N)$.

Además suponga que $e_i \sim N(0, \sigma^2)$ y que las distribuciones previa de cada una de las variables de interés y del parámetro θ son

$$\begin{aligned} Y_i \mid \theta &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\mu, \tau^2) \end{aligned}$$

Donde σ^2, μ y τ^2 se asumen conocidas. Asuma entonces que cada uno de los errores del modelo e_i es independiente de θ para todo $i = 1, \ldots, N$.

Resultado 11.2.1. *Si* $\mathbf{Y}$ *se particiona como* $\mathbf{Y} = (\mathbf{Y}_r', \mathbf{Y}_s')$*, donde* $\mathbf{Y}_r$ *y* $\mathbf{Y}_s$ *son de tamaño* $N - n$ *y* n*, respectivamente. Entonces se tiene que* $\mathbf{Y}_r \mid \mathbf{Y}_s$ *tiene distribución normal multivariante con esperanza común* $\mu_r = \dfrac{n\bar{y}_s\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}$*, varianza común* $\tau_r = \sigma^2 + \dfrac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$ *y covarianza común* $\dfrac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$.

Prueba. En primer lugar, tenemos las siguientes propiedades acerca de las variables $Y_1, \ldots, Y_N$. Para $i = 1, \ldots, N$, tenemos que

$$E(Y_i) = E(\theta + e_i) = \mu,$$

$$Var(Y_i) = Var(\theta) + Var(e_i) = \tau^2 + \sigma^2,$$

y

$$\begin{aligned} Cov(Y_i, Y_j) &= Cov(\theta + e_i, \theta + e_j) \\ &= Var(\theta) + Cov(e_i, e_j) \\ &= \tau^2 \end{aligned}$$

para $i \neq j$.

Ahora consideramos el estimador de θ en la muestra observada $\bar{Y}_s$. Usando $Y_i \mid \theta \sim N(\theta, \sigma^2)$ para $i = 1, \ldots, n$, se tiene que $\bar{Y}_s \mid \theta \sim N(\theta, \sigma^2/n)$, además $\theta \sim N(\mu, \tau^2)$, entonces utilizando el resultado A.3.5., se tiene que el vector $(\bar{Y}_s, \theta)$ tiene distribución normal bivariante. Para encontrar el vector de medias y matriz de varianzas y covarianzas, tenemos

$$E(\bar{Y}_s) = E(Y_1) = \mu,$$

$$\begin{aligned} Var(\bar{Y}_s) &= \frac{1}{n^2} Var\Big(\sum_{k \in s} Y_k\Big) \\ &= \frac{1}{n^2}\left[\sum_{k \in s} Var(Y_k) + \sum\sum_{i \neq j} Cov(Y_i, Y_j)\right] \\ &= \frac{1}{n^2}\left[n(\tau^2 + \sigma^2) + (n^2 - n)\tau^2\right] \\ &= \tau^2 + \frac{\sigma^2}{n} \end{aligned}$$

y

$$\begin{aligned} Cov(\bar{Y}_s, \theta) &= \frac{1}{n} \sum_{k \in s} Cov(Y_k, \theta) \\ &= \frac{1}{n} \sum_{k \in s} \tau^2 \\ &= \tau^2. \end{aligned}$$

En conclusión, se tiene que

$$\begin{pmatrix} \bar{Y}_s \\ \theta \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \tau^2 + \frac{\sigma^2}{n} & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix} \right).$$

Ahora, usando la anterior distribución y la propiedad 4 del resultado A.3.3 de la distribución multivariante, se tiene que

$$\theta \mid \bar{Y}_s \sim N(\mu + \tau^2(\tau^2 + \frac{\sigma^2}{n})^{-1}(\bar{y}_s - \mu), \tau^2 - \tau^2(\tau^2 + \frac{\sigma^2}{n})^{-1}\tau^2).$$

Para simplificar la esperanza y la varianza de esta distribución, tenemos

$$\begin{aligned} \mu + \tau^2(\tau^2 + \frac{\sigma^2}{n})^{-1}(\bar{y}_s - \mu) &= \mu + \frac{n\tau^2}{n\tau^2 + \sigma^2}(\bar{y}_s - \mu) \\ &= \frac{n\tau^2 \bar{y}_s}{n\tau^2 + \sigma^2} + (1 - \frac{n\tau^2}{n\tau^2 + \sigma^2})\mu \\ &= \frac{n\tau^2 \bar{y}_s + \sigma^2 \mu}{n\tau^2 + \sigma^2}, \end{aligned}$$

y

$$\begin{aligned} \tau^2 - \tau^2(\tau^2 + \frac{\sigma^2}{n})^{-1}\tau^2 &= \tau^2 - \frac{n\tau^4}{n\tau^2 + \sigma^2} \\ &= \tau^2(1 - \frac{n\tau^2}{n\tau^2 + \sigma^2}) \\ &= \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}. \end{aligned}$$

En conclusión,

$$\theta \mid \bar{Y}_s \sim N(\frac{n\tau^2 \bar{y}_s + \sigma^2 \mu}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}).$$

Ahora, por hipótesis, se tiene que

$$\begin{pmatrix} \mathbf{Y}_r \\ \mathbf{Y}_s \end{pmatrix} \mid \theta \sim N \left(\begin{pmatrix} \theta \mathbf{1}_{N-n} \\ \theta \mathbf{1}_n \end{pmatrix}, \begin{pmatrix} \sigma^2 \mathbf{I}_{N-n} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \mathbf{I}_n \end{pmatrix} \right),$$

de donde

$$\mathbf{Y}_r \mid \mathbf{Y}_s, \theta \sim N(\theta \mathbf{1}_{N-n}, \sigma^2 \mathbf{I}_{N-n}).$$

Por otro lado, la estadística $\mathbf{Y}_s$ es suficiente bayesianamente, (**?**) entonces la distribución de interés $\mathbf{Y}_r \mid \mathbf{Y}_s$ es la misma de $\mathbf{Y}_r \mid \bar{Y}_s$. Y podemos encontrar la distribución de $\mathbf{Y}_r \mid \mathbf{Y}_s$ usando la distribución de $(\mathbf{Y}_r', \mathbf{Y}_s)'$ encontrada anteriormente. Tenemos que

$$\begin{aligned} p(\mathbf{Y}_r \mid \bar{Y}_s) &= \frac{p(\mathbf{Y}_r, \bar{Y}_s)}{p(\bar{Y}_s)} \\ &= \frac{\int p(\mathbf{Y}_r \mid \bar{Y}_s, \theta) p(\theta \mid \bar{Y}_s) p(\bar{Y}_s) d\theta}{p(\bar{Y}_s)} \\ &= \int p(\mathbf{Y}_r \mid \bar{Y}_s, \theta) p(\theta \mid \bar{Y}_s) d\theta. \end{aligned} \tag{11.2.2}$$

De nuevo, usando la suficiencia de $\bar{Y}_s$, se tiene que la distribución de $\mathbf{Y}_r \mid \bar{Y}_s, \theta$ coincide con la de $\mathbf{Y}_r \mid \mathbf{Y}_s, \theta$ que corresponde a una distribución normal multivariante. Entonces la expresión $p(\mathbf{Y}_r \mid \bar{Y}_s, \theta)$ contiene una forma cuadrática de $\mathbf{Y}_r$, y por consiguiente, también lo contiene la expresión (**??**), de donde se concluye que la distribución de $\mathbf{Y}_r \mid \bar{Y}_s$ es normal multivariante. Para encontrar la respectiva esperanza y varianza. Tenemos que

$$\begin{aligned} E(\mathbf{Y}_r \mid \bar{Y}_s) &= E(E(\mathbf{Y}_r \mid \mathbf{Y}_s, \theta) \mid \bar{Y}_s) \\ &= E(\theta \mathbf{1}_{N-n} \mid \bar{Y}_s) \\ &= \mathbf{1}_{N-n} E(\theta \mid \bar{Y}_s) \\ &= \mathbf{1}_{N-n} \frac{n\tau^2 \bar{y}_s + \sigma^2 \mu}{n\tau^2 + \sigma^2}. \end{aligned}$$

Y

$$\begin{aligned} Var(\mathbf{Y}_r \mid \bar{Y}_s) &= E(Var(\mathbf{Y}_r \mid \mathbf{Y}_s, \theta) \mid \bar{Y}_s) + Var(E(\mathbf{Y}_r \mid \mathbf{Y}_s, \theta) \mid \bar{Y}_s) \\ &= E(\sigma^2 \mathbf{I}_{N-n} \mid \bar{Y}_s) + Var(\theta \mathbf{I}_{N-n} \mid \bar{Y}_s) \\ &= \sigma^2 \mathbf{I}_{N-n} + \mathbf{1}_{(N-n)\times(N-n)} \frac{\tau^2 \sigma^2}{n\tau^2 + \sigma^2}, \end{aligned}$$

donde $\mathbf{1}_{(N-n)\times(N-n)}$ denota la matriz de dimensión $(N-n)\times(N-n)$ de entradas iguales a 1. En conclusión, $\mathbf{Y}_r \mid \mathbf{Y}_s$ tiene distribución normal multivariante con esperanza común $\dfrac{n\bar{y}_s\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2}$, varianza común $\sigma^2 + \dfrac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$ y covarianza común $\dfrac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$. ■

Con base en este resultado, se procede a la utilización de XXXXXX para poder calcular las predicciones para los parámetros de interés en la encuesta o el estudio por muestreo.

Resultado 11.2.2. *Para cualquier cantidad lineal poblacional* $\mathbf{l}'\mathbf{Y}$*, el predictor Bayesiano que minimiza la pérdida del error cuadrático bajo el modelo simple está dado por*

$$\mathbf{l}_s' \mathbf{Y}_s + \mathbf{l}_r' \boldsymbol{\mu}_r \tag{11.2.3}$$

donde $\boldsymbol{\mu}_r = (\mu_r, \ldots, \mu_r)'$

En particular la predicción para el total poblacional T_y está dada por la siguiente expresión

$$\hat{T}_y = \sum_{k \in s} Y_k + (N-n)\mu_r \tag{11.2.4}$$

Si el parámetro de interés es la media poblacional, entonces el predictor Bayesiano estaría dado por

$$\bar{Y}_U = \frac{\hat{T}_y}{N} = \frac{\sum_{k \in s} Y_k + (N-n)\mu_n}{N} \tag{11.2.5}$$

11.3 Modelo general

Asuma que el modelo de superpoblación ξ que rige el comportamiento de la estructura probabilística de la variable de interés en la población finita está dado por

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{11.3.1}$$

Además suponga que las distribuciones previa de cada una del vector de variables de interés y del vector de parámetros $\boldsymbol{\beta}$ son

$$\begin{aligned} \mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X} &\sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \\ \boldsymbol{\beta} &\sim N(\mathbf{b}, \mathbf{B}) \end{aligned}$$

Donde $\mathbf{V}$ es la matriz de varianzas poblacional y se supone conocida. Además $\mathbf{e}$ y $\boldsymbol{\beta}$ son independientes.

Resultado 11.3.1. *Si* $\mathbf{Y}$ *se particiona como* $\mathbf{Y} = (\mathbf{Y}_r', \mathbf{Y}_s')$*, donde* $\mathbf{Y}_r$ *y* $\mathbf{Y}_s$ *son de tamaño* $N-n$ *y* n*, respectivamente. Entonces se tiene que* $\mathbf{Y}_r \mid \mathbf{Y}_s$ *tiene distribución normal multivariante con esperanza dada por*

$$E(\mathbf{Y}_r \mid \mathbf{Y}_s) = \boldsymbol{\mu}_r = \mathbf{X}_r\hat{\boldsymbol{\beta}}_B + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s)\hat{\boldsymbol{\beta}}_B$$

y matriz de varianzas dada por

$$\begin{aligned} Var(\mathbf{Y}_r \mid \mathbf{Y}_s) = \boldsymbol{\Sigma}_r = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr} + \\ (\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{X}_s)(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \mathbf{B}^{-1})^{-1}(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{X}_s)', \end{aligned}$$

donde $\mathbf{X}_s$*,* $\mathbf{X}_r$*,* $\mathbf{V}_r$*,* $\mathbf{V}_{rs}$ *y* $\mathbf{V}_{sr}$ *corresponden a las submatrices de* $\mathbf{X}$ *y* $\mathbf{V}$ *correspondiente a la partición de* $\mathbf{Y}$*. Y*

$$\hat{\boldsymbol{\beta}}_B = (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \mathbf{B}^{-1})^{-1}(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s + \mathbf{B}^{-1}\mathbf{b}).$$

Prueba. En primer lugar consideramos el estimador de $\boldsymbol{\beta}$ que es suficiente en el sentido Bayesiano (**?**) dado por

$$\hat{\boldsymbol{\beta}}_s = (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{Y}_s.$$

Por hipótesis, se tiene que $\mathbf{Y}_s \mid \boldsymbol{\beta}$ tiene distribución normal multivariante, entonces $\hat{\boldsymbol{\beta}}_s \mid \boldsymbol{\beta}$ también tiene distribución normal multivariante con

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_s \mid \boldsymbol{\beta}) &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}E(\mathbf{Y}_s \mid \boldsymbol{\beta}) \\ &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

y

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_s \mid \boldsymbol{\beta}) &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}Var(\mathbf{Y}_s \mid \boldsymbol{\beta})\mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} \\ &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{V}_s\mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} \\ &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}. \end{aligned}$$

En conclusión, $\hat{\boldsymbol{\beta}}_s \mid \boldsymbol{\beta} \sim N(\boldsymbol{\beta}, (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1})$. Por otro lado, $\boldsymbol{\beta} \sim N(\mathbf{b}, \mathbf{B})$, entonces por el resultado A.3.5, se tiene que $(\hat{\boldsymbol{\beta}}_s', \boldsymbol{\beta}')$ tiene distribución normal multivariante, con

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_s) &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}E(\mathbf{Y}_s) \\ &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_sE(\boldsymbol{\beta}) \\ &= E(\boldsymbol{\beta}) \\ &= \mathbf{b}, \end{aligned}$$

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_s) &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}Var(\mathbf{Y}_s)\mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} \\ &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}(\mathbf{X}_s\mathbf{B}\mathbf{X}_s' + \mathbf{V}_s)\mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} \\ &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} + \mathbf{B} \end{aligned}$$

y

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_s, \boldsymbol{\beta}) &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}Cov(\mathbf{Y}_s, \boldsymbol{\beta}) \\ &= (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_sCov(\boldsymbol{\beta}, \boldsymbol{\beta}) \\ &= \mathbf{B}. \end{aligned}$$

En conclusión, se tiene que

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_s \\ \boldsymbol{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{b} \\ \mathbf{b} \end{pmatrix}, \begin{pmatrix} (\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} + \mathbf{B} & \mathbf{B} \\ \mathbf{B} & \mathbf{B} \end{pmatrix} \right).$$

Usando la propiedad 4 del resultado A.3.3, se tiene que $\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s$ tiene distribución normal multivariante con esperanza

$$E(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s) = \mathbf{b} + \mathbf{B}[(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} + \mathbf{B}]^{-1}(\hat{\boldsymbol{\beta}}_s - \mathbf{b})$$

al definir $\hat{\boldsymbol{\beta}}_B = E(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s)$, se puede ver que $\hat{\boldsymbol{\beta}}_B = (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \mathbf{B}^{-1})^{-1}(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s + \mathbf{B}^{-1}\mathbf{b})$ y matriz de covarianzas

$$Var(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s) = \mathbf{B} - \mathbf{B}[(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s)^{-1} + \mathbf{B}]^{-1}\mathbf{B}$$

que resulta ser igual a $(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s + \mathbf{B}^{-1})^{-1}$.

Ahora, por hipótesis se tiene que

$$\begin{pmatrix}\mathbf{Y}_r \\ \mathbf{Y}_s\end{pmatrix} \mid \boldsymbol{\beta}, \mathbf{X} \sim N\left(\begin{pmatrix}\mathbf{X}_r\boldsymbol{\beta} \\ \mathbf{X}_s\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_r & \mathbf{V}_{rs} \\ \mathbf{V}_{sr} & \mathbf{V}_s\end{pmatrix}\right),$$

entonces la distribución de $\mathbf{Y}_r \mid \mathbf{Y}_s, \boldsymbol{\beta}$ es una distribución normal multivariante con

$$E(\mathbf{Y}_r \mid \mathbf{Y}_s, \boldsymbol{\beta}) = \mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})$$

y

$$Var(\mathbf{Y}_r \mid \mathbf{Y}_s, \boldsymbol{\beta}) = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr}.$$

Ahora, $\hat{\boldsymbol{\beta}}_s$ es simplemente una matriz de constantes multiplicado por $\mathbf{Y}_s$, por lo tanto, la distribución de $\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s, \boldsymbol{\beta}$ también es normal multivariante.

Por otro lado, la suficiencia de $\hat{\boldsymbol{\beta}}_s$ implica que la distribución de interés $\mathbf{Y}_r \mid \mathbf{Y}_s$ es la misma que $\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s$. Tenemos que

$$\begin{aligned} p(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s) &= \frac{p(\mathbf{Y}_r, \hat{\boldsymbol{\beta}}_s)}{p(\hat{\boldsymbol{\beta}}_s)} \\ &= \frac{\int p(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s) p(\hat{\boldsymbol{\beta}}_s) d\boldsymbol{\beta}}{p(\hat{\boldsymbol{\beta}}_s)} \\ &= \int p(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s) d\boldsymbol{\beta}. \end{aligned} \tag{11.3.2}$$

Como se comentó antes, $\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s, \boldsymbol{\beta}$ tiene distribución normal multivariante, y por consiguiente, $p(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s, \boldsymbol{\beta})$ contiene una forma cuadrática de $\mathbf{Y}_r$. Luego, $p(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s)$ también contiene esa forma cuadrática, y se concluye que la distribución de $\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s$ es normal multivariante. Entonces se tiene que $\mathbf{Y}_r \mid \mathbf{Y}_s$ también tiene distribución normal multivariante.

Ahora,

$$\begin{aligned} E(\mathbf{Y}_r \mid \mathbf{Y}_s) &= E(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_r \mid \mathbf{Y}_s) \\ &= E(\mathbf{X}_r\boldsymbol{\beta} \mid \mathbf{Y}_s) + E(\mathbf{V}_r \mid \mathbf{Y}_s) \\ &= \mathbf{X}_r E(\boldsymbol{\beta} \mid \mathbf{Y}_s) + \mathbf{V}_r \\ &= \mathbf{X}_r E(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s) + E(\mathbf{V}_r \mid \hat{\boldsymbol{\beta}}_s) \\ &= E(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_r \mid \hat{\boldsymbol{\beta}}_s) \\ &= E(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s) \end{aligned}$$

pero

$$\begin{aligned} E(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s) &= E(E(\mathbf{Y}_r \mid \mathbf{Y}_s, \boldsymbol{\beta}) \mid \hat{\boldsymbol{\beta}}_s) \\ &= E(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \mid \hat{\boldsymbol{\beta}}_s) \\ &= \mathbf{X}_r E(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s) + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s E(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s)) \\ &= \mathbf{X}_r\hat{\boldsymbol{\beta}}_B + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}_B). \end{aligned}$$

De donde se concluye que $E(\mathbf{Y}_r \mid \mathbf{Y}_s) = \mathbf{X}_r\hat{\boldsymbol{\beta}}_B + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}_B)$.

Análogamente, se tiene que

$$\begin{aligned} Var(\mathbf{Y}_r \mid \mathbf{Y}_s) &= Var(\mathbf{Y}_r \mid \hat{\boldsymbol{\beta}}_s) \\ &= E(Var(\mathbf{Y}_r \mid \mathbf{Y}_s, \boldsymbol{\beta}) \mid \hat{\boldsymbol{\beta}}_s) + Var(E(\mathbf{Y}_r \mid \mathbf{Y}_s, \boldsymbol{\beta}) \mid \hat{\boldsymbol{\beta}}_s) \\ &= E(\mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr} \mid \hat{\boldsymbol{\beta}}_s) + Var(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \mid \hat{\boldsymbol{\beta}}_s) \\ &= \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr} + Var(\mathbf{X}_r\boldsymbol{\beta} - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{X}_s\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s) \\ &= \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr} + (\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{X}_s)Var(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_s)(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{X}_s)' \\ &= \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr} + (\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{X}_s)(\mathbf{X}_s'\mathbf{V}_s\mathbf{X}_s + \mathbf{B}^{-1})^{-1}(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{X}_s)'. \end{aligned}$$

Con lo anterior, se completa la demostración. ■

Con base en el anterior resultado, se procede a la utilización de XXXXXX para poder calcular las predicciones para los parámetros de interés en la encuesta o el estudio por muestreo.

Resultado 11.3.2. *Para cualquier cantidad lineal poblacional* $\mathbf{l}'\mathbf{Y}$*, el predictor Bayesiano que minimiza la pérdida del error cuadrático bajo el modelo simple está dado por*

$$\mathbf{l}_s'\mathbf{Y}_s + \mathbf{l}_r'\boldsymbol{\mu}_r \tag{11.3.3}$$

En particular la predicción para el total poblacional T_y está dada por la siguiente expresión

$$\hat{T}_y = \sum_{k\in s} Y_k + \mathbf{1}_r'\mu_r \tag{11.3.4}$$

donde $\mathbf{1}_r'$ es un vector de unos de tamaño $N - n$.

12 Modelos de sobrevivencia

12.1 Tópicos básicos

Suponga que se tiene acceso a una muestra de variables aleatorias $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$. Los elementos de $\mathbf{Y}$ denotan los tiempos de sobrevida de n individuos y la distribución de cada variable Y_i está parámetrizada por el vector $\boldsymbol{\theta}$. Además, se tiene para cada uno de los individuos que participan en el estudio un vector indicador de censura $\mathbf{d} = (d_1, d_2, \ldots, d_n)'$ que toma los siguientes valores

$$d_i = \begin{cases} 0, & \text{si } Y_i \text{ está censurado a la derecha,} \\ 1, & \text{si } Y_i \text{ representa el tiempo de falla o de muerte.} \end{cases}$$

Uno de los objetivos fundamentales para el investigador es inferir acerca de la función de sobrevida[1], definida para todos los valores de y como $S(y \mid \boldsymbol{\theta})$ dada por

$$S(y \mid \boldsymbol{\theta}) = Pr(Y > y) = \int_y^\infty p(u \mid \theta)\, du \tag{12.1.1}$$

Antes de especificar una distribución previa para el vector de parámetros, nótese que en realidad el vector de observaciones incluye las realizaciones de las variables de interés, así como del indicador de censura y el tamaño de muestra. De esta manera, y siguiendo la notación de **?**, el modelamiento bayesiano auténtico debe ser realizado en términos de todos los datos observados notados como $\mathbf{D} = (\mathbf{Y}, \mathbf{d})$. Bajo el contexto del análisis de sobrevivencia, la función de verosimilitud de los datos está dada por el siguiente resultado.

Resultado 12.1.1. *Bajo censura a la derecha, la verosimilitud conjunta de los datos observados* $\mathbf{D} = (\mathbf{Y}, \mathbf{d})$*está dada por la siguiente expresión:*

$$p(\mathbf{D} \mid \boldsymbol{\theta}) = c \prod_{i=1}^{n} \left[p(y_i \mid \boldsymbol{\theta})\right]^{d_i} \left[S(y_i \mid \boldsymbol{\theta})\right]^{1-d_i} \tag{12.1.2}$$

donde c *es una constante y* $S(\cdot \mid \boldsymbol{\theta})$ *es la función de sobrevida.*

Prueba. La demostración de este resultado se sale de los alcances de este libro pero puede ser consultada en (**?**). ■

[1] **?** afirma que ésta es el corazón del análisis de sobrevida.

12.1.1 Censura y truncamiento

12.2 Modelo Exponencial

Para este caso se supone una muestra aleatoria $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ de tiempos de sobrevida cuyos componentes tienen distribución exponencial con parámetro θ. Bajo este marco de referencia, la función de sobrevida está dada por

$$S(y \mid \theta) = \exp\{-\theta y\} \tag{12.2.1}$$

De acuerdo al resultado 12.2.1, la función de verosimilitud de los datos observados estaría dada por las siguientes expresiones.

$$\begin{aligned}
p(\mathbf{D} \mid \boldsymbol{\theta}) &\propto \prod_{i=1}^{n} [p(y_i \mid \theta)]^{d_i} [S(y_i \mid \theta)]^{1-d_i} \\
&= \prod_{i=1}^{n} [\theta \exp\{-\theta y_i\}]^{d_i} [\exp\{-\theta y_i\}]^{1-d_i} \\
&= \prod_{i=1}^{n} \left[\theta^{d_i} \exp\{-\theta y_i d_i\}\right]^{d_i} [\exp\{-\theta y_i + \theta d_i y_i\}]^{1-d_i} \\
&= \prod_{i=1}^{n} \theta^{d_i} \exp\{-\theta y_i\} \\
&= \theta^{\sum_{i=1}^{n} d_i} \exp\left\{-\theta \sum_{i=1}^{n} y_i\right\}
\end{aligned}$$

Dado lo anterior, se tiene que la verosimilitud depende de θ y por consiguiente, de acuerdo al enfoque bayesiano, es necesario proponer una distribución previa para este parámetro. Como θ vuiene de la distribución exponencial, se propone que su distribución previa sea modelada mediante una Gamma de parámetros (α, β). De acuerdo a este contexto, se tiene la siguiente deducción.

Resultado 12.2.1. *La distribución posterior del parámetro θ es*

$$\theta \mid \mathbf{D} \sim Gamma\left(\alpha + \sum_{i=1}^{n} d_i, \beta + \sum_{i=1}^{n} y_i\right)$$

Prueba.

$$\begin{aligned}
p(\theta \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \theta) p(\theta \mid \alpha, \beta) \\
&= \theta^{\sum_{i=1}^{n} d_i} \exp\left\{-\theta \sum_{i=1}^{n} y_i\right\} \frac{\beta^{\alpha} \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)} \\
&\propto \theta^{\sum_{i=1}^{n} d_i + \alpha - 1} \exp\left\{-\theta \left(\sum_{i=1}^{n} y_i + \beta\right)\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma\,(\alpha + \sum_{i=1}^{n} d_i, \beta + \sum_{i=1}^{n} y_i)$. ■

Como lo afirman **?**, un aspecto importante del enfoque bayesiano es la predicción. Más aún en el análisis de sobrevivencia, es muy útil predecir la probabilidad de que un individuo sobreviva en un determinado lapso de tiempo. Para este contexto se deja de lado la censura y se trabaja con la verosimilitud de la muestra aleatoria.

Resultado 12.2.2. *Después de la recolección de los datos, la distribución predictiva posterior para un nuevo tiempo $\tilde{y}$ está dada por*

$$p(\tilde{y} \mid \mathbf{D}) = \frac{\Gamma(\sum_{i=1}^{n} d_i + \alpha + 1)}{(\tilde{y} + \sum_{i=1}^{n} y_i + \beta)^{\sum_{i=1}^{n} d_i + \alpha + 1}} \tag{12.2.2}$$

Prueba. De la definición de la distribución predictiva posterior dada en la expresión (1.4.9), se tiene que

$$\begin{aligned}
p(\tilde{y} \mid \mathbf{D}) &= \int p(\tilde{y} \mid \theta) p(\theta \mid \mathbf{D})\, d\theta \\
&\propto \int_0^{\infty} \theta \exp\{-\theta\tilde{y}\} \theta^{\sum_{i=1}^{n} d_i + \alpha - 1} \exp\left\{-\theta\left(\sum_{i=1}^{n} y_i + \beta\right)\right\}\, d\theta \\
&= \frac{\Gamma(\sum_{i=1}^{n} d_i + \alpha + 1)}{(\tilde{y} + \sum_{i=1}^{n} y_i + \beta)^{\sum_{i=1}^{n} d_i + \alpha + 1}} \\
&\qquad \times \int_0^{\infty} \frac{(\tilde{y} + \sum_{i=1}^{n} y_i + \beta)^{\sum_{i=1}^{n} d_i + \alpha + 1}}{\Gamma(\sum_{i=1}^{n} d_i + \alpha + 1)} \\
&\qquad \times \exp\left\{-\theta\left(\tilde{y} + \sum_{i=1}^{n} y_i + \beta\right)\right\} \theta^{\sum_{i=1}^{n} d_i + \alpha + 1 - 1}\, d\theta \\
&= \frac{\Gamma(\sum_{i=1}^{n} d_i + \alpha + 1)}{(\tilde{y} + \sum_{i=1}^{n} y_i + \beta)^{\sum_{i=1}^{n} d_i + \alpha + 1}}
\end{aligned}$$

■

Con los anteriores resultados, basta con simular observaciones de la distribución posterior de θ para encontrar estimaciones puntuales, que bajo el criterio de mínima pérdida cuadrática sería igual a la media de dichas observaciones, intervalos de credibilidad y varianza. Nótese que con cada una de estas observaciones simuladas de θ se tiene un valor para la función de sobrevivencia S en un tiempo t; es decir, se tiene un número igual de realizaciones de S que el número de observaciones simuladas para θ. Con esto, es posible obtener estimaciones puntuales e intervalos de credibilidad para S en un tiempo determinado.

También es posible modelar la media de la distribución exponencial mediante una relación lineal. Suponga que se tiene un conjunto de n variables aleatorias

intercambiables $\mathbf{Y} = \{Y_i, \ldots, Y_n\}$. Cada variable aleatoria Y_i se rige por una distribución de probabilidad exponencial de parámetro θ_i. Además se supone que q vectores auxiliares, $(\mathbf{X}_1, \ldots, \mathbf{X}_q)'$ tienen una relación de causalidad con una función de la media de Y_i; en esta sección se trabajará con la función exponencial[2], de tal forma que $\theta_i = \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}$. El modelo planteado se da a continuación.

$$\boldsymbol{\theta} = (\exp\{\mathbf{X}_1'\boldsymbol{\beta}\}, \ldots, \exp\{\mathbf{X}_n'\boldsymbol{\beta}\})' \tag{12.2.3}$$

Donde

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} \qquad \text{y} \qquad \mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nq} \end{pmatrix}$$

Bajo esta perspectiva, los datos observados ahora incluyen al vector de observaciones de tiempos de sobrevivencia, el vector indicador de censura y la matriz de observaciones de las variables auxiliares; por lo tanto, la inferencia bayesiana de debe realizar en términos de $\mathbf{D} = (\mathbf{Y}, \mathbf{X}, \mathbf{d})$. Con base en lo anterior, la verosimilitud toma la siguiente forma.

$$\begin{aligned}
p(\mathbf{D} \mid \boldsymbol{\beta}) &= \prod_{i=1}^{n} \left[p(y_i \mid \boldsymbol{\beta})\right]^{d_i} \left[S(y_i \mid \boldsymbol{\beta})\right]^{1-d_i} \\
&= \prod_{i=1}^{n} \left[\exp\{\mathbf{X}_i'\boldsymbol{\beta}\} \exp\{-\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}y_i\}\right]^{d_i} \left[\exp\{-\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}y_i\}\right]^{1-d_i} \\
&= \prod_{i=1}^{n} \left[\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}^{d_i} \exp\{-\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}y_i d_i\}\right]^{d_i} \left[\exp\{-\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}y_i + \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}d_i y_i\}\right]^{1-d_i} \\
&= \prod_{i=1}^{n} \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}^{d_i} \exp\{-\exp\{\mathbf{X}_i'\boldsymbol{\beta}\}y_i\} \\
&= \exp\left\{\sum_{i=1}^{n} d_i \mathbf{X}_i'\boldsymbol{\beta}\right\} \exp\left\{-\sum_{i=1}^{n} y_i \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}\right\}
\end{aligned}$$

El siguiente paso es la escogencia de la distribución previa para el vector de parámetros $\boldsymbol{\beta}$. Una sugerencia natural para esta distribución es la normal multivariante. De esta manera, se tiene que:

$$\boldsymbol{\beta} \sim Normal_q(\mathbf{b}, \mathbf{B})$$

Nótese que en el caso en donde no se tenga conocimiento previa del problema, es posible asignar a $\mathbf{B}$ como una matriz diagonal cuyas entradas son lo suficientemente grandes como para reflejar la incertidumbre de la situación. Bajo este contexto se tiene el siguiente resultado.

[2]Aunque es posible trabajar con otras funciones como $\theta_i = (\mathbf{X}_i'\boldsymbol{\beta})^{-1}$

Resultado 12.2.3. *La distribución posterior del vector de hiperparámetros* $\boldsymbol{\beta}$ *es*

$$p(\boldsymbol{\beta} \mid \mathbf{D}) = p(\beta_1, \beta_2 \ldots, \beta_q \mid \mathbf{D}) \propto \exp\left\{\sum_{i=1}^{n} d_i \mathbf{X}_i'\boldsymbol{\beta} - \sum_{i=1}^{n} y_i \exp(\mathbf{X}_i'\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b})\right\}$$

Por el anterior resultado, la distribución posterior no tiene una forma conocida, y es necesario utilizar la técnica del condicionamiento sucesivo junto con el método de la grilla para simular observaciones correspondientes a esta distribución multivariante. Con esto en mente, un algoritmo permite la simulación de estas observaciones es el siguiente[3].

Algoritmo de Gibbs

- Fijar valores iniciales para el vector de parámetros de interés; por ejemplo, estos valores iniciales pueden estar dados por $\boldsymbol{\beta}^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_q^0)'$.
- Para simular un valor de la distribución posterior condicional para β_0 se tiene en cuenta que

$$p(\beta_1 \mid \beta_2, \ldots, \beta_q, \mathbf{D}) \propto (\beta_1, \underbrace{\beta_2 \ldots, \beta_q}_{fijos} \mid \mathbf{D})$$

  Por tanto, utilizando los valores iniciales $\beta_2^0, \ldots, \beta_q^0$ en la distribución conjunta y dejando a β_1 como una variable, entonces es posible utilizar el método de la grilla para simular una nueva observación, β_1^1, de esta distribución univariada. Este nuevo valor β_1^1 reemplaza al valor β_1^0.
- Realizar el anterior procedimiento para simular una nueva observación β_2^1 de

$$p(\beta_2 \mid \beta_1, \beta_3 \ldots, \beta_q, \mathbf{D}) \propto (\beta_2, \underbrace{\beta_1, \beta_3 \ldots, \beta_q}_{fijos} \mid \mathbf{D})$$

- Repetir este proceso hasta que todos los componentes del vector de valores iniciales $\boldsymbol{\beta}^0$ sean reemplazados en su totalidad por un nuevo vector de valores dado por $\boldsymbol{\beta}^1 = (\beta_1^1, \beta_2^1, \ldots, \beta_q^1)'$.
- ...
- Simular un número grande de vectores $\boldsymbol{\beta}$ hasta obtener convergencia. Al final, todos los vectores simulados son considerados como realizaciones de la distribución multivariada posterior conjunta del vector de parámetros de interés dada por el anterior resultado.

Una vez que se tengan las observaciones simuladas de la distribución posterior del vector $\boldsymbol{\beta}$, es posible realizar todo tipo de inferencias posterior, no sólo para $\boldsymbol{\beta}$

[3]Para garantizar que este algoritmo pueda ser utilizado, es necesario corroborar que las distribuciones condicionales posterior son log-concavas.

sino también para la función de sobrevida que, para un individuo con vector de covariables $\mathbf{z}$ en un tiempo t, estaría dada por la siguiente expresión

$$S(t \mid \boldsymbol{\beta}) = \exp\{-\exp(\mathbf{z}'\boldsymbol{\beta})t\}. \tag{12.2.4}$$

12.3 Modelo Weibull

Para utilizar este modelo, es conveniente reparametrizar la función de densidad de la siguiente manera, haciendo $\lambda = -\theta\ln(\gamma)$. Con base en lo anterior se tiene que la densidad para una variable Y que se distribuye Weibull, está dada por

$$p(y \mid \theta, \lambda) = \theta y^{\theta-1}\exp\{\lambda - y^{\alpha}\exp(\lambda)\} \tag{12.3.1}$$

De esta forma, la función de sobrevivencia está dada por

$$S(y \mid \theta, \lambda) = \exp\left\{\ -y^{\theta}\exp(\lambda)\right\} \tag{12.3.2}$$

La función de verosimilitud de los datos observados está dada por las siguientes expresiones.

$$\begin{aligned}
p(\mathbf{D} \mid \theta, \lambda) &= \prod_{i=1}^{n}\left[p(y_i \mid \theta)\right]^{d_i}\left[S(y_i \mid \theta)\right]^{1-d_i} \\
&= \prod_{i=1}^{n}\left[\theta y^{\theta-1}\exp\{\lambda - y^{\alpha}\exp(\lambda)\}\right]^{d_i}\left[\exp\left\{\ -y^{\theta}\exp(\lambda)\right\}\right]^{1-d_i} \\
&= \theta^{\sum_{i=1}^{n} d_i}\prod_{i=1}^{n} y_i^{d_i(\theta-1)}\exp\left\{\lambda\sum_{i=1}^{n} d_i - \sum_{i=1}^{n} y_i^{\theta}\exp(\lambda)\right\} \\
&= \theta^{\sum_{i=1}^{n} d_i}\exp\left\{\sum_{i=1}^{n} d_i(\theta-1)\ln(y_i) + \lambda\sum_{i=1}^{n} d_i - \sum_{i=1}^{n} y_i^{\theta}\exp(\lambda)\right\}
\end{aligned}$$

Luego, la verosimilitud depende de θ y de λ y por consiguiente es necesario proponer una distribución previa para estos parámetros. Como θ debe ser positivo, se propone que su distribución previa sea modelada mediante una Gamma de parámetros (α, β). Por otro lado, como λ es el logaritmo de un número positivo, de acuerdo a la parametrización, entonces su rango son los reales, por lo tanto es plausible proponer una distribución previa normal de parámetros (μ, σ^2). De acuerdo a la anterior formulación, se tiene la siguiente deducción.

Resultado 12.3.1. *Considerando a θ independiente de λ, su distribución posterior conjunta es*

$$\begin{aligned}
p(\theta, \alpha \mid \mathbf{D}) &\propto \theta^{\sum_{i=1}^{n} d_i + \alpha - 1} \\
&\quad \times \exp\left\{\lambda\sum_{i=1}^{n} d_i + \sum_{i=1}^{n}\left[d_i(\theta-1)\ln(y_i) + y_i^{\theta}\exp(\lambda)\right] - \frac{1}{2\sigma^2}(\lambda-\mu)^2 - \beta\theta\right\}
\end{aligned}$$

Prueba. De la definición de distribución posterior, se tiene que

$$
\begin{aligned}
p(\theta,\lambda \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \theta,\lambda)p(\theta \mid \alpha,\beta)p(\lambda \mid \mu,\sigma^2) \\
&= \theta^{\sum_{i=1}^n d_i} exp\left\{\sum_{i=1}^n d_i(\theta-1)\ln(y_i) + \lambda\sum_{i=1}^n d_i - \sum_{i=1}^n y_i^\theta \exp(\theta)\right\} \\
&\qquad \times \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp\{-\beta\theta\} \\
&\qquad \times \frac{1}{\sqrt{2\pi\sigma}}\exp\left\{\frac{-1}{2\sigma^2}(\lambda-\mu)^2\right\} \\
&\propto \theta^{\sum_{i=1}^n d_i} \\
&\times \exp\left\{\lambda\sum_{i=1}^n d_i + \sum_{i=1}^n \left[d_i(\theta-1)\ln(y_i) + y_i^\theta\exp(\lambda)\right] - \frac{1}{2\sigma^2}(\lambda-\mu)^2 - \beta\theta\right\}
\end{aligned}
$$

■

La anterior distribución bivariada no tiene una forma conocida y no está implementada en ningún paquete estadístico. Sin embargo, utilizando el algoritmo de Gibbs descrito en la sección anterior es posible simular observaciones provenientes de esta distribución posterior conjunta. Basado en esto, basta con simular observaciones de la distribución posterior de $(\theta,\lambda)'$ para encontrar estimaciones puntuales, que bajo el criterio de mínima pérdida cuadrática sería igual a la media de dichas observaciones, intervalos de credibilidad y varianza. Nótese que con cada una de estas observaciones simuladas de θ se tiene un valor para la función de sobrevivencia S.

Cuando se cuentan con covariables, es posible modelar el parámetro λ mediante una relación lineal. Si se tiene un conjunto de n variables aleatorias intercambiables $\mathbf{Y} = \{Y_i,\ldots,Y_n\}$, cada variable aleatoria Y_i se rige por una distribución de probabilidad Weibull de parámetros (θ,λ_i). Además se supone que q vectores auxiliares, $(\mathbf{X}_1,\ldots,\mathbf{X}_q)'$ tienen una relación de causalidad con una función de la media de Y_i de tal forma que $\lambda_i = \mathbf{X}_i'\boldsymbol{\beta}$. Por lo tanto, la verosimilitud de los datos toma la siguiente forma.

$$
\begin{aligned}
p(\mathbf{D} \mid \theta,\boldsymbol{\beta}) &= \prod_{i=1}^n \left[p(y_i \mid \theta)\right]^{d_i}\left[S(y_i \mid \theta)\right]^{1-d_i} \\
&= \prod_{i=1}^n \left[\theta y^{\theta-1}\exp\left\{\mathbf{X}_i'\boldsymbol{\beta} - y^\alpha\exp(\mathbf{X}_i'\boldsymbol{\beta})\right\}\right]^{d_i}\left[\exp\left\{\ -y^\theta\exp(\mathbf{X}_i'\boldsymbol{\beta})\right\}\right]^{1-d_i} \\
&= \theta^{\sum_{i=1}^n d_i}\prod_{i=1}^n y_i^{d_i(\theta-1)}\exp\left\{\sum_{i=1}^n d_i\mathbf{X}_i'\boldsymbol{\beta} - \sum_{i=1}^n y_i^\theta\exp(\mathbf{X}_i'\boldsymbol{\beta})\right\} \\
&= \theta^{\sum_{i=1}^n d_i}\exp\left\{\sum_{i=1}^n d_i(\theta-1)\ln(y_i) + \sum_{i=1}^n d_i\mathbf{X}_i'\boldsymbol{\beta} - \sum_{i=1}^n y_i^\theta\exp(\mathbf{X}_i'\boldsymbol{\beta})\right\}
\end{aligned}
$$

Como es usual, la distribución previa para el vector de parámetros $\boldsymbol{\beta}$ es la normal multivariante de vector de medias $\mathbf{b}$ y matriz de covarianzas $\mathbf{B}$. Suponiendo que los parámetros de $\boldsymbol{\beta}$ son independientes del parámetro θ y asignándole a este una distribución previa gamma (α, β). Bajo este contexto se tiene el siguiente resultado.

Resultado 12.3.2. *La distribución posterior del vector de hiperparámetros $\boldsymbol{\beta}$ es*

$$\begin{aligned} p(\boldsymbol{\beta}, \theta \mid \mathbf{D}) &= p(\beta_1, \beta_2 \ldots, \beta_q, \theta \mid \mathbf{D}) \\ &\propto \theta^{\sum_{i=1}^{n} d_i + \alpha - 1} \exp\left\{ \sum_{i=1}^{n} \left[d_i \mathbf{X}_i' \boldsymbol{\beta} + d_i(\theta - 1)\ln(y_i) + y_i^{\theta} \exp(\mathbf{X}_i' \boldsymbol{\beta}) \right] \right\} \\ &\quad \times \exp\left\{ -\beta\theta - \frac{1}{2}(\boldsymbol{\beta} - b)' \mathbf{B}^{-1} (\boldsymbol{\beta} - b) \right\} \end{aligned}$$

Por el anterior resultado, la distribución posterior no tiene una forma conocida, y es necesario utilizar la técnica del condicionamiento sucesivo junto con el método de la grilla para simular observaciones correspondientes a esta distribución multivariante. Con esto en mente, el algoritmo de Gibbs permite la simulación de estas observaciones.

12.4 Modelo de valor-extremo

El modelo probabilístico de valor extremo es reconocido como una familia de distribuciones de probabilidad que intentan modelar eventos raros cuya ocurrencia no es frecuente. Entre las distribuciones de probabilidad que pertenecen a esta familia se encuentran la distribución de Weibull, Rayleigh, entre otras. Una parametrización conveniente y ampliamente usada se da ac continuación. Se supone que se tiene una muestra aleatoria de variables de sobrevivencia $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ cada una distribuida de acuerdo a la siguiente densidad

$$p(y \mid \theta, \lambda) = \theta \exp(\theta y) \exp\{\lambda - \exp(\lambda + \theta y)\} \tag{12.4.1}$$

Para este modelo, la función de sobrevivencia está dada por

$$S(y \mid \theta, \lambda) = \exp\{ -\exp(\lambda + \theta y)\} \tag{12.4.2}$$

La función de verosimilitud de los datos observados está dada por las siguientes

expresiones.

$$
\begin{aligned}
p(\mathbf{D} \mid \theta, \lambda) &= \prod_{i=1}^{n} [p(y_i \mid \theta)]^{d_i} [S(y_i \mid \theta)]^{1-d_i} \\
&= \theta^{\sum_{i=1}^{n} d_i} \exp(\theta \sum_{i=1}^{n} y_i) \exp\left\{ \lambda \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} d_i \exp(\lambda + \theta y_i) \right\} \\
&\qquad \times \exp\left\{ -\sum_{i=1}^{n} \exp(\lambda + \theta y_i) \right\} \exp\left\{ -\sum_{i=1}^{n} d_i \exp(\lambda + \theta y_i) \right\} \\
&= \theta^{\sum_{i=1}^{n} d_i} \exp\left\{ \theta \sum_{i=1}^{n} y_i d_i + \lambda \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} \exp(\lambda + \theta y_i) \right\}
\end{aligned}
$$

Como la verosimilitud depende de θ y de λ, es necesario proponer una distribución previa para estos parámetros. Como θ debe ser positivo, se propone que su distribución previa sea modelada mediante una Gamma de parámetros (α, β). Por otro lado, para λ es posible proponer una distribución previa normal de parámetros (μ, σ^2). De acuerdo a la anterior formulación, se tiene el siguiente resultado.

Resultado 12.4.1. *Considerando a θ independiente de λ, su distribución posterior conjunta es*

$$
\begin{aligned}
p(\theta, \alpha \mid \mathbf{D}) &\propto \theta^{\sum_{i=1}^{n} d_i + \alpha - 1} \\
&\qquad \times \exp\left\{ \sum_{i=1}^{n} [d_i y_i \theta + \lambda d_i - \exp(\lambda + \theta y_i)] - \frac{1}{2\sigma^2}(\lambda - \mu)^2 - \beta\theta \right\}
\end{aligned}
$$

Prueba. De la definición de distribución posterior, se tiene que

$$
\begin{aligned}
p(\theta, \lambda \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \theta, \lambda) p(\theta \mid \alpha, \beta) p(\lambda \mid \mu, \sigma^2) \\
&= \theta^{\sum_{i=1}^{n} d_i} \exp\left\{ \theta \sum_{i=1}^{n} y_i d_i + \lambda \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} \exp(\lambda + \theta y_i) \right\} \\
&\qquad \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha - 1} \exp\{-\beta\theta\} \\
&\qquad \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-1}{2\sigma^2}(\lambda - \mu)^2 \right\} \\
&\propto \theta^{\sum_{i=1}^{n} d_i + \alpha - 1} \\
&\qquad \times \exp\left\{ \sum_{i=1}^{n} [d_i y_i \theta + \lambda d_i - \exp(\lambda + \theta y_i)] - \frac{1}{2\sigma^2}(\lambda - \mu)^2 - \beta\theta \right\}
\end{aligned}
$$

■

La anterior distribución bivariada no tiene una forma conocida, pero utilizando el algoritmo de Gibbs descrito en las secciones anteriores es posible simular observaciones provenientes de esta distribución posterior conjunta. Basado en esto,

basta con simular observaciones de la distribución posterior de $(\theta, \lambda)'$ para encontrar estimaciones puntuales, que bajo el criterio de mínima pérdida cuadrática sería igual a la media de dichas observaciones, intervalos de credibilidad y varianza. Nótese que con cada una de estas observaciones simuladas de θ se tiene un valor para la función de sobrevivencia S.

Al igual que en los anteriores modelos, cuando se cuenta con covariables, es posible modelar el parámetro λ mediante una relación lineal. Si se tiene un conjunto de n variables aleatorias intercambiables $\mathbf{Y} = \{Y_i, \ldots, Y_n\}$, cada variable aleatoria Y_i se rige por una distribución de probabilidad Weibull de parámetros (θ, λ_i). Además se supone que q vectores auxiliares, $(\mathbf{X}_1, \ldots, \mathbf{X}_q)'$ tienen una relación de causalidad con una función de la media de Y_i de tal forma que $\lambda_i = \mathbf{X}_i'\boldsymbol{\beta}$. Por lo tanto, la verosimilitud de los datos toma la siguiente forma.

$$
\begin{aligned}
p(\mathbf{D} \mid \theta, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \left[p(y_i \mid \theta)\right]^{d_i} \left[S(y_i \mid \theta)\right]^{1-d_i} \\
&= \theta^{\sum_{i=1}^{n} d_i} \exp(\theta \sum_{i=1}^{n} y_i) \exp\left\{\sum_{i=1}^{n} d_i \mathbf{X}_i'\boldsymbol{\beta} - \sum_{i=1}^{n} d_i \exp(\mathbf{X}_i'\boldsymbol{\beta} + \theta y_i)\right\} \\
&\qquad \times \exp\left\{ -\sum_{i=1}^{n} \exp(\mathbf{X}_i'\boldsymbol{\beta} + \theta y_i)\right\} \exp\left\{ -\sum_{i=1}^{n} d_i \exp(\mathbf{X}_i'\boldsymbol{\beta} + \theta y_i)\right\} \\
&= \theta^{\sum_{i=1}^{n} d_i} \exp\left\{\theta \sum_{i=1}^{n} y_i d_i + \mathbf{X}_i'\boldsymbol{\beta} \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} \exp(\mathbf{X}_i'\boldsymbol{\beta} + \theta y_i)\right\}
\end{aligned}
$$

Como es usual, la distribución previa para el vector de parámetros $\boldsymbol{\beta}$ es la normal multivariante de vector de medias $\mathbf{b}$ y matriz de covarianzas $\mathbf{B}$. Suponiendo que los parámetros de $\boldsymbol{\beta}$ son independientes del parámetro θ y asignándole a este una distribución previa gamma (α, β). Bajo este contexto se tiene la siguiente consecuencia.

Resultado 12.4.2. *La distribución posterior del vector de hiperparámetros $\boldsymbol{\beta}$ es*

$$
\begin{aligned}
p(\boldsymbol{\beta}, \theta \mid \mathbf{D}) &= p(\beta_1, \beta_2 \ldots, \beta_q, \theta \mid \mathbf{D}) \\
&\propto \theta^{\sum_{i=1}^{n} d_i + \alpha - 1} \exp\left\{\sum_{i=1}^{n} \left[\theta y_i d_i + d_i \mathbf{X}_i'\boldsymbol{\beta} - \exp(\mathbf{X}_i'\boldsymbol{\beta} + \theta y_i)\right]\right\} \\
&\qquad \times \exp\left\{-\beta\theta - \frac{1}{2}(\boldsymbol{\beta} - b)'\mathbf{B}^{-1}(\boldsymbol{\beta} - b)\right\}
\end{aligned}
$$

Por el anterior resultado, la distribución posterior no tiene una forma conocida, y es necesario utilizar la técnica del condicionamiento sucesivo junto con el método de la grilla para simular observaciones correspondientes a esta distribución multivariante. Con esto en mente, el algoritmo de Gibbs permite la simulación de estas observaciones.

12.5 Modelo log-normal

Suponga que se tiene una muestra aleatoria de variables de sobrevivencia $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ cada una distribuida de acuerdo a la siguiente densidad

$$p(Y \mid \theta, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\{\frac{-1}{2\sigma^2}(\ln(y) - \theta)^2\} \tag{12.5.1}$$

Para este modelo log-normal, la función de sobrevivencia está dada por

$$S(y \mid \theta, \sigma^2) = 1 - \Phi\left(\frac{\ln(y) - \theta}{\sigma}\right) \tag{12.5.2}$$

En donde Φ es la función de distribución acumulativa para la distribución normal estándar. La función de verosimilitud de los datos observados está dada por:

$$\begin{aligned} p(\mathbf{D} \mid \theta, \sigma^2) &= \prod_{i=1}^{n} \left[p(y_i \mid\mid \theta, \sigma^2)\right]^{d_i} \left[S(y_i \mid \theta, \sigma^2)\right]^{1-d_i} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\sum_{i=1}^{n} d_i} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n} d_i(\ln(y) - \theta)^2\right\} \\ &\qquad \times \prod_{i=1}^{n}(y_i)^{-d_i}\left[1 - \Phi\left(\frac{\ln(y) - \theta}{\sigma}\right)\right]^{1-d_i} \end{aligned}$$

Bajo esta situación, y siguiendo el enfoque del capítulo XXXXXX, la inferencia posterior de los parámetros de interés debe ser llevada a cabo en dos etapas: En la primera, se establece la distribución previa conjunta para ambos parámetros tal que

$$p(\theta, \sigma^2) = p(\theta \mid \sigma^2)p(\sigma^2)$$

Luego, en la segunda etapa ya es posible analizar posterior propiamente cada uno de los parámetros de interés puesto que

$$p(\theta, \sigma^2 \mid \mathbf{D}) \propto p(\mathbf{D} \mid \theta, \sigma^2)p(\theta \mid \sigma^2)p(\sigma^2)$$

La anterior formulación conlleva a asignar una distribución previa para θ dependiente del parámetro σ^2. Esto quiere decir que en la distribución $p(\theta \mid \sigma^2)$ el valor de σ^2 se considera una constante fija y conocida. Siguiendo los argumentos de la Sección 2.6, es plausible que la distribución previa de θ sea modelada como

$$p(\theta \mid \sigma^2) \sim Normal(\mu, \sigma^2/c_0)$$

Donde c_0 es una constante. Por otro lado, y siguiendo los argumentos de la sección 2.7, una posible opción para la distribución previa de σ^2, que no depende de θ, corresponde a

$$p(\sigma^2) \sim Gamma - inversa(n_0/2, n_0\sigma_0^2/2)$$

Resultado 12.5.1. *Considerando a θ dependiente de σ^2, su distribución posterior conjunta es*

$$p(\theta, \sigma^2 \mid \mathbf{D}) \propto (\sigma^2)^{\left(\frac{\sum_{i=1}^n d_i}{2} - \frac{1}{2} - \frac{n_0}{2}\right) - 1} \times \exp\left\{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^n d_i(\ln(y) - \theta)^2 + c_0(\theta - \mu)^2 + n_0\sigma_0^2\right]\right\} \times \prod_{i=1}^n (y_i)^{-d_i}\left[1 - \Phi\left(\frac{\ln(y) - \theta}{\sigma}\right)\right]^{1-d_i}$$

La anterior distribución bivariada no tiene una forma conocida, y **?** afirman que se debe utilizar un algoritmo de Metrópolis para simular observaciones provenientes de esta distribución posterior conjunta y así encontrar estimaciones puntuales, intervalos de credibilidad y varianza. Nótese que con cada una de estas observaciones simuladas de (θ, σ^2) se tiene un valor para la función de sobrevivencia S.

Al igual que en los anteriores modelos, cuando se cuenta con covariables, es posible modelar el parámetro θ mediante una relación lineal. Si se tiene un conjunto de n variables aleatorias intercambiables $\mathbf{Y} = \{Y_i, \ldots, Y_n\}$, cada variable aleatoria Y_i se rige por una distribución de probabilidad log-normal de parámetros (θ_i, σ^2). Además se supone que q vectores auxiliares, $(\mathbf{X}_1, \ldots, \mathbf{X}_q)'$ tienen una relación de causalidad con una función de la media de Y_i de tal forma que $\theta_i = \mathbf{X}_i'\boldsymbol{\beta}$. Por lo tanto, la verosimilitud de los datos toma la siguiente forma.

$$\begin{aligned} p(\mathbf{D} \mid \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^n \left[p(y_i \mid \boldsymbol{\beta}, \sigma^2)\right]^{d_i}\left[S(y_i \mid \boldsymbol{\beta}, \sigma^2)\right]^{1-d_i} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\sum_{i=1}^n d_i} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^n d_i(\ln(y) - \mathbf{X}_i'\boldsymbol{\beta})^2\right\} \\ &\quad \times \prod_{i=1}^n (y_i)^{-d_i}\left[1 - \Phi\left(\frac{\ln(y) - \mathbf{X}_i'\boldsymbol{\beta}}{\sigma}\right)\right]^{1-d_i} \end{aligned}$$

Los parámetros de interés son $\boldsymbol{\beta}$ y σ^2 y su distribuciones previa conjunta se supone que está dada por

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta} \mid \sigma^2)p(\sigma^2)$$

Específicamente, la distribución previa del parámetro $\boldsymbol{\beta}$ condicionada a σ^2 es informativa y está regida por la siguiente estructura probabilística

$$\boldsymbol{\beta} \mid \sigma^2 \sim Normal_q(\mathbf{b}, \sigma^2\mathbf{B})$$

en donde $\mathbf{b}$ es un vector de medias y $\mathbf{B}$ es una matriz de varianzas simétrica y definida positiva. Por otro lado, la distribución previa del parámetro σ^2 también se considera informativa y dada por

$$\sigma^2 \sim Inversa - Gamma\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$$

Resultado 12.5.2. *La distribución posterior del vector de hiperparámetros* $\boldsymbol{\beta}$ *es*

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{D}) \propto (\sigma^2)^{\left(\frac{\sum_{i=1}^{n} d_i}{2} - \frac{1}{2} - \frac{n_0}{2}\right) - 1} \times \exp\left\{ \frac{-1}{2\sigma^2} \left[\sum_{i=1}^{n} d_i (\ln(y) - \mathbf{X}_i' \boldsymbol{\beta})^2 + (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b}) + n_0 \sigma_0^2 \right] \right\} \times \prod_{i=1}^{n} (y_i)^{-d_i} \left[1 - \Phi\left(\frac{\ln(y) - \mathbf{X}_i' \boldsymbol{\beta}}{\sigma} \right) \right]^{1 - d_i}$$

Por el anterior resultado, la distribución posterior no tiene una forma conocida, y es posible utilizar la técnica del condicionamiento sucesivo junto con el método de la grilla para simular observaciones correspondientes a esta distribución multivariante. Con esto en mente, el algoritmo de Gibbs permite la simulación de estas observaciones.

13 Modelos en tablas de contingencia

13.1 Análisis de proporciones

Es cierto, a veces algunas técnicas estadísticas fallan. Más aun, a veces fallan técnicas que, por ser consideradas clásicas y robustas, no deberían de hacerlo. Es bien sabido que, con respecto a la diferencia de dos proporciones, los intervalos de confianza de Wald no son la mejor opción. Pues bien, el mismo espíritu de pensamiento que concibió estos intervalos es el que está detrás de la prueba de hipótesis clásica para dos proporciones: sí, la prueba aquella de dos colas que utiliza la normalidad, la prueba clásica, la del Canavos, la del comando prop.test en R.

Imagínese que a usted lo contratan en un juicio de discriminación racial. Una gran compañía metalúrgica enfrentada contra un sindicato de trabajadores de raza negra. La información es la siguiente: en el último periodo de contratación, de 80 personas de raza blanca, 41 fueron admitidos y 39 fueron rechazados; mientras que de 44 personas de raza negra, 14 fueron admitidos y 30 fueron rechazados. La proporción de admitidos de raza blanca es de casi el 50 %, pero la proporción de admitidos de raza negra es de apenas el 30 %.

Si utilizamos una prueba de proporciones clásica, llegaríamos a la conclusión de que la diferencia de proporciones es estadísticamente igual a cero. Por tanto, concluiríamos que no existe evidencia de discriminación racial. Sin emabargo, al utilizar un enfoque bayesiano, las conclusiones y el resultado final cambiarían la historia del juicio rotundamente (si quiere conocer el final de la historia de clic acá).

El objetivo de esta entrada es introducir al lector a un conjunto de soluciones computacionales, programadas en el ambiente de R, que permiten analizar, de forma bayesiana, los problemas concernientes al juzgamiento de hipótesis para dos proporciones. El conjunto de funciones (paquete propbayes) está disponible gratuitamente acá junto con un conjunto de ejemplos que le permitirán analizar de manera consistente sus datos.

En las últimas décadas el enfoque bayesiano ha sido uno de los tópicos más desarrollados en la ciencia estadística. Los avances computacionales, como los algoritmos de cadenas de Markov basados en simulaciones de Monte Carlo (MCMC), han hecho que la utilización de los métodos bayesianos sea cada vez mas común por parte del investigador. Sin embargo, como lo afirman **?**, las técnicas bayesianas no son muy usadas cuando se trata de la inferencia de tablas 2×2, siendo este

uno de los problemas más comunes en la práctica, específicamente el análisis de la diferencias de proporciones.

Este artículo está enfocado en el caso donde las distribuciones previa para cada una de las dos proporciones se rige por una distribución de tipo Beta[1], siguiendo los importantes resultados de (**?**), en donde se encontró la distribución exacta posterior de la diferencia de proporciones, la cual está en términos de la primera función hipergeométrica de Appell. Sin embargo, en el software estadístico `R`, a nuestro conocimiento actual, no existen funciones, rutinas o paquetes que calculen dicha función; en otras palabra, no es posible realizar inferencias bayesianas exactas para el problema en cuestión.

Basado en lo anterior, este articulo presenta un código computacional eficiente que permite realizar inferencia bayesiana tanto exacta como simulada mediante métodos MCMC[2]. Estas funciones son parte de un nuevo paquete del ambiente computacional de `R` desarrollado por los autores del presente artículo.

El artículo está divido de la siguiente forma: en la Sección 2 se presentan los resultados básicos de la inferencia bayesiana que son usados para la estimación y elaboración de los códigos computacionales. En la Sección 3 se presentan los resultados acerca de la distribución exacta posterior para la diferencia de proporciones y el algoritmo MCMC, que utiliza el muestreo de Gibbs, para la distribución simulada. En la Sección 4 se introduce la descripción de cada una de las funciones pertenecientes al paquete propuesto. En la Sección 5, se analiza un conjunto de datos usando estas funciones computacionales. Finalmente, en la Sección 6 se abordan algunas conclusiones acerca de la inferencia para la diferencia de proporciones.

Para llevar a cabo la inferencia bayesiana sobre la diferencia de proporciones, es posible calcular la función de densidad exacta previa y posterior de este parámetro, o bien usar el muestreo de Gibbs para obtener la función de densidad posterior simulada. Sin embargo, el software estadístico `R` actualmente no implementa rutinas o paquetes que calculen las funciones hipergeométricas de Appell que son usadas para obtener la densidad posterior de la diferencia de proporciones, y por consiguiente tampoco existen rutinas o paquetes que lleven a cabo el respectivo procedimiento bayesiano de forma exacta. Por esta razón, en este artículo se desarrollan funciones que permiten efectuar dicho procedimiento. Antes de introducir estas funciones, se presentan los resultados concernientes a la densidad exacta previa y posterior de la diferencia de proporciones y el muestreo de Gibbs.

13.1.1 Distribución exacta

Suponga que la distribución previa para las proporciones es $Beta(a_i, b_i)$ para el parámetro θ_i con $i = 1, 2$, es decir, $\theta_i \sim Beta(a_i, b_i)$. Como el parámetro de interés es la diferencia de las proporciones $\theta = \theta_1 - \theta_2$, el siguiente resultado de (**?**) provee la solución exacta para encontrar la distribución previa de θ.

[1] Cubriendo casos importantes como la distribución uniforme o la distribución no informativa de Jeffreys

[2] Dado que los resultados finales, aunque similares, no son idénticos.

Resultado 13.1.1. *Sea* $\theta_i \sim Beta(a_i, b_i)$ *con* $i = 1, 2$ *variables aleatorias independientes, entonces* $\theta = \theta_1 - \theta_2$ *tiene la siguiente función de densidad de probabilidad*

$$\pi(\theta \mid a_1, b_1, a_2, b_2) = \begin{cases} \dfrac{1}{A} B(a_2, b_1)\theta^{b_1+b_2-1}(1-\theta)^{a_2+b_1-1} \\ \qquad F_1(b_1, a_1+b_1+a_2+b_2-2, 1-a_1, b_1+a_2, 1-\theta, 1-\theta^2) \\ \qquad\qquad \textit{para} \quad 0 < \theta \leq 1 \\ \dfrac{1}{A} B(a_1+a_2-1, b_1+b_2-1) \qquad \textit{para} \quad \theta = 0 \\ \dfrac{1}{A} B(a_1, b_2)(-\theta)^{b_1+b_2-1}(1+\theta)^{a_1+b_2-1} \\ \qquad F_1(b_2, 1-a_2, a_1+b_1+a_2+b_2-2, b_2+a_1, 1-\theta^2, 1+\theta) \\ \qquad\qquad \textit{para} \quad -1 \leq \theta < 0 \end{cases} \tag{13.1.1}$$

donde $A = B(a_1, b_1)B(a_2, b_2)$*, con* $B(a, b)$ *la función beta evaluada en* a *y* b*, es decir,*

$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt. \tag{13.1.2}$$

Por otro lado $F_1(\varphi, \eta_1, \eta_2, \psi, w_1, w_2)$ *corresponde a la primera función hipergeométrica de Appell, dada por*

$$\frac{\Gamma(\psi)}{\Gamma(\varphi)\Gamma(\psi-\varphi)} \int_0^1 u^{\varphi-1}(1-u)^{\psi-\varphi-1}(1-uw_1)^{-\eta_1}(1-uw_2)^{-\eta_2} du, \tag{13.1.3}$$

cuando las partes reales de φ *y* $\psi - \varphi$ *son positivos, tal como lo muestra Bailey (1934).*

Ahora suponga que se observan los valores que toman las variables $X_1, \cdots, X_{n_1}$ y $Y_1, \cdots, Y_{n_2}$ que denotan el éxito o fracaso en cada uno de los n_1 y n_2 ensayos independientes. Entonces la distribución posterior de las proporciones es $p(\theta_1 \mid \mathbf{x}) = Beta(\alpha_1, \beta_1)$ y $p(\theta_2 \mid \mathbf{y}) = Beta(\alpha_2, \beta_2)$ donde $\alpha_i = a_i + x_i$ y $\beta_i = b_i + n_i - x_i$ para $i = 1, 2$.

Para el parámetro de interés $\theta = \theta_1 - \theta_2$ es posible hallar su distribución posterior usando el Teorema 1. Por tanto la función de densidad posterior dada las

observaciones de θ es

$$p(\theta \mid \mathbf{x}, \mathbf{y}) = \begin{cases} \dfrac{1}{A} B(\alpha_2, \beta_1)\theta^{\beta_1+\beta_2-1}(1-\theta)^{\alpha_2+\beta_1-1} \\ \qquad F_1(\beta_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, 1-\alpha_1, \beta_1+\alpha_2, 1-\theta, 1-\theta^2) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{para} \;\; 0 < \theta \leq 1 \\ \dfrac{1}{A} B(\alpha_1+\alpha_2-1, \beta_1+\beta_2-1) \qquad\qquad\qquad\qquad \text{para} \;\; \theta = 0 \\ \dfrac{1}{A} B(\alpha_1, \beta_2)(-\theta)^{\beta_1+\beta_2-1}(1+\theta)^{\alpha_1+\beta_2-1} \\ \qquad F_1(\beta_2, 1-\alpha_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, \beta_2+\alpha_1, 1-\theta^2, 1+\theta) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{para} \;\; -1 \leq \theta < 0 \end{cases} \tag{13.1.4}$$

donde la definición de A y F_1 son enunciados en el Resultado 1.

Dadas las distribuciones previa y posterior de θ es posible calcular la estimación puntual previa y posterior junto con el intervalo de credibilidad a *previa* y posterior. Para calcular la estimación puntual previa, tenemos

$$E(\theta) = E(\theta_1) - E(\theta_2) = \int_0^1 \theta_1 \pi(\theta_1 \mid a_1, b_1) d\theta_1 - \int_0^1 \theta_2 \pi(\theta_2 \mid a_2, b_2) d\theta_2. \tag{13.1.5}$$

Para calcular el intervalo de credibilidad previa, se debe encontrar dos valores l y u tales que

$$Pr(l \leq \theta \leq u) = 1 - \frac{\alpha}{2}.$$

En la práctica se escoge l y u de tal manera que $Pr(\theta < l) = Pr(\theta > u) = \alpha/2$. En consecuencia, se buscan valores l y u con

$$\int_{-1}^{l} \pi(\theta \mid a_1, b_1, a_2, b_2)\; d\theta = \int_u^1 \pi(\theta \mid a_1, b_1, a_2, b_2)\; d\theta = \frac{\alpha}{2}, \tag{13.1.6}$$

donde $\pi(\theta \mid a_1, b_1, a_2, b_2)$ es la función de densidad previa de θ en (**??**).

Cuando los valores de las variables han sido observadas, la estimación puntual posterior se define como

$$E(\theta \mid \mathbf{x}, \mathbf{y}) = E(\theta_1 \mid \mathbf{x}) - E(\theta_2 \mid \mathbf{y}) = \int_0^1 \theta_1 p(\theta_1 \mid \mathbf{x})\; d\theta_1 - \int_0^1 \theta_2 p(\theta_2 \mid \mathbf{y})\; d\theta_2. \tag{13.1.7}$$

Similarmente, el intervalo de credibilidad está dado por dos valores l y u tales que

$$Pr(l \leq \theta \leq u \mid \mathbf{x}, \mathbf{y}) = 1 - \frac{\alpha}{2}.$$

En consecuencia, se buscan valores l y u con

$$\int_{-1}^{l} p(\theta \mid \mathbf{x}, \mathbf{y})\; d\theta = \int_u^1 p(\theta \mid \mathbf{x}, \mathbf{y})\; d\theta = \frac{\alpha}{2}, \tag{13.1.8}$$

donde $p(\theta \mid \mathbf{x}, \mathbf{y})$ es la distribución posterior de θ dada por (**??**).

Además de estimar el parámetro θ, es posible obtener predicciones acerca de posibles resultados en nuevas muestras observadas por medio de la distribución predictiva posterior dada en (8). Para tal fin es necesario calcular la función de verosimilitud de los datos dada por

$$f(\mathbf{x}, \mathbf{y} \mid \theta_1, \theta_2) = \theta_1^{s_x}(1-\theta_1)^{n_1-s_x}\theta_2^{s_y}(1-\theta_2)^{n_2-s_y}. \tag{13.1.9}$$

donde $s_x = \sum_{i=1}^{n_1} x_i$, $s_y = \sum_{i=1}^{n_2} y_i$, n_1 y n_2 es el tamaño de muestras de las dos poblaciones y, x_i, y_i son las realizaciones de variables aleatorias con distribción Bernoulli. A continuación se ilustra el cálculo de la función predictiva con un ejemplo sencillo: suponga que se vuelven a observar dos muestras, ambas de tamaños 1; es decir, ahora existen dos nuevas variables $\tilde{X}$ y $\tilde{Y}$ que denotan el éxito o fracaso en cada ensayo de la muestra, respectivamente. Entonces, la probabilidad de que ambos ensayos tengan como resultado éxito está dada por

$$Pr(\tilde{X} = \tilde{Y} = 1) = \int_0^1 \int_0^1 \theta_1\theta_2 p(\theta_1 \mid \mathbf{x}) p(\theta_2 \mid \mathbf{y})\ d\theta_1 d\theta_2 \tag{13.1.10}$$

13.1.2 Distribución simulada

Desde otro punto de vista, es posible aplicar el muestreo de Gibbs para encontrar la estimación puntual y el respectivo intervalo de credibilidad. En este contexto las distribuciones tanto previa como posterior de θ_1 y θ_2 son independientes. Por lo tanto, las distribuciones condicionales posterior de θ_i, con $i = 1, 2$ en la j-ésima iteración son iguales a las respectivas distribuciones posterior de θ_i con $i = 1, 2$. Es decir $p(\theta_1 \mid \theta_2^{(j-1)}, \mathbf{x}, \mathbf{y}) = p(\theta_1 \mid \mathbf{x})$ y $p(\theta_2 \mid \theta_1^{(j-1)}, \mathbf{x}, \mathbf{y}) = p(\theta_2 \mid \mathbf{x})$. Nótese que en este caso, los resultados de la j-ésima iteración y los de la $j-1$-ésima iteración son independientes, y por consiguiente no hay necesidad de fijar valores iniciales. Esto conlleva a que, debido a la independencia, la convergencia del algoritmo se tenga desde su primera iteración. De esta manera, el algoritmo del muestreo de Gibbs se convierte en:

1. Fijar el número de iteraciones J.
2. Simular J observaciones de la distribuciones $p(\theta_1 \mid \mathbf{x})$ y $p(\theta_2 \mid \mathbf{y})$ respectivamente de manera que se dispone de valores $\theta_1^{(1)}, \cdots, \theta_1^{(J)}$ y $\theta_2^{(1)}, \cdots, \theta_2^{(J)}$.
3. Calcular la diferencia entre los valores simulados: $\theta_1^{(1)} - \theta_2^{(1)}, \cdots, \theta_1^{(J)} - \theta_2^{(J)}$. Cada uno de estos valores corresponde al valor de la distribución posterior de $\theta_1 - \theta_2$

Una vez termine el anterior algoritmo, es posible calcular la estimación puntual de $\theta = \theta_1 - \theta_2$ como el promedio $\sum_{j=1}^{N}(\theta_1^{(j)} - \theta_2^{(j)})/J$, y el intervalo de credibilidad donde los límites inferior y superior corresponden a los percentiles $\alpha/2$ y $1-\alpha/2$ de los valores $\theta_1^{(1)} - \theta_2^{(1)}, \cdots, \theta_1^{(J)} - \theta_2^{(J)}$, respectivamente.

13.2 funciones en R

Este artículo presenta una colección de funciones computacionales que son utilizadas en la implementación de un análisis bayesiano exhaustivo para la diferencia de dos proporciones. Con este fin, se discute la estimación puntual, estimación mediante intervalos de credibilidad y la inferencia predictiva desde dos escenarios: el primero basado en las densidades exactas previa y posterior (construidas mediante la primera función hipergeométrica de Appell) y el segundo basado en densidades simuladas (mediante un algoritmo de cadenas de Markov con métodos de Monte Carlo). La implementación de estas funciones se realiza en el software estadístico `R` puesto que es un software libre, funciona bien en múltiples plataformas y permite enmarcar estas funciones bajo un objeto computacional denominado paquete.

13.2.1 Recursos en la Internet

La página WEB `http://CRAN.R-project.org/` es la página oficial del software estadístico R **?**. En ésta se encuentran la descarga y la actualización del software `R`, además de numerosas librerías y paquetes específicos como el que se presenta en el presente artículo.

Por otro lado, está la página `http://predictive.wordpress.com/stats/propbayes/`, donde se encuentra la documentación y ayuda completa del paquete que contiene las funciones resultantes de este artículo. Los lectores interesados pueden descargar el paquete en esta página. Alternativamente, es posible importar las funciones del paquete directamente proveyendo la dirección URL apropiada como un argumento a la función fuente de `R`.

13.2.2 Descripción del código computacional

A continuación se introducen las funciones principales para llevar a cabo la metodología de este artículo.

- `F1` Esta función calcula la primera función de Appell dada por (**??**). La forma de uso de la función es `F1(A,B1,B2,C,X1,X2)`, donde `A`, `B1`, `B2` y `C` corresponden a los valores ϕ, η_1, η_2 y ψ; `X1` y `X2` corresponden a w_1 y w_2 respectivamente. Esta función utiliza la función `integrate` propia del ambiente `R`, la cual hace eficiente computacionalmente el análisis bayesiano para la diferencia de proporciones.
- `plot.dist` Esta función calcula y grafica la distribución exacta previa y posterior del parámetro de interés, $\theta = \theta_1 - \theta_2$, usando (**??**) y (**??**), respectivamente. La forma de uso de la función es `plot.dist(a1,b1,a2,b2,plot)`. Para calcular la distribución previa, `a1`, `b1`, `a2` y `b2` deben corresponder a los parámetros de la distribución previa de θ_1 y θ_2 respectivamente. Para calcular la distribución posterior, `a1`, `b1`, `a2` y `b2` deben corresponder a los parámetros

de la distribución posterior de θ_1 y θ_2 respectivamente. El argumento `plot` corresponde a la opción de graficar la función, cuando `plot=TRUE`, elabora la gráfica; y cuando `plot=FALSE`, sólo realiza los cálculos, omitiendo la gráfica de la función.

- `p.est` Esta función calcula la estimación puntual bayesiana previa o posterior de $\theta = \theta_1 - \theta_2$ dada por (**??**) y (**??**), respectivamente. Ésta usa los resultados de la función `plot.dist`. La forma de uso de la función es `p.est(a1,b1,a2,b2)` donde los cuatro argumentos corresponden a los mismos `a1,b1,a2,b2` de la función `plot.dist`. La estimación puntual bayesiana tanto previa como posterior se lleva a acabo usando esta función.
- `percentil` Esta función calcula los percentiles de la distribución previa o posterior del parámetro $\theta = \theta_1 - \theta_2$, y usa los resultados de la función `plot.dist`. Dada una probabilidad v, el percentil asociado con v es aquel valor a con $P(\theta < a) = v$. La forma de uso de la función es `percentil(val,a1,b1,a2,b2)`, donde `val` corresponde a la probabilidad v y los restantes cuatro argumentos corresponden a los mismos de la función `plot.dist`. El cálculo del intervalo de credibilidad previa o posterior se lleva a cabo usando esta función.
- `prob` Esta función calcula la probabilidad previa o posterior de que $\theta > d$ usando los resultados de la función `plot.dist`. La forma de uso de la función es `prob(val,a1,b1,a2,b2)`, donde `val` corresponde al valor d, y los restantes cuatro `a1,b1,a2,b2` corresponden a los mismos `a1,b1,a2,b2` de la función `plot.dist`. El cálculo del factor de Bayes, dao en (**??**), se lleva a cabo usando esta función.
- `plot.pred` Esta función calcula y grafica la función de densidad predictiva previa o posterior de la variable $S_x - S_y$ dada por (**??**) y (**??**), respectivamente, cuando se tienen dos nuevas muestras de tamaño n_1 y n_2. La forma de uso de la función es
  `plot.pred(a1,b1,a2,b2,n1,n2,plot)`. Los primeros cuatro argumentos corresponden a los argumentos de la función `plot.dist`; los dos argumentos siguientes corresponden a los tamaños de muestra n_1 y n_2; y el último argumento `plot` corresponde a la opción de graficación. Cuando `plot=TRUE`, elabora la gráfica; y cuando `plot=FALSE`, sólo realiza los cálculos, omitiendo la gráfica de la función predictiva.
- `plot.gibbs` Esta función calcula la estimación puntual, el intervalo de credibilidad posterior y grafica la función de densidad posterior para θ usando el muestreo de Gibbs descrito en la secciones anteriores. La forma de uso de la función es
  `plot.gibbs(a1,b1,a2,b2,nsim,plot,chain)`. Los argumentos `a1`, `b1`, `a2` y `b2` corresponden a los parámetros de las distribuciones posterior de θ_1 y θ_2, respectivamente. `nsim` corresponde al número de iteraciones del algoritmo. `plot` corresponde a la opción de graficar la distribución posterior simulada. Cuando `plot=TRUE`, elabora la gráfica y cuando `plot=FALSE`, omite la gráfica. `chain` corresponde a la opción de graficar los valores simulados de la distribución posterior, cuando `chain=TRUE`, muestra gráficamente estos valores simulados, y cuando `chain=FALSE`, omite la gráfica.

13.3 Aplicación: discriminación racial

En esta sección, se presenta una aplicación a datos reales de las funciones creadas en este artículo. La documentación de estas funciones está descrito en la Sección 4.2. y están disponibles en la página de internet dada en la Sección 4.1.

Los datos mostrados en la Tabla 1 fueron la base de una demanda en un caso de discriminación racial en 1980 entre los solicitantes de empleo en una fábrica de placas metálicas **?**.

Raza	Admitido	Rechazados	Total
Blanca	41	39	80
Negra	14	30	44
Total	55	69	124

Tabla 13.1: Tabla de conteos como sustento de discriminación racial.

Mediante el análisis estadístico de estos datos se debe responder a la siguiente pregunta: ¿Existe evidencia de discriminación racial?

Análisis frecuentista

El análisis estadístico de los datos necesita de un modelo cuyas características generales se dan a continuación: Suponga que θ_1 es la probabilidad de admisión al empleo de una persona de raza blanca y que θ_2 es la probabilidad de admisión al empleo de una persona de raza negra. Sea $X_i = 1$ si la i-ésima persona de raza blanca es admitida en el empleo y $X_i = 0$, en otro caso. De la misma manera, se define $Y_i = 1$ si la i-ésima persona de raza negra es admitida en el empleo y $Y_i = 0$, en otro caso.

Si asumimos que existe independencia entre y dentro de cada raza, y que θ_1 y θ_2 son constantes entre los individuos, entonces su admisión al empleo forma una secuencia de ensayos Bernoulli. Definiendo, $S_x = \sum_{i=1}^{80} X_i$ y $S_y = \sum_{i=1}^{44} Y_i$, se concluye que

$$S_x \mid \theta_1 \sim Binomial(80, \theta_1) \qquad S_y \mid \theta_2 \sim Binomial(44, \theta_1) \tag{13.3.1}$$

Bajo este marco de referencia, casi todos los textos básicos de inferencia estadística **?** proponen que la diferencia de dos proporciones, dada por

$$D = \frac{S_x}{n_1} - \frac{S_y}{n_1} \tag{13.3.2}$$

donde n_1 es el tamaño de muestra de los individuos de raza blanca, y n_2 es el tamaño de muestra de los individuos de raza negra, está bien aproximada mediante una distribución normal de media nula y de varianza $V_\theta(D) = (1/n_1 + 1/n_2)\theta(1-\theta)$. Se supone que los conteos tienen una distribución binomial con el mismo parámetro $\theta = \theta_1 = \theta_2$. Por tanto, para juzgar la hipótesis $\theta_1 = \theta_2$, se construye una nueva

variable aleatoria $U = D/\sqrt{V_{\hat{\theta}}(D)}$ que aproximadamente[3] tiene distribución normal estándar. También es posible utilizar la variable U^2 que aproximadamente tiene una distribución chi-cuadrado con un grado de libertad.

Para responder a la pregunta de interés, el investigador estaría tentado a realizar una prueba de diferencia de dos proporciones y a tomar una decisión con respecto al valor p arrojado por dicha prueba. A continuación se muestran los resultados arrojados por la función `prop.test` propia del ambiente computacional del software `R` **?**.

```
> n1<-80  ;  x1<-41
> n2<-44  ;  x2<-14
> prop.test(c(x1,x2),c(n1,n2))

        2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 3.5913, df = 1, p-value = 0.05808
alternative hypothesis: two.sided
```

De esta manera, para un nivel de significación del 5 %, no se rechaza la hipótesis de igualdad de proporciones. En otras palabras, no se encuentra evidencia de discriminación racial en la fábrica de placas metálicas.

Análisis bayesiano previa

Sin duda alguna, una de las herramientas más poderosas de la inferencia bayesiana es la definición de la distribución previa de los parámetros y como lo afirma (**?**) de la misma manera que no existe un principio general al definir una verosimilitud para una muestra aleatoria, en el caso frecuentista, tampoco existe un principio general para definir una distribución previa, en el caso bayesiano. Por lo anterior, esta etapa del análisis bayesiano debe ser recorrida con mucho cuidado.

Se supone que la distribución previa para la proporción de admisión θ_i es $Beta(a_i, b_i)$ con $i = 1, 2$. En la Figura 1 se observan dos candidatos miembros de la familia de las distribuciones Beta; estos son, $Beta(1, 1)$ y $Beta(2, 2)$. Nótese que la distribución $Beta(1, 1)$ se reduce a la distribución uniforme continua sobre el intervalo $(0, 1)$, la cual es una distribución previa no informativa y parece natural pensar que esta distribución pueda adecuarse a este contexto. Por otro lado, la distribución $Beta(2, 2)$ da mayor peso al valor 0.5 para la proporción de admisión y da menor peso a los valores extremos reflejando así, que no debería existir discriminación en ninguna raza.

Mediante las distribuciones previa de cada uno de los parámetros, se calcula la distribución previa de la diferencia de proporciones $\theta = \theta_1 - \theta_2$, para cada una de las dos distribuciones mencionadas anteriormente para θ_i con $i = 1, 2$. Este cálculo hace uso del Resultado 1 y del siguiente código. La gráfica de estas

[3]Puesto que $\hat{\theta} = (S_x + S_y)/(n_x + n_y)$.

Figura 13.1: Dos distribuciones previa para la proporción θ_i con $i = 1, 2$.

Figura 13.2: Distribución previa para la diferencia de proporciones $\theta_1 - \theta_2$.

distribuciones exactas se muestra en la Figura 2 y se realiza mediante el uso de la función `plot.dist`.

```
> previa1<-plot.dist(1,1,1,1, plot=FALSE)
> previa2<-plot.dist(2,2,2,2, plot=FALSE)
```

Nótese que las dos distribuciones previa son simétricas con respecto al valor cero. Sin embargo, nuestra atención estará centrada en la distribución $Beta(2,2)$ como distribución previa para ambas proporciones; por tanto si quisiéramos hallar una estimación puntual o por intervalo previa para la diferencia de proporciones θ, simplemente recurriríamos a las funciones `p.est` y `percentil`, respectivamente.

```
> a1<-2  ;  b1<-2
> a2<-2  ;  b2<-2
> p.est(a1,b1,a2,b2)
[1] -6.589832e-18
> percentil(0.05,a1,b1,a2,b2)
[1] -0.525
> percentil(0.95,a1,b1,a2,b2)
[1] 0.525
```

La estimación puntual previa es muy cercana al valor 0, indicando que la proporción de admisión de una persona de raza blanca debería ser igual a la de una persona de raza negra. El intervalo de credibilidad es simétrico con respecto al valor 0, confirmando la suposición «actual» de que no existe discriminación de ninguna raza; y según esa suposición la probabilidad de que se admiten más empleados de raza blanca que de raza negra debe ser equivalente a la probabilidad de que se admitan más empleados de raza negra que de raza blanca. Haciendo uso de la función `prob` se tiene que $Pr(\theta > 0) = Pr(\theta < 0) \approx 0.5$

```
> prob(0,a1,b1,a2,b2)
[1] 0.503
```

Análisis bayesiano posterior

A continuación se realiza el análisis posterior para θ incorporando la información contenidas en las muestras observadas (Tabla 1). En primer lugar, se especifican los parámetros de las distribuciones posterior de θ_1 y θ_2, las cuales son $Beta(43,41)$ y $Beta(16,32)$, respectivamente.

```
> al1<-a1+x1  ;  be1<-b1+n1-x1
> al2<-a2+x2  ;  be2<-b2+n2-x2
```

A partir de éstas, se calcula la distribución posterior exacta para $\theta = \theta_1 - \theta_2$ usando (**??**). La gráfica de esta distribución posterior, que se realiza usando la función `plot.dist`, se muestra en la Figura 3.

```
> plot.dist(al1,be1,al2,be2, plot=TRUE)
```

Con este simple paso hemos «actualizado» nuestras suposiciones con respecto a θ. La Figura 3 muestra una distribución posterior que no está centrada en cero y por consiguiente pone en tela de juicio la no existencia de discriminación racial. Una estimación puntual exacta del parámetro θ está dada por la media de la distribución. Haciendo uso de la función `p.est` se encuentra que esta estimación corresponde a un número positivo, sugiriendo que la probabilidad de admisión de una persona de raza blanca es superior a la de una persona de raza negra.

```
> p.est(al1,be1,al2,be2)
[1] 0.1785714
```

Figura 13.3: Distribución posterior para la diferencia de proporciones $\theta_1 - \theta_2$.

La suposición de que sí existe evidencia de discriminación racial se verifica al calcular el intervalo de credibilidad al 95 %, usando la función `percentil`, puesto que éste intervalo (0.035,0.32) no contiene al valor cero. Más aún, la probabilidad de que la diferencia de proporciones sea positiva, calculada mediante la función `prob`, resulta ser $Pr(\theta > 0) \approx 0.98$.

```
> percentil(0.05,al1,be1,al2,be2)
[1] 0.035
> percentil(0.95,al1,be1,al2,be2)
[1] 0.32
> prob(0,al1,be1,al2,be2)
[1] 0.9797013
```

Desde otro punto de vista, el valor crítico que θ debe exceder para que exista discriminación es el cero. Dado este valor de corte, es natural comparar las hipótesis

$M_1 : \theta > 0$ y $M_2 : \theta \leq 0$. De esta manera, el factor de Bayes en favor de M_1 se calcula fácilmente, usando (**??**) y la función `prob`, mediante el siguiente código computacional

```
> num <- prob(0,al1,be1,al2,be2)/(1-prob(0,al1,be1,al2,be2))
> den <- prob(0,a1,b1,a2,b2)/(1-prob(0,a1,b1,a2,b2))
> FB <- num/den
> FB
[1] 47.68851
```

(**?**) propuso una escala empírica para clasificar la evidencia a favor de M_1 cuando se utilizan los factores de Bayes. Según esta escala, existe una muy fuerte evidencia de discriminación racial a favor de personas de raza blanca. La Figura 4 muestra la distribución predictiva posterior, utilizando la función `plot.pred`, cuando se consideran muestras de tamaño $n_1 = n_2 = 1$ y $n_1 = n_2 = 3$, respectivamente. En ambos casos la balanza se inclina a favor de la admisión de las personas de raza blanca.

```
> plot.pred(al1,be1,al2,be2,1,1,plot=TRUE)
> plot.pred(al1,be1,al2,be2,3,3,plot=TRUE)
```

Figura 13.4: Distribución predictiva posterior para la diferencia de proporciones.

Los resultados del análisis simulado usando el muestreo de Gibbs son equivalentes a los encontrados con el análisis exacto. Como se observa en la Figura 5, la cadena de Markov converge en la primera iteración y la distribución posterior es equivalente a la distribución encontrada de manera exacta. Este análisis simulado

se realizo mediante la función `plot.gibbs` la cual también devuelve la estimación puntual para θ y el respectivo intervalo de credibilidad cuyos resultados fueron muy similares a los anteriormente mencionados.

```
> plot.gibbs(al1,be1,al2,be2,10000,plot=FALSE,chain=TRUE)
> plot.gibbs(al1,be1,al2,be2,10000,plot=TRUE,chain=FALSE)
```

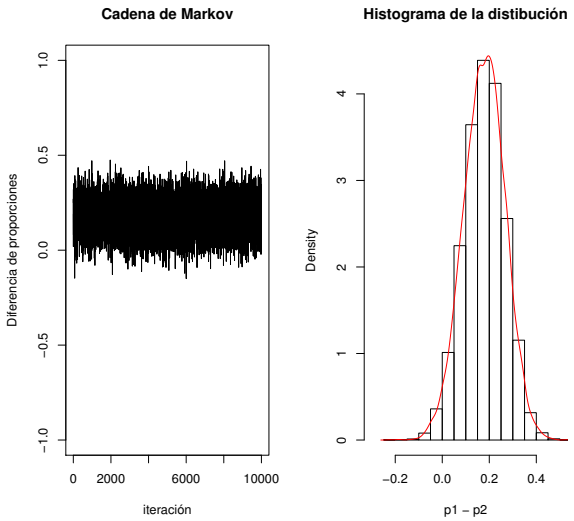

Figura 13.5: Convergencia de la cadena y respectiva distribución posterior usando el muestreo de Gibbs.

13.4 Independencia en tablas de contingencia

La prueba Ji-cuadrado (más conocida como el test de Pearson) usualmente tiene dos connotaciones prácticas importantes en el día a día del oficio del estadístico. Una de ellas es chequear la bondad del ajuste de una distribución propuesta a los datos reales y la otra se trata de probar la independencia de dos variables aleatorias categóricas cuyos conteos se reúnen en una Tabla de contingencia. Sin embargo, esta prueba utiliza resultados de teoría asintótica y por tanto sólo debe ser utilizada cuando el total de conteos marginales es grande (¿qué es grande? algunos autores afirman que es inapropiado utilizar esta prueba cuando los valores esperados por cada celda son menores que diez). Por otra parte, tampoco es apropiado utilizarlo en tablas de contingencia 2X2 puesto que, en este caso particular, la estadística de prueba «asintótica» tendría un solo grado de libertad.

Se cuenta que Fisher pensó en este problema cuando una señorita afirmó (The lady tasting tea) que era capaz de discernir cuándo el té inglés era preparado

adicionando primero la leche, luego el té y revolviendo o viceversa. La solución de Fisher fue la prueba exacta que lleva su nombre (Fishers exact test) la cual guía a la probabilidad exacta, basada en una distribución hipergeométrica, de obtener un arreglo particular en unta Tabla 2X2. Sin embargo, el bayesiano Andrew Gelman afirma:

Yo odio el Fishers exact test puesto que tiene problemas de orden práctico, citando a Agresti y Coull CITAR, y que la presunción de que es ≪exacta≫ se da en circunstancias muy raras. O, para decirlo de otra manera, es una solución exacta a un problema que nunca se verá.

Gelman también propone un método bayesiano directo y sencillo (basado en la diferencia de dos proporciones inducida por la Tabla 2X2) que es práctico en los problemas fáciles y fácilmente se generaliza a problemas más complejos. Curiosamente, el método del análisis de la diferencia de proporciones fue uno de mis primeros acercamientos a la teoría bayesiana y hace poco tiempo, en este blog, publiqué un artículo que trata de cómo realizar un análisis de este tipo en R. En resumidas cuentas, la técnica se trata de suponer que las entradas de la Tabla vienen de distribuciones binomiales ≪independientes≫ (Nótese el símil con la hipótesis nula de independencia), una con parámetro θ_1 y la otra con parámetro θ_2. Estos parámetros, que se asumen variables aleatorias, tienen asignada una función de probabilidad previa, que usualmente es Beta puesto que admite varios casos particulares como la uniforme. Como se asume independencia, entonces la densidad posterior de los parámetros será Beta. En resumen se tienen dos densidades posterior para las cuales se simulan un número grande (digamos dos mil) observaciones para tener dos vectores de tamaño 2000. Por la independencia, un vector de observaciones provenientes de la distribución posterior de $\theta_1 - \theta_2$ está dada por la resta de los anteriores vectores. Las inferencias estarán dadas en términos de este nuevo vector. Eso es todo.

Sin embargo, una cosa es que la diferencia de los parámetros de las binomiales sea igual a cero con una credibilidad muy alta y otra es que las filas y las columnas en la Tabla 2X2 sean independientes. Luego, lo que Gelman no responde es ¿de qué manera se relacionan estas dos situaciones? Pues bien, la respuesta está en este sencillo documento que encuentro muy didáctico. Ahora que está claro que sí existe una relación directa entonces no queda nada más sino bajar las funciones y realizar la inferencia, obviamente bayesiana, en las tablas 2X2 que son tan usadas por este gremio.

13.4.1 Independencia en tablas 2×2 y diferencia de proporciones

Suponga que se tienen dos variables aleatorias X, Y con dos categorías cada una. Las categorías de X son x_1 y x_2, las categorías de Z son y_1 y y_2 Basado en la siguiente Tabla 2×2, las columnas representan las categorías de X, mientras que las filas, las categorías de Y.

En donde, n_{ij} denota el número de individuos que tienen la característica x_j y y_i para $i, j = 1, 2$. Suponga que el objetivo del investigador es averiguar acerca de la existencia de independencia entre las dos variables de estudio. En primer lugar,

	x_1	x_2
y_1	n_{11}	n_{12}
y_2	n_{21}	n_{22}

Tabla 13.2: Tabla de contingencia 2×2

note que, por la definición clásica de independencia, X y Y, son independientes si y sólo si

$$Pr(X = x_i) = Pr(X = x_i \mid Y = y_j) \qquad i, j = 1, 2. \tag{13.4.1}$$

Por otra parte, suponga que el análisis bayesiano de diferencia de proporciones arroja como conclusión que $theta_1$ y $theta_2$ son estadísticamente iguales[4]. Bajo este supuesto, por el teorema de probabilidad total

$$\begin{aligned} Pr(X = x_1) &= Pr(X = x_1 \mid Y = y_1)Pr(Y = y_1) + Pr(X = x_1 \mid Y = y_2)Pr(Y = y_2) \\ &= Pr(X = x_1 \mid Y = y_1)Pr(Y = y_1) + Pr(X = x_1 \mid Y = y_1)Pr(Y = y_2) \\ &= Pr(X = x_1 \mid Y = y_1)[Pr(Y = y_1) + Pr(Y = y_2)] \\ &= Pr(X = x_1 \mid Y = y_1) \end{aligned}$$

Puesto que $Pr(Y = y_1) + Pr(Y = y_2) = 1$. Con esto se concluye que $Pr(X = x_1) = Pr(X = x_1 \mid Y = y_1)$. Análogamente se tiene también que $Pr(X = x_i) = Pr(X = x_i \mid Y = y_j)$ para $i, j = 1, 2$. Con lo anterior se concluye que si $\theta_1 = \theta_2$, entonces existe independencia entre las filas y columnas.

13.5 Una aplicación al mercadeo empresarial

(**?**) afirma que el empaque de un producto juega un papel muy importante en la decisión de compra de los consumidores, pues éste sirve como mecanismo para captar la atención, recordar a los compradores actuales y crea expectativa sobre lo que está adentro, entre otras. Lo anterior implica que un mejor empaque puede ser de significación para un producto, en términos de su posicionamiento en el mercado y/o en el aumento de las ventas del producto. Por esta razón, es indispensable realizar una prueba de empaque antes de lanzar oficialmente un nuevo producto o cambiar la presentación de un producto comercializado en la actualidad. En esta sección, se presenta una aplicación de las funciones creadas en este artículo aplicado a datos reales resultante de una prueba de empaque. La documentación de estas funciones está descrito en la Sección 4.2. y están disponibles en la página de internet dada en la Sección 4.1.

[4] En el contexto de la Tabla de contingencia, $\theta_1 = Pr(X = x_1 \mid Y = y_1)$ y $\theta_2 = Pr(X = x_1 \mid Y = y_2)$. También es útil notar que n_{11} es una realización de la variable S_x, mientras que n_{21} es una realización de la variable S_y.

De esta manera, siguiendo a (**?**), suponga que una empresa desea cambiar el empaque y la forma de presentación de un producto particular que está regularmente posicionado en el mercado. Para evaluar el impacto de la nueva presentación en la intención de compra del producto, el gerente de *marketing* planea una prueba de empaque por medio de la recolección de información en una sesión de grupo (*focus group*). La prueba fue realizada en 124 consumidores, donde a cada uno de ellos se le pregunta sobre la preferencia entre el empaque nuevo y el actual en términos de la intención de compra, y los resultados de la prueba de empaque se muestran en la Tabla 1.

Tabla 13.3: *Tabla de conteos resultante de una prueba de empaque*

Empaque	Compra	No compra	Total
Nuevo	32	31	63
Actual	11	24	35

Mediante el análisis estadístico de estos datos se debe responder a la siguiente pregunta: ¿El cambio de empaque afecta la intención de compra de los consumidores en la categoría?

Análisis frecuentista

El análisis estadístico de los datos necesita de un modelo cuyas características generales se dan a continuación: Suponga que θ_1 es la probabilidad de que se venda un producto con empaque nuevo y que θ_2 es la probabilidad de venta de un producto con empaque actual. Sea $X_i = 1$ si el i-ésimo consumidor encuestado tiene intención de comprar el producto con empaque nuevo y $X_i = 0$, si no tiene intención de comprar el producto con empaque nuevo. De la misma manera, se define $Y_i = 1$ si el i-ésimo consumidor tiene intención de comprar el producto con empaque actual y $Y_i = 0$, en otro caso.

Si asumimos que existe independencia entre y dentro de cada tipo de empaque, y que θ_1 y θ_2 son constantes entre los consumidores, entonces su decisión de compra forma una secuencia de ensayos Bernoulli. Definiendo, $S_x = \sum_{i=1}^{63} X_i$ y $S_y = \sum_{i=1}^{35} Y_i$, se concluye que

$$S_x|\theta_1 \sim Binomial(63, \theta_1) \qquad S_y|\theta_2 \sim Binomial(35, \theta_2) \tag{13.5.1}$$

Bajo este marco de referencia, casi todos los textos básicos de inferencia estadística **?** proponen que la distribución de la diferencia de dos proporciones muestrales, dada por

$$D = \frac{S_x}{n_1} - \frac{S_y}{n_2} \tag{13.5.2}$$

donde n_1 es el tamaño de muestra de los consumidores de empaque nuevo, y n_2 es el tamaño de muestra de los consumidores de empaque actual, es aproximada mediante una distribución normal de media nula y de varianza $V_\theta(D) = (1/n_1 +$

$1/n_2)\theta(1-\theta)$. Se supone que los conteos tienen una distribución binomial con el mismo parámetro $\theta = \theta_1 = \theta_2$. Por tanto, para juzgar la hipótesis $\theta_1 = \theta_2$, se construye una nueva variable aleatoria $U = D/\sqrt{V_{\hat{\theta}}(D)}$ que aproximadamente[5] tiene distribución normal estándar. También es posible utilizar la variable U^2 que aproximadamente tiene una distribución chi-cuadrado con un grado de libertad.

Para responder a la pregunta de interés, el investigador estaría tentado a realizar una prueba de diferencia de dos proporciones y a tomar una decisión con respecto al valor p arrojado por dicha prueba. A continuación se muestran los resultados arrojados por la función `prop.test` propia del ambiente computacional del software `R` **?**.

```
> n1 <- 63  ;  x1 <- 32
> n2 <- 35  ;  x2 <- 11
> prop.test(c(x1,x2),c(n1,n2))

        2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 2.6852, df = 1, p-value =  0.1013
alternative hypothesis: two.sided
```

De esta manera, para un nivel de significación del 5 %, no se rechaza la hipótesis de igualdad de proporciones. En otras palabras, no se encuentra evidencia de que el cambio al empaque nuevo tenga algún efecto sobre la decisión de compra comparado con el empaque actual.

Análisis bayesiano previa

Sin duda alguna, una de las herramientas más poderosas de la inferencia bayesiana es la definición de la distribución previa de los parámetros y como lo afirma (**?**) de la misma manera que no existe un principio general al definir una verosimilitud para una muestra aleatoria, en el caso frecuentista, tampoco existe un principio general para definir una distribución previa, en el caso bayesiano. Por lo anterior, esta etapa del análisis bayesiano debe ser recorrida con mucho cuidado.

Se supone que la distribución previa para la proporción de admisión θ_i es $Beta(a_i, b_i)$ con $i = 1, 2$. En la Figura 1 se observan dos candidatos miembros de la familia de las distribuciones Beta; estos son, $Beta(1, 1)$ y $Beta(2, 2)$. Nótese que la distribución $Beta(1, 1)$ se reduce a la distribución uniforme continua sobre el intervalo $(0, 1)$, la cual es una distribución previa no informativa y parece natural pensar que esta distribución pueda adecuarse a este contexto. Por otro lado, la distribución $Beta(2, 2)$ da mayor peso al valor 0.5 para la probabilidad de venta y da menor peso a los valores extremos reflejando así, que la probabilidad de vender el producto es la misma sin importar el empaque. Lo anterior conduce a una modesta percepción del investigador hacia el nuevo empaque.

[5] Puesto que $\hat{\theta} = (S_x + S_y)/(n_1 + n_2)$.

Figura 13.6: Dos distribuciones previa para la proporción θ_i con $i = 1, 2$.

Figura 13.7: Distribución previa para la diferencia de proporciones $\theta_1 - \theta_2$.

Mediante las distribuciones previa de cada uno de los parámetros, se calcula la distribución previa de la diferencia de proporciones $\theta = \theta_1 - \theta_2$, para cada una de las dos distribuciones mencionadas anteriormente para θ_i con $i = 1, 2$.

Este cálculo hace uso del Resultado 1 y del siguiente código. La gráfica de estas distribuciones exactas se muestra en la Figura 2 y se realiza mediante el uso de la función `plot.dist`:

```
> previa1 <- plot.dist(1,1,1,1, plot=FALSE)
> previa2 <- plot.dist(2,2,2,2, plot=FALSE)
```

Nótese que las dos distribuciones previa son simétricas con respecto al valor cero. Sin embargo, nuestra atención estará centrada en la distribución $Beta(2,2)$ como distribución previa para ambas proporciones; por tanto si quisiéramos hallar una estimación puntual o por intervalo previa para la diferencia de proporciones θ, simplemente recurriríamos a las funciones `p.est` y `percentil`, respectivamente, tal como lo indica el siguiente código:

```
> a1 <- 2  ;  b1 <- 2
> a2 <- 2  ;  b2 <- 2
> p.est(a1,b1,a2,b2)
[1] -6.589832e-18
> percentil(0.05,a1,b1,a2,b2)
[1] -0.525
> percentil(0.95,a1,b1,a2,b2)
[1] 0.525
```

La estimación puntual previa es muy cercana al valor 0, indicando que la probabilidad de venta con el empaque nuevo debería ser igual a la del empaque actual. El intervalo de credibilidad es simétrico con respecto al valor 0, confirmando la suposición ≪actual≫ de que no existe diferencia significativa en los dos tipos de empaques; y según esa suposición la probabilidad de que se venden más productos de empaque nuevo que de empaque actual debe ser equivalente a la probabilidad de que se venden más productos de empaque actual que de empaque nuevo. Haciendo uso de la función `prob` se tiene que $Pr(\theta > 0) = Pr(\theta < 0) \approx 0.5$

```
> prob(0,a1,b1,a2,b2)
[1] 0.503
```

Análisis bayesiano posterior

A continuación se realiza el análisis posterior para θ incorporando la información contenidas en las muestras observadas (Tabla 1). En primer lugar, se especifican los parámetros de las distribuciones posterior de θ_1 y θ_2, las cuales son $Beta(34,33)$ y $Beta(13,26)$, respectivamente.

```
> al1 <- a1+x1  ;  be1 <- b1+n1-x1
> al2 <- a2+x2  ;  be2 <- b2+n2-x2
```

A partir de éstas, se calcula la distribución posterior exacta para $\theta = \theta_1 - \theta_2$ usando (**??**). La gráfica de esta distribución posterior, que se realiza usando la función `plot.dist`, se muestra en la Figura 3.

```
> plot.dist(al1,be1,al2,be2, plot=TRUE)
```

Con este simple paso hemos «actualizado» nuestras suposiciones con respecto a θ. La Figura 3 muestra una distribución posterior que no está centrada en cero y por consiguiente pone en tela de juicio la igualdad entre los tipos de empaque. Una estimación puntual exacta del parámetro θ está dada por la media de la distribución. Haciendo uso de la función `p.est` se encuentra que esta estimación corresponde a un número positivo, sugiriendo que la probabilidad de venta de un producto con el empaque nuevo es superior a la de un producto con el empaque actual.

```
> p.est(al1,be1,al2,be2)
[1] 0.1741294
```

Figura 13.8: Distribución posterior para la diferencia de proporciones $\theta_1 - \theta_2$.

La suposición de que sí existe evidencia de que el nuevo empaque afecta de manera positiva la intención de compra del comsumidor se verifica al calcular el intervalo de credibilidad al 95 %, usando la función `percentil`, puesto que este intervalo (0.015,0.33) no contiene al valor cero. Más aún, la probabilidad de que la diferencia de proporciones sea positiva, calculada mediante la función `prob`, resulta ser $Pr(\theta > 0) \approx 0.964$.

```
> percentil(0.05,al1,be1,al2,be2)
[1] 0.015
> percentil(0.95,al1,be1,al2,be2)
[1] 0.33
```

```
> prob(0,al1,be1,al2,be2)
[1] 0.964069
```

Desde otro punto de vista, el valor crítico que θ debe exceder para que exista diferencia entre los dos tipos de empaque es el valor cero. Dado este valor de corte, es natural comparar las hipótesis $M_1 : \theta > 0$ y $M_2 : \theta \leq 0$. De esta manera, el factor de Bayes en favor de M_1 se calcula fácilmente, usando (11) y la función `prob`, mediante el siguiente código computacional

```
> num <- prob(0,al1,be1,al2,be2)/(1-prob(0,al1,be1,al2,be2))
> den <- prob(0,a1,b1,a2,b2)/(1-prob(0,a1,b1,a2,b2))
> FB <- num/den
> FB
[1] 26.511
```

(**?**) propuso una escala empírica para clasificar la evidencia a favor de M_1 cuando se utilizan los factores de Bayes. Según esta escala, existe una fuerte evidencia de que los efectos de los tipos de empaque sobre la decisión de compra son diferentes a favor del empaque nuevo. La Figura 4 muestra la distribución predictiva posterior, utilizando la función `plot.pred`, cuando se consideran muestras futuras de tamaño $n_1 = n_2 = 5$, $n_1 = n_2 = 10$ y $n_1 = n_2 = 15$, respectivamente. En estos tres casos la balanza se inclina a favor de la venta de los productos con el empaque nuevo.

```
> plot.pred(al1,be1,al2,be2,1,1,plot=TRUE)
> plot.pred(al1,be1,al2,be2,3,3,plot=TRUE)
```

Los resultados del análisis simulado usando el muestreo de Gibbs son equivalentes a los encontrados con el análisis exacto. Como se observa en la Figura 5, la cadena de Markov converge en la primera iteración y la distribución posterior es equivalente a la distribución encontrada de manera exacta. Este análisis simulado se realizo mediante la función `plot.gibbs` la cual también devuelve la estimación puntual para θ y el respectivo intervalo de credibilidad cuyos resultados fueron muy similares a los anteriormente mencionados.

```
> plot.gibbs(al1,be1,al2,be2,10000,plot=FALSE,chain=TRUE)
> plot.gibbs(al1,be1,al2,be2,10000,plot=TRUE,chain=FALSE)
```

13.5.1 Validación del nuevo empaque usando independencia

El problema de evaluar el impacto del cambio de empaque sobre la venta del producto también puede ser resuelto considerando la información recolectada como datos categóricos pertenecientes a una Tabla de contingencia 2×2, donde se tienen las categorías en filas y columnas y éstas conducen a la definición de dos variables aleatorias discretas, C y F. Las realizaciones de C se denotan como c_1 y c_2 y las de F como f_1 y f_2 (ver Tabla 2). Bajo esta perspectiva, el investigador desea

Figura 13.9: Distribución predictiva posterior para la diferencia de proporciones en muestras de tamaño 5, 10 y 15.

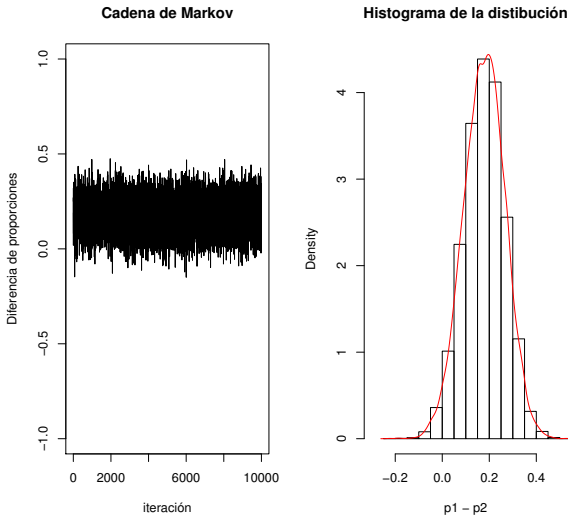

Figura 13.10: Convergencia de la cadena y respectiva distribución posterior usando el muestreo de Gibbs.

saber si las filas son independientes de las columnas. Si esto sucede, en el caso de la prueba de producto, es posible concluir que el cambio de empaque no tiene un

efecto significativo en la intención de compra del producto.

Tabla 13.4: *Tabla de contingencia* 2×2

	c_1	c_2	Total
f_1	s_x	$n_1 - s_x$	n_1
f_2	s_y	$n_2 - s_y$	n_2

Existen varios métodos estadísticos utilizados para verificar la independencia entre filas y columnas; dos de los más conocidos, en el enfoque frecuentista, son la prueba Ji-cuadrado y la prueba exacta de Fisher. La prueba Ji-cuadrado (*Pearson's Test*) utiliza resultados de teoría asintótica y por tanto sólo debe ser utilizada cuando los totales marginales, n_1 y n_2, son grandes. Por otra parte, tampoco es apropiado utilizarlo en tablas de contingencia 2×2 puesto que, en este caso particular, la estadística de prueba «asintótica» tendría un solo grado de libertad. Por otro lado, Fisher propuso una solución a este inconveniente (*Fisher's Exact Test*) la cual guía a la probabilidad «exacta», basada en una distribución hipergeométrica, de obtener un arreglo particular en unta Tabla 2×2. Sin embargo, esta solución frecuentista tiene problemas de orden práctico **?**.

Retomando el caso de la prueba de empaque, usamos los comandos `chisq.test` y `fisher.test` en `R` para llevar a cabo estos dos procedimientos de verificación de independencia entre filas y columnas. Es posible observar que en ambos procedimientos, el valor p es mayor que el nivel de significación usual del 5 %, indicando que hay independencia entre las filas y las columnas. De esta forma, los dos métodos en el enfoque frecuentista coinciden en que la decisión de compra no está influenciada por el tipo de empaque; esto es, el cambio de empaque no tiene efecto sobre la intención de compra de los consumidores.

```
> Datos <- matrix(c(x1,n1-x1,x2,n2-x2),2,2)
> chisq.test(Datos)

        Pearson's Chi-squared test with Yates' continuity correction

data:  Datos
X-squared = 2.6852, df = 1, p-value = 0.1013

> fisher.test(Datos)

        Fisher's Exact Test for Count Data

data:  Datos
p-value = 0.08914
```

Una alternativa bayesiana para analizar la independencia entre filas y columnas en una Tabla de contingencia 2×2 es analizar la diferencia de las dos proporciones utilizando el método descrito en las secciones anteriores. En primer lugar, note

que, por la definición clásica de independencia, C y F son independientes si y sólo si

$$Pr(F = f_i) = Pr(F = f_i \mid C = c_j) \qquad i, j = 1, 2. \tag{13.5.3}$$

Por otra parte, suponga que el análisis bayesiano de diferencia de proporciones arroja como conclusión que θ_1 y θ_2 son estadísticamente iguales[6]. Bajo este supuesto, por el teorema de probabilidad total, se tiene que

$$\begin{aligned} Pr(F = f_1) &= Pr(F = f_1 \mid C = c_1)Pr(C = c_1) + Pr(F = f_1 \mid C = c_2)Pr(C = c_2) \\ &= Pr(F = f_1 \mid C = c_1)Pr(C = c_1) + Pr(F = f_1 \mid C = c_1)Pr(C = c_2) \\ &= Pr(F = f_1 \mid C = c_1)[Pr(C = c_1) + Pr(C = c_2)] \\ &= Pr(F = f_1 \mid C = c_1) \end{aligned}$$

Puesto que $Pr(C = c_1) + Pr(C = c_2) = 1$. Con esto se concluye que $Pr(F = f_1) = Pr(F = f_1 \mid C = c_1)$. Análogamente se tiene también que $Pr(F = f_i) = Pr(F = f_i \mid C = c_j)$ para $i, j = 1, 2$. Con lo anterior se concluye que si $\theta_1 = \theta_2$, entonces existe independencia entre las filas y columnas.

En el caso de la prueba de empaque, el análisis bayesiano condujo a la conclusión de que $\theta_1 > \theta_2$ puesto que:

1. Siendo $\theta = \theta_1 - \theta_2$, se tiene que $Pr(\theta > 0) \approx 0.964$.
2. El factor de Bayes a favor del modelo $M_1 : \theta > 0$ es 26.511, indicando que los datos muestran una fuerte evidencia a favor de que el nuevo empaque afecta de manera positiva la intención de compra.

Por lo anterior, las filas y las columnas de la Tabla de contingencia no se consideran independientes; esto es, el cambio de empaque sí influye significativamente en la decisión de compra de los consumidores. En este caso, el empaque nuevo promueve favorablemente la venta del producto comparado con el empaque actual y la recomendación gerencial debería estar enfocada en el lanzamiento del producto con el nuevo empaque.

13.6 Conclusión

A pesar de los avances teóricos y computacionales en la estadística bayesiana en las últimas décadas, poca atención se ha prestado a uno de los problemas más comunes en la investigación estadística, el análisis de la diferencia de proporciones. Lo anterior implica que, para este problema específico, la distribución exacta posterior, encontrada por (**?**), sea muy difícil de implementar en la práctica, dada su forma compleja.

[6] En el contexto de la Tabla de contingencia, $\theta_1 = Pr(F = f_1 \mid C = c_1)$ y $\theta_2 = Pr(F = f_1 \mid C = c_2)$. También es útil notar que s_x es una realización de la variable S_x, mientras que s_y es una realización de la variable S_y.

En este artículo se plantea la solución computacional a este problema, mediante la creación de una serie de funciones, enmarcadas en el software estadístico R, que permiten realizar un análisis bayesiano exhaustivo: desde la definición de la distribución previa para el parámetro de interés hasta la realización de pruebas de hipótesis bayesianas. Este paquete de funciones no sólo se remite a los cálculos numéricos exactos para las distribuciones sino que también permite representar gráficamente funciones predictivas que reflejan las suposiciones «actuales» acerca de la diferencia de proporciones.

Como una aplicación empírica, se proponen dos soluciones, de tipo bayesiano y frecuentista, a un problema empresarial referente al cambio del empaque de un producto en una categoría de mercado. Como resultado de esta práctica se concluye que las técnicas estadísticas clásicas frecuentistas guían a una conclusión errónea acerca del juzgamiento de la hipótesis de interés. Sin embargo, al utilizar las técnicas bayesianas, aparte de llegar a las conclusiones correctas, es posible obtener información adicional acerca del comportamiento de los parámetros que el enfoque clásico no brinda.

14 Tópicos avanzados en modelos bayesianos

14.1 Modelación de media y varianza conjunta en el modelo lineal general

En términos de modelamiento estadístico, la relación entre una variable de interés Y y un conjunto de variables auxiliares $\mathbf{X}$, es una de las herramientas estadísticas más utilizadas por los investigadores. La regresión simple, la regresión múltiple y el análisis de varianza, entre otros, forman parte de las opciones que este tipo de relaciones de asociación pueden establecer. Como lo menciona (**?**), es muy útil adoptar la notación matricial para el desarrollo posterior del análisis bayesiano; entonces, se definen

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{y} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nq} \end{pmatrix}$$

y se supone que existe una relación de asociación entre $\mathbf{X}$ y $\mathbf{y}$ que puede ser descrita mediante el siguiente modelo probabilístico

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{14.1.1}$$

en donde $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ es un vector aleatorio tal que cada una de sus componentes sigue una distribución de probabilidad, que en la mayoría de los casos suele ser normal de media cero y de varianza constante. Sin embargo, tal como lo afirman (**?**), cuando la varianza no es la misma en cada componente de $\boldsymbol{\varepsilon}$, no siempre es posible recurrir a las técnicas clásicas de estabilización de varianza y se hace conveniente considerar un análisis con un modelamiento propio para la varianza del modelo. De esta manera, a nivel individual, el modelo que se considera es el siguiente.

$$Y_i = \mu_i + \varepsilon_i \qquad i = 1, \dots, n \tag{14.1.2}$$

Donde $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$, ε_i sigue una distribución normal con media cero y varianza σ_i^2. Además, $Cov(\varepsilon_i, \varepsilon_j) = 0$ para todo $i \neq j$. Para analizar la heteroscedasticidad del modelo, se propone una transformación de σ_i^2 mediante una función conveniente[1] $g(\cdot)$ tal que

$$g(\sigma_i^2) = \mathbf{z}_i'\boldsymbol{\gamma} \tag{14.1.3}$$

[1] La función $g(\cdot)$ debe ser monótona, diferenciable y asegurar la positividad de las varianzas. Nótese que una función conveniente con las anteriores características puede ser $g(\cdot) = \log(\cdot)$.

Donde $\mathbf{z}_i$ es un vector de variables auxiliares para el i-ésimo individuo cuyos componentes son todas o algunas variables en $\mathbf{x}_i'$, y $\boldsymbol{\gamma}$ es un vector de parámetros. En (**?**) se aclara que es posible realizar inferencia frecuentista en términos de estimación conjunta para el vector de parámetros de interés $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ utilizando métodos numéricos y el algoritmo de Score (**??**). Sin embargo, dadas las condiciones del modelo y la gran cantidad de parámetros que se deben estimar, se esperaría que el enfoque Bayesiano tenga un mejor comportamiento en términos de eficiencia computacional dadas las demostradas propiedades óptimas de los métodos de Monte Carlo vía Cadenas de Markov (MCMC) (**???**).

En términos de inferencia bayesiana, se debe proponer una distribución previa $p(\boldsymbol{\beta}, \boldsymbol{\gamma})$ para el vector de parámetros de interés; es usual asignar una distribución normal multivariante dada por

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} \sim Normal \left[\boldsymbol{\theta}_0 = \begin{pmatrix} \mathbf{b} \\ \mathbf{g} \end{pmatrix}, \boldsymbol{\Sigma}_0 = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0}' & \mathbf{G} \end{pmatrix} \right] \tag{14.1.4}$$

Lo anterior implica que se considera independencia previa entre los parámetros del modelo de media $\boldsymbol{\beta}$ y los parámetros del modelo de dispersión $\boldsymbol{\gamma}$. Lo anterior tiene sentido, más aún cuando (**?**) ha demostrado que la matriz de información de Fisher es diagonal por bloques; luego, haciendo uso de los resultados de la teoría asintótica, se tiene la independencia de los parámetros en cuestión. Por otro lado, la verosimilitud de los datos está dada por

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\} \tag{14.1.5}$$

Donde, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ depende de $\boldsymbol{\beta}$ pues $\mu_i = \mathbf{x}_i' \boldsymbol{\beta}$ y $\boldsymbol{\Sigma} = diag(\sigma_i)$ depende de $\boldsymbol{\gamma}$ puesto que $\sigma_i = g^{-1}(\mathbf{z}_i' \boldsymbol{\gamma})$. Luego, haciendo uso del teorema de Bayes, la distribución posterior está dada por

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}) \right\} \\ \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right\} \end{aligned} \tag{14.1.6}$$

Es claro que la anterior distribución de probabilidad multivariante no tiene una forma conocida y no es sencillo simular vectores de observaciones provenientes de esta. Sin embargo, desde la anterior expresión, es posible conocer las distribuciones condicionales posterior tanto de $\boldsymbol{\beta}$ como de $\boldsymbol{\gamma}$. Luego, se tiene que para $\boldsymbol{\beta}$ la distribución condicional posterior es

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{y}) &\propto \exp\left\{ -\frac{1}{2} \left[(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}) + (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b}) \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{b}^*)' \mathbf{B}^{*-1} (\boldsymbol{\beta} - \mathbf{b}^*) \right\} \end{aligned} \tag{14.1.7}$$

Fácilmente, factorizando convenientemente, se encuentra una expresión idéntica a la función de densidad de un vector aleatorio con distribución multivariante $Normal(\mathbf{b}^*, \mathbf{B}_q^*)$. En donde,

$$\mathbf{B}^* = \left(\mathbf{B}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}$$
$$\mathbf{b}^* = \mathbf{B}^*\left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\right)$$

De otra parte, la distribución posterior condicional para el vector de parámetros de interés $\boldsymbol{\gamma}$ toma la siguiente forma

$$p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\left[(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \mathbf{g})'\mathbf{G}^{-1}(\boldsymbol{\gamma} - \mathbf{g})\right]\right\} \tag{14.1.8}$$

Algoritmo de estimación

Dado el rápido avance de los métodos computacionales y su aplicación en la inferencia estadística, es posible utilizar algunos procedimientos enmarcados en los métodos MCMC. En particular sería plausible pensar en que dado que las distribuciones condicionales posterior son conocidas, se podría usar el algoritmo de Gibbs (**??**) para dos cadenas multivariantes cuya distribución estacionaria este dada por (16) y por (17), respectivamente. Este procedimiento funciona bien para crear la cadena desde (16); sin embargo, es posible mostrar que (17) no es un kernel log-concavo y por lo tanto la utilización de Gibbs no es conveniente pues la cadena presentaría serios problemas de convergencia.

Para palear estos inconvenientes, es posible utilizar el algoritmo de Metropolis-Hastings (**??**), que es útil justamente en casos en donde obtener observaciones directamente de la distribución posterior conjunta (15) es difícil. Este método requiere de la asignación, por parte del investigador, de una distribución de probabilidad (llamada por (**?**, p. 289) distribución de salto o *jumping distribution*) de donde se simulan valores en cada iteración y luego se decide si ese valor particular puede tomarse como un valor posiblemente generado de la distribución posterior. Dependiendo de la escogencia de la distribución de salto, el algoritmo converge rápido o lento; al respecto, (**?**) han mostrado teóricamente que una tasa de aceptación aceptable es del orden del 50 % y que ésta depende también de la semejanza de la distribución posterior con la distribución de salto.

Sobre la base expuesta, y teniendo en cuenta que la distribución posterior (15) no tiene una forma cerrada, es indispensable la buena escogencia de la distribución de salto. Al respecto, (**?**) ha diseñado, bajo el contexto de la estimación bayesiana en modelos lineales generalizados mixtos, un algoritmo que permite escoger una distribución de salto que guía no sólo a una alta eficiencia computacional sino a tasas de aceptación muy altas. Este algoritmo fue modificado por (**?**) y ajustado en el contexto del modelamiento bayesiano conjunto de media y varianza. En esa investigación se encontró que, cuando la función de vinculo $g(\cdot) = \ln(\cdot)$, una distribución de salto óptima estaba dada en términos de pseudo-variables dadas

por

$$\tilde{y}_i = \mathbf{z}_i'\boldsymbol{\gamma} + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1 \qquad i = 1, \ldots, n.$$

Con esta construcción, (**?**) demostraron que, siendo $\boldsymbol{\beta}^{(c)}$ y $\boldsymbol{\gamma}^{(c)}$ los valores actuales de los parámetros de interés, la distribución de salto con kernel gaussiano $q_{\boldsymbol{\gamma}}$ se obtiene como

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(c)}) = Normal(\mathbf{g}^*, \mathbf{G}^*) \tag{14.1.9}$$

En donde,

$$\mathbf{G}^* = \left(\mathbf{G}^{-1} + \frac{1}{2}\mathbf{Z}'\mathbf{Z}\right)^{-1}$$
$$\mathbf{g}^* = \mathbf{G}^*\left(\mathbf{G}^{-1}\mathbf{g} + \frac{1}{2}\mathbf{Z}'\tilde{\mathbf{y}}\right)$$

Con $\mathbf{Z}$ la matriz que contiene las observaciones de las variables auxiliares para el modelo de varianzas y $\tilde{\mathbf{y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)'$. (**?**) utilizaron el anterior enfoque para crear un algoritmo de Metropolis-Hastings modificado (puesto que la distribución de salto sólo es utilizada para proponer valores de $\boldsymbol{\gamma}$), el cual permite la simulación de valores provenientes de las distribución posterior (15) mediante la escogencia de una distribución de salto apropiada para el modelamiento en las varianzas. A continuación, se dará paso al algoritmo que permite la realización de la debida inferencia posterior para el vector de parámetros $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

1. Iniciar el contador de iteraciones en $j = 1$.
2. Fijar valores iniciales para la cadena dados por $(\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\gamma}^{(j-1)})$.
3. Actualizar el vector $\boldsymbol{\beta}$ a un nuevo valor $\boldsymbol{\beta}^{(j)}$, generado desde (16).
4. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\gamma}}$, generado desde la distribución de salto (18).
5. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\gamma}}$, de otra manera $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$
6. Actualizar el contador de la cadena de j a $j + 1$.
7. Volver al paso 3 y repetir el procedimiento hasta que la cadena alcance la convergencia deseada.

La convergencia de la cadena resultante, cuya distribución estacionaria está dada por la posterior conjunta (15), puede ser constatada de distintas maneras; por ejemplo, con ayuda de métodos tradicionales de series de tiempo o mediante la implementación de varias cadenas paralelas (**?**). Por otro lado, cuando esta metodología se ha aplicado a conjuntos de datos reales y simulados, se nota que el periodo de transición de la cadena es muy corto y la convergencia se alcanza casi que inmediatamente.

14.1.1 Modelos no lineales generalizados

Consideramos el modelo de regresión no lineal

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + e_i, \qquad i = 1, \ldots, n \tag{14.1.10}$$

donde y_i es la variable respuesta, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})'$ es un vector de información auxiliar asociado con la variable respuesta por medio de un vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$. $f(\cdot, \cdot)$ es una función no lineal diferenciable para $\boldsymbol{\beta}$, y los errores e_i se asumen independientes con media cero, pero no necesariamente idénticamente distribuidos. Al mismo tiempo, se asume que la distribución de y_i se encuentra dentro de la familia exponencial con verosimilitud dada por la siguiente expresión general

$$p(y_i \mid \theta_i) = a(y_i) \exp\{d(\theta_i)T(y_i) + c(\theta_i)\}$$

donde θ_i es el parámetro de interés, y las formas funcionales de $a(\cdot)$, $d(\cdot)$, $T(\cdot)$ y $c(\cdot)$ se suponen conocidas. De esta manera, la media de y_i está dada por

$$E(y_i) = \mu_i = h^{-1}(f(\mathbf{x}_i, \boldsymbol{\beta})) \tag{14.1.11}$$

donde $h(\cdot)$ es una función de vínculo definida similarmente como en (**?**). Bajo estos supuestos, se dice que y_i sigue un modelo no lineal generalizado. Note que si $f(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, entonces y_i sigue un modelo lineal generalizado. Sin embargo, en situaciones prácticas puede existir un componente de sobredispersión y variación extra que no se tiene en cuenta en la formulación anterior. En esta directriz, (**?**) propone un enfoque clásico en el análisis de regresión normal y más tarde (**?**) propone el uso de variables de trabajo para la estimación bayesiana de los parámetros tal como en (**?**). Recientemente, (**?**) han reportado avances en el desarrollo de una metodología bayesiana para el modelamiento de parámetros en la familia exponencial biparamétrica con linealidad en la media. Ese supuesto de linealidad fue relajado por (**?**) en un artículo donde estudia modelos normales no lineales heterocedásticos. La contraparte bayesiana de estos enfoques pueden ser vistos como casos particulares de los resultados de este trabajo doctoral puesto que los algoritmos resultantes son implícitos en la metodología MCMC propuestos en este trabajo, considerando densidades no lineales y no normales con efectos aleatorios.

14.1.2 Modelos mixtos

Modelo mixto lineal

La formulación del modelo mixto tiene sus orígenes en el artículo de (**?**) que, bajo el contexto del análisis inferencial para modelos de datos longitudinales, propuso la siguiente expresión para la modelación de la variable respuesta:

$$\mathbf{y}_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{W}_i'\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i \tag{14.1.12}$$

donde $\mathbf{y}_i$ es un vector de respuestas de tamaño n_i para la i-ésima unidad experimental ($i = 1, \ldots, n$), $\mathbf{X}_i'$ es una matriz de diseño de tamaño $n_i \times p$, la cual caracteriza la parte sistemática (efectos fijos) de la respuesta que depende de las covariables y del tiempo. $\boldsymbol{\beta}$ es un vector de parámetros de los efectos fijos de tamaño p, $\mathbf{W}_i$ es otra matriz de diseño que caracteriza la parte estocástica (efectos aleatorios) en la respuesta, que se puede atribuir a las fuentes de variación dentro de las unidades experimentales. $\boldsymbol{\lambda}_i$ corresponde a un vector de parámetros de efectos aleatorios de tamaño R (p y R no necesariamente iguales). Por último, $\boldsymbol{\varepsilon}_i$ es un vector de errores de tamaño n_i, el cual caracteriza la variación propia de la forma de medición de la unidad i-ésima.

(**?**) afirma que los supuestos generales acerca del modelo clásico anterior se dan a continuación:

- $\boldsymbol{\varepsilon}_i \sim Normal(\mathbf{0}, \boldsymbol{\Sigma}_i)$, donde $\boldsymbol{\Sigma}_i$ es una matriz de covarianzas de tamaño $n_i \times n_i$ que caracteriza la variación y la correlación dentro de las unidades. Esta variación incluye el error de medición en la respuesta y la posible correlación inducida por la naturaleza serial de la recolección de los datos. La escogencia más simple y común para $\mathbf{W}_i$ es el modelo que afirma que la varianza es la misma en todos los puntos del tiempo y que las mediciones están suficientemente apartadas en el tiempo; de tal manera que no hay ninguna correlación inducida por la recolección de los datos. El anterior caso se escribe como $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$
- $\boldsymbol{\lambda}_i \sim Normal(\mathbf{0}, \mathbf{L})$, donde $\mathbf{L}$ es una matriz de covarianzas que caracteriza la variación entre individuos. Es posible que $\mathbf{L}$ tenga una forma particular o que no esté estructurada. También, es posible tener diferentes matrices $\mathbf{L}$ para diferentes grupos.
- $\boldsymbol{\varepsilon}_i$ es independiente de $\boldsymbol{\lambda}_i$ para todo $i = 1, \ldots, n_i$.

Como consecuencia de los anteriores supuestos, se tiene que, el modelo queda completamente especificado como se escribe a continuación:

$$\begin{aligned} E(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{W}_i'\boldsymbol{\lambda}_i \\ Var(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \sigma^2 \boldsymbol{\Sigma}_i \end{aligned}$$

La escritura del modelo lineal mixto toma la siguiente forma general

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} + \boldsymbol{\varepsilon} \tag{14.1.13}$$

En donde la esperanza y varianza del vector de variables repuesta está dada por

$$\begin{aligned} E(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} \\ Var(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \boldsymbol{\Sigma} \end{aligned}$$

Nótese que, bajo diferentes contextos, el modelo se acomoda de acuerdo a la estructura que tomen las matrices $\mathbf{X}$, $\mathbf{W}$, $\boldsymbol{\Sigma}$ y los vectores $\mathbf{y}$ y $\boldsymbol{\varepsilon}$. En particular, el modelo de (**?**) toma su lugar cuando se escriben los anteriores componentes de

la siguiente manera:

$$\begin{aligned}
\mathbf{X}' &= (\mathbf{X}_1' \ldots, \mathbf{X}_n')' \\
\mathbf{W} &= \bigoplus_{i=1}^{n} \mathbf{W}_i = diag(\mathbf{W}_1 \ldots, \mathbf{W}_n) \\
\boldsymbol{\Sigma} &= diag(\boldsymbol{\Sigma}_1 \ldots, \boldsymbol{\Sigma}_n) \\
\mathbf{y} &= (\mathbf{y}_1' \ldots, \mathbf{y}_n')' \\
\boldsymbol{\varepsilon} &= (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_n')
\end{aligned}$$

Modelo mixto no lineal

La definición general del modelo no lineal mixto está dada por la relación del vector de variables respuesta $\mathbf{y} = (y_1, \ldots, y_n)'$ con un vector de medias $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ de la siguiente manera

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{14.1.14}$$

donde, $\boldsymbol{\varepsilon}$ es un vector de errores y las componentes de $\boldsymbol{\mu}$ son funciones no lineales de los parámetros fijos y aleatorios ligados con variables de información auxiliar en $\mathbf{X}$ y en $\mathbf{W}$, definidas en la sección anterior. De esta forma, se define la componente media del modelo para el i-ésimo individuo como

$$\mu_i = f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) \tag{14.1.15}$$

donde f es una función no lineal y diferenciable. Nótese que la esperanza y la varianza del vector de vriables respuesta es En donde la esperanza y varianza del vector de variables repuesta está dada por

$$\begin{aligned}
E(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \boldsymbol{\mu} \\
Var(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \boldsymbol{\Sigma}
\end{aligned}$$

De la misma manera, se asume que los componentes del vector de errores tienen distribución normal con media nula y varianza constante. Además, se tiene que el vector de parámetros de los efectos aleatorios tiene distribución normal y es independiente del vector de errores.

Modelo mixto lineal generalizado

Existen varios autores que han trabajado este modelo general; sin embargo, el lector puede referirse a (**?**), (**?**), (**?**) y (**?**). Un puñado de las muy variadas aplicaciones se pueden encontrar en (**?**), (**?**), (**?**), (**?**), (**?**) y (**?**). Cuando la variable respuesta pertenece a la familia exponencial bi-paramétrica, una forma general, aunque no única, de un modelo lineal generalizado mixto está dado por las siguientes expre-

siones

$$p(y_i \mid \theta_i, \tau_i) = a(y) \exp\{d_1(\theta_i, \tau_i)T_1(y_i) + d_2(\theta_i, \tau_i)T_2(y_i) + b(\theta_i, \tau_i)\} \quad (14.1.16)$$

$$E(y_i \mid \theta_i, \tau_i) = \mu_i \quad (14.1.17)$$

$$h(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} \quad (14.1.18)$$

Bajo esta formulación, se asume que que el espacio del parámetro natural contiene un rectángulo bidimensional que, por propiedades de traslación, contiene el punto $\tau = 0$. Luego, la familia exponencial uniparamétrica surge cuando $\tau = 0$ y la verosimilitud de las observaciones toma la siguiente forma

$$p(y_i \mid \theta_i) = a(y_i) \exp\{d(\theta_i)T(y_i) + c(\theta_i)\}$$

De manera tradicional, la función de vínculo h se considera diferenciable y está relacionada con la modelación de la media de la variable aleatoria relacionada con parámetros fijo y aleatorios.

14.1.3 Modelos jerárquicos y multinivel

En términos de notación, se establece que las observaciones individuales, llamadas *casos* se observan dentro de conglomerados o grupos, llamados *unidades*. Luego, los casos son para las unidades como los estudiantes a las escuelas o los hogares a las ciudades. De esta manera, siempre y cuando se tenga acceso a información auxiliar en el nivel de las unidades y también de los casos, es posible plantear un modelo que explique el fenómeno de estudio en los casos y al mismo tiempo tener en cuenta un submodelo que explique la presencia de los casos dentro de las unidades.

Existe muy variada literatura sobre el tratamiento frecuentista y bayesiano de los modelos multinivel. Nosotros referimos al lector a (**?**), (**?**), (**?**) , quienes dan una excelente introducción basada en modelos lineales. (**?**) y (**?**), proveen una discusión acerca de la modelación no lineal bajo el contexto de aplicaciones bioestadísticas. En el caso de modelos lineales generalizados (**?**), (**?**), (**?**) y (**?**) son un buen punto de partida en términos de inferencia estadística y aplicaciones prácticas cuando existe una estructura multinivel.

Relación entre un modelo multinivel y un modelo mixto

En este apartado, se considera un modelo jerárquico con dos niveles. En primer lugar, se considera que existen n individuos agrupados (no necesariamente mediante una jerarquía directa) dentro de J unidades; en cada grupo existen n_j casos, tal que $\sum_{j=1}^{J} n_j = n$, y se supone que para las J unidades, es posible modelar el vector de variables respuesta como

$$\mathbf{y}_j = \mathbf{X}_j'\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \quad (14.1.19)$$

donde cada matriz $\mathbf{X}_j$ tiene dimensiones $n_j \times P$ y $\boldsymbol{\varepsilon}_j$ se distribuye normal con media $\mathbf{0}$ y matriz de covarianzas $\boldsymbol{\Sigma}_j$. En el siguiente nivel, se quiere modelar la variación del vector $\boldsymbol{\beta}_j$ de una unidad a otra. En otras palabras, se quiere un modelo entre unidades que vincula a los casos del modelo anterior. De esta manera, suponiendo que existe información auxiliar $\mathbf{u}_{jp}$ para el p-ésimo parámetro $(p = 1, \ldots, P)$ perteneciente a la j-ésima unidad $(p = 1, \ldots, P)$, entonces se tiene el siguiente modelo

$$\beta_{jk} = \mathbf{u}'_{jk}\boldsymbol{\alpha}_k + \delta_{jk} \tag{14.1.20}$$

donde δ_{ij} tiene distribución normal con media cera y varianza σ_p^2. Nótese que, en forma más general, el anterior modelo puede escribirse de la siguiente manera, llegando a un modelo entre unidades

$$\boldsymbol{\beta}_j = \mathbf{U}'_j\boldsymbol{\alpha} + \boldsymbol{\delta}_j \tag{14.1.21}$$

Donde $\boldsymbol{\delta}_j$ tiene distribución normal con media $\mathbf{0}$ y varianza $\mathbf{L}_j$, además se tiene que

$$\begin{aligned}
\mathbf{U}_j &= \begin{pmatrix} \mathbf{u}'_{j1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{u}'_{j2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{u}'_{jP} \end{pmatrix} \\
&= \bigoplus_{p=1}^{P} \mathbf{u}'_{jp} \\
\boldsymbol{\alpha} &= (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_P)' \\
\boldsymbol{\delta}_j &= (\delta_{j1}, \ldots, \delta_{jP})'
\end{aligned}$$

Por lo tanto, se tiene la siguiente expresión

$$\mathbf{y}_j = \mathbf{X}'_j\mathbf{U}'_j\boldsymbol{\alpha} + \mathbf{X}'_j\boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j \tag{14.1.22}$$

La anterior formulación indica que un modelo multinivel es un caso particular de un modelo mixto puesto que puede escribirse también como

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} + \boldsymbol{\varepsilon} \tag{14.1.23}$$

En donde

$$\begin{aligned}
\mathbf{y} &= (\mathbf{y}_1, \ldots, \mathbf{y}_J)' \\
\mathbf{X} &= (\mathbf{X}'_1\mathbf{U}'_1, \ldots, \mathbf{X}'_J\mathbf{U}'_J)' \\
\mathbf{W} &= \bigoplus_{j=1}^{J} \mathbf{X}'_j \\
\boldsymbol{\lambda} &= (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_J)' \\
\boldsymbol{\varepsilon} &= (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_J)'
\end{aligned}$$

Luego,

$$\boldsymbol{\Sigma} = \bigoplus_{j=1}^{J} \boldsymbol{\Sigma}_j$$
$$\mathbf{L} = \bigoplus_{j=1}^{J} \mathbf{L}_j$$

Modelos multinivel generalizados

Los modelos multinivel pueden ser generalizados de la misma forma que (**?**) generalizaron el modelo lineal. Lo anterior permite que el investigador pueda modelar la variable respuesta como binomial, poisson, logístico, entre otras al mismo tiempo que se utilizan los efectos aleatorios para modelar el anidadamiento de las observaciones o sobredispersión. (**?**) introducen algunas técnicas bayesianas para encontrar la densidad posterior de los parámetros de interés mediante la técnica de Gibbs. El anterior acercamiento generaliza el trabajo de (**?**), (**?**), (**?**).

Cuando la variable respuesta pertenece a la familia exponencial bi-paramétrica, un modelo lineal generalizado multinivel se puede escribir de la siguiente manera

$$\begin{aligned} p(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= a(y_{ij}) \exp\left\{d_1(\theta_{ij}, \tau_{ij}) T_1(y_{ij}) + d_2(\theta_{ij}, \tau_{ij}) T_2(y_{ij}) + b(\theta_{ij}, \tau_{ij})\right\} \\ E(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= \mu_{ij} \\ h(\mu_{ij}) &= \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_j \end{aligned}$$

Nótese que el vector $\boldsymbol{\eta}$ en la j-ésima unidad está dado por

$$\boldsymbol{\eta}_j = \mathbf{X}'_j \boldsymbol{\beta}_j$$

En el siguiente nivel, se quiere modelar la variación del vector $\boldsymbol{\beta}_j$ de una unidad a otra. Luego, suponiendo que existe información auxiliar $\mathbf{U}_j$ para la j-ésima unidad, entonces se tiene que

$$\boldsymbol{\beta}_j = \mathbf{U}'_j \boldsymbol{\alpha} + \boldsymbol{\delta}_j$$

Donde $\boldsymbol{\delta}_j$ tiene distribución normal con media $\mathbf{0}$ y varianza $\mathbf{L}_j$. Al combinar los modelos en los dos niveles, se tiene la siguiente expresión

$$\boldsymbol{\eta}_j = \mathbf{X}'_j \mathbf{U}'_j \boldsymbol{\alpha} + \mathbf{X}'_j \boldsymbol{\delta}_j$$

La anterior formulación indica que un modelo multinivel generalizado es un caso particular de un modelo mixto generalizado puesto que el modelo global puede escribirse también como

$$\boldsymbol{\eta} = \mathbf{X}' \boldsymbol{\beta} + \mathbf{W}' \boldsymbol{\lambda}$$

En donde las anteriores cantidades se definen análogamente al anterior apartado.

14.1.4 Estimación en áreas pequeñas

La referencia principal de la inferencia frecuentista y bayesiana en problemas de áreas pequeñas es el trabajo reciente de (**?**). Otras referencias de importancia en el tema son (**?**), (**?**), (**?**), (**?**), (**?**), (**?**) y (**?**).

Estimadores directos y sintéticos

Bajo la teoría de inferencia en poblaciones finitas, se considera un parámetro de interés a una función conocida de los valores que toma la variable respuesta en cada uno de los elementos de una población $U = \{1, \ldots, k, \ldots, N\}$. Siendo y_k el valor de la variable respuesta en el individuo i, algunos ejemplos de parámetros de interés son el total poblacional dado por

$$Y = \sum_{k \in U} y_k$$

el promedio poblacional dado por

$$\bar{Y} = \frac{Y}{N}$$

entre otros. Suponga que la población finita está particionada en J subgrupos poblacionales llamados dominios U_j tales que

$$U = \bigcup_{j=1}^{J} U_j \qquad U_j \bigcap U_{j'} = \emptyset$$

Es de interés para el investigador obtener estimaciones de los parámetros de interés (totales, promedios, razones) en cada uno de los dominios. La teoría de muestreo provee estimadores de estas cantidades con el supuesto de que al menos un elemento de cada uno de los dominios fue seleccionado en una muestra seleccionada s. El lector puede referirse a los textos de (**?**), (**?**) y (**?**) para una introducción al tema. Además de lo anterior, se denota por $\hat{Y}_j$ el estimador del total en el dominio U_j dado por Y_j y se supone que este estimador es insesgado con respecto al diseño de muestreo p utilizado en la selección de s; en otras palabras $E_p(\hat{Y}_j) = Y_j$ para todo $j = 1, \ldots, J$. Cuando el tamaño de muestra n_j en el dominio U_j es muy pequeño, los estimadores directos pueden arrojar resultados incoherentes influenciados por una gran variación.

Una solución al anterior inconveniente está dada por los estimadores sintéticos que son usados para derivar estimativos de los parámetros de interés bajo el supuesto de que estos dominios se comportan de igual forma que la población finita. De esta manera, se asume que existe un estimador directo $\hat{Y}$ para el total poblacional Y que puede ser calculado mediante los datos obtenidos en la muestra s y la estimación tendrá un error estándar pequeño puesto que emplea toda la información recolectada en el estudio. Al mismo tiempo, se asume que existe información

auxiliar de los totales en los dominios para alguna característica de información auxiliar x. Luego $X_j = \sum_{k \in U_j} x_k$ se asume conocida para todo U_j. Un estimador sintético para Y_j está dado por

$$\hat{Y}_j = \frac{X_j}{X} \hat{Y}$$

donde X es el total poblacional de la característica de la información auxiliar. Nótese que $\sum_{j=1}^{J} \hat{Y}_j = \hat{Y}$ y además que incluso si la muestra seleccionada s no contiene elementos de un dominio determinado, siempre es posible calcular una estimación para tal dominio. En el caso particular en que $X_j = N_j$, el tamaño del dominio U_j, este estimador asumiría que la media global $\bar{Y}$ es equivalente a las medias $\bar{Y}_j$. Es claro que el anterior supuesto es muy fuerte y por tanto este tipo de estimadores pueden ser ampliamente sesgados para algunos subgrupos de interés.

Modelos tipo A

Los modelos tipo A hacen referencia a modelos especificados para la información de un dominio per se; es decir, totales o medias en cada dominio. En primer lugar, se asume que $\theta_j = v(\bar{Y}_j)$ hace referencia al parámetro poblacional de interés en el dominio U_j; por ejemplo, si el parámetro de interés es el promedio, entones v será la función idéntica. Si el parámetro de interés es el total, entonces $v(\bar{Y}_j) = N_j \bar{Y}_j$. Asuma que θ_j está relacionado con un vector de variables auxiliares $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})'$ para el dominio U_j a través de un modelo lineal

$$\theta_j = \mathbf{x}_j' \boldsymbol{\beta} + u_j \delta_j \tag{14.1.24}$$

donde los u_j son constantes positivas conocidas, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ es un vector de coeficientes de regresión, δ_j son efectos aleatorios especificados para cada subgrupo poblacional U_j que se asumen independientes e idénticamente distribuidos con esperanza $E_m(\delta_j) = 0$ y varianza $Var_m(\delta_j) = \sigma_\delta^2 \geq 0$. El sufijo m denota que la esperanza o la varianza se toma al respecto del modelo. El parámetro σ_δ^2 es una medida de homogeneidad de los dominios una vez se ha contemplado el efecto de las covariables en $\mathbf{x}_j$. (**?**) afirma que en algunos estudios no es posible seleccionar a todos los dominios en la muestra seleccionada, sin embargo, se asume que el modelo supuesto obedece a un modelo poblacional para todo $j = 1, \ldots, J$ y que los dominios seleccionados en la muestra siguen esa estructura. En algunas palabras, el sesgo de selección de los dominios es nulo.

Para realizar inferencias acerca del parámetro de interés θ_j, es necesario asumir que existe un estimador directo $\hat{\theta}_j$ que se comporta de la siguiente manera:

$$\hat{\theta}_j = \theta_j + e_j \tag{14.1.25}$$

donde e_j representan los errores de muestreo, distribuidos normalmente con $E_p(e_j|\theta_j) = 0$ y $Var_p(e_j|\theta_j) = \sigma_j^2$. Luego, se tiene que

$$\hat{\theta}_j = \mathbf{x}_j' \boldsymbol{\beta} + u_j \delta_j + e_j \tag{14.1.26}$$

Nótese que si asumimos independencia entre e_j y δ_j, la anterior formulación puede ser vista como un caso especial de un modelo lineal mixto, puesto que

$$\hat{\boldsymbol{\theta}} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} + \mathbf{e} \tag{14.1.27}$$

con $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$, $\mathbf{W} = diag(u_1, \ldots, u_J)$, $\boldsymbol{\lambda} = (\delta_1, \ldots, \delta_J)'$ y $\mathbf{e} = (e_1, \ldots, e_J)'$. Si además suponemos que σ_j^2 puede ser modelada de la siguiente manera

$$g(\sigma_j^2) = \mathbf{z}_j'\boldsymbol{\gamma}$$

entonces el modelo se convierte en un caso particular de la propuesta para modelos lineales mixtos tal como se expuso en la sección 4.1. Nótese que bajo el contexto de estimación en áreas pequeñas es muy justificable asignar una variación distinta en cada dominio. De esta forma, $\mathbf{z}_j$ debe corresponder a un vector de variables auxiliares relacionadas directamente con el dominio U_j. De esta manera, un estimador natural para estos parámetros nace de la esperanza condicional de la variable respuesta, que en este caso particular son los estimadores directos en cada dominio, dado los parámetros del modelo como se indica a continuación.

$$\begin{aligned} \hat{\theta}_j^* &= E(\hat{\theta}_j | \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\lambda^2) \\ &= \mathbf{x}_j'\boldsymbol{\beta} + u_j\delta_j \end{aligned} \tag{14.1.28}$$

Modelos tipo B

A diferencia de los modelos tipo A, este tipo de modelos tienen en cuenta la información individual de los elementos pertenecientes a la población U y por lo tanto la información auxiliar está directamente relacionada con estos individuos, mas no con el dominio como ocurre en los modelos tipo A. De esta forma, para el individuo i-ésimo perteneciente al j-ésimo dominio ($i = 1, \ldots, N_j$), ($j = 1, \ldots, J$) se supone conocido un vector de información auxiliar $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijP})'$ el cual se relaciona con la variable respuesta y_{ij} a través de un modelo de regresión dado por

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \delta_j + e_{ij} \tag{14.1.29}$$

con δ_j denotando el efecto aleatorio del dominio U_j cuya distribución es $Normal(0, \sigma_\delta^2)$ y e_{ij} son variables aleatorias independientes y normalmente distribuidas con media $E_m(e_{ij}) = 0$ y $Var_m(e_{ij}) = \sigma_{ij}^2$. Asumiendo que el diseño de muestreo garantiza que para cada dominio U_j de tamaño N_j exista una submuestra s_j de tamaño n_j y que los valores muestrales obedezcan al modelo anterior. Nótese entonces que para la población finita éste puede ser reescrito como

$$\mathbf{y}_j = \mathbf{X}_j'\boldsymbol{\beta} + \delta_j\mathbf{1}_j + \mathbf{e}_j \tag{14.1.30}$$

$$= \begin{pmatrix} \mathbf{y}_j^s \\ \mathbf{y}_j^r \end{pmatrix} = \begin{pmatrix} \mathbf{X}_j^s \\ \mathbf{X}_j^r \end{pmatrix}\boldsymbol{\beta} + \delta_j \begin{pmatrix} \mathbf{1}_j^s \\ \mathbf{1}_j^r \end{pmatrix} + \begin{pmatrix} \mathbf{e}_j^s \\ \mathbf{e}_j^r \end{pmatrix} \tag{14.1.31}$$

donde el superíndice s denota las unidades que fueron seleccionadas en la muestra y r denota las unidades que no fueron seleccionadas en la muestra. Cabe

aclarar que el modelo anterior es válido siempre y cuando el diseño de muestreo se pueda considerar no informativo y por lo tanto ignorable (**??**). Por ejemplo, el modelo considerado no es apropiado si el diseño de muestreo utilizado es por conglomerados o en varias etapas puesto que el efecto de los conglomerados no está incorporado. Por otra parte, nótese que si el parámetro de interés es el total Y_j, entonces

$$Y_j = \sum_{k \in U_j} y_k = \sum_{k \in s_j} y_k + \sum_{k \in r_j} y_k$$

donde s_j denota el conjunto de elementos de U_j que fueron seleccionados en la muestra y $r_j = U_j - s_j$. Por lo tanto, el problema de estimar el total Y_j se convierte en un problema de predicción de y_k para todo $k \in r_j$. De la misma manera, si el parámetro de interés es la media $\bar{Y}_j$, entonces

$$\bar{Y}_j = f_j \bar{y}_{s_j} + (1 - f_j)\bar{y}_{r_j}$$

En este caso, la estimación del promedio de la variable respuesta en el dominio U_j es equivalente a la predicción de $\bar{y}_{r_j}$ dados los datos de la muestra $\mathbf{y}_j^s$ y el vector de variables auxiliares $\mathbf{X}_j = (\mathbf{X}_j^s, \mathbf{X}_j^r)'$. Nótese que este modelo es un caso particular de un modelo lineal mixto de la forma $\mathbf{y}_j = \mathbf{x}_j'\boldsymbol{\beta} + \mathbf{W}_j'\boldsymbol{\lambda}_j + \mathbf{e}_j$.

Modelos en la familia exponencial bi-paramétrica

Siguiendo el enfoque de (**?**), es posible proponer modelos lineales generalizados con efectos aleatorios concernientes a los dominios de la población U condicionado a μ_{ij}, τ_{ij}, los parámetros de la familia exponencial bi-paramétrica, la variable respuesta y_{ij} se asume independientes con función de densidad de probabilidad perteneciendo a la familia exponencial de la siguiente manera

$$\begin{aligned} p(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= a(y_{ij}) \exp\{d_1(\theta_{ij}, \tau_{ij})T_1(y_{ij}) + d_2(\theta_{ij}, \tau_{ij})T_2(y_{ij}) + b(\theta_{ij}, \tau_{ij})\} \\ E(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= \mu_{ij} \\ h(\mu_{ij}) &= \eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_j + \delta_j \end{aligned}$$

La anterior formulación incluye distribuciones bien conocidas como la binomial, normal, Poisson, Gamma, entre otras. Así mismo, es posible modelar el parámetro τ_{ij} mediante información auxiliar $\mathbf{z}_{ij}$. Por ejemplo, si $\mu_{ij} = logit(\pi_{ij})$, entones π_{ij} puede denotar una proporción asociada a una variable binaria en el i-ésimo barrio de la j-ésima ciudad.

14.1.5 Modelos mixtos

El modelo lineal descrito en la sección anterior tiene sólo un componente de efecto aleatorio dado por $\boldsymbol{\varepsilon}$. Los parámetros del modelos son los componentes del vector de coeficientes de regresión $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ y, si el modelo es homoscedástico, la varianza del error σ^2. El modelo mixto se puede ver como una extensión del modelo lineal general que incluye términos adicionales en los componentes aleatorios y son

apropiados en muchos casos en donde es explícita la naturaleza jerárquica de los datos, o cuando las respuestas presentan algún tipo de correlación, o cuando se registran varias mediciones de los mismos individuos a través del tiempo. Según (**?**), los modelos mixtos son la base de modelos más complejo como los de intercepto aleatorio, pendiente aleatoria, efectos anidados aleatorios, modelos aditivos, entre otros.

En este apartado se revisará sin profundizar varias acepciones a los modelos mixtos. En particular, a) modelos mixtos lineales, partiendo del modelo de (**?**), pero generalizando su escritura al caso más general, b) modelos mixtos no lineales y c) modelos mixtos generalizados.

Es abrumadora la cantidad de literatura existente acerca de las diferentes variantes de un modelo mixto (lineal, no lineal o generalizado) y las distintas técnicas de estimación y predicción. Para no perder de vista el objetivo de este trabajo doctoral, no se ha querido analizar a profundidad ni en las métodos ni en las aplicaciones del modelo mixto, sino que más bien, el lector es referido a algunas importantes libros o artículos en cada modelo.

Es importante aseverar que el desarrollo de esta propuesta, así como su justificación están basadas en la escritura del modelo mixto en su forma más general. Con base en lo anterior, esta pequeña revisión no está supeditada a ninguna notación particular. Por otro lado, la razón de incluir estas tres acepciones del modelo mixto está dada por la misma fundamentación de las propuestas metodológicas de inferencia bayesiana que en este trabajo se realizan a modelos mixtos lineales, modelos mixtos no lineales y modelos mixtos generalizados.

Modelo mixto lineal

La formulación del modelo mixto tiene sus orígenes en el famoso artículo de (**?**) que, bajo el contexto del análisis inferencial para modelos de datos longitudinales, propuso la siguiente expresión para la modelación de la variable respuesta:

$$\mathbf{y}_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{W}_i'\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i \tag{14.1.32}$$

donde $\mathbf{y}_i$ es un vector de respuestas de tamaño n_i para la i-ésima unidad experimental ($i = 1, \ldots, n$), $\mathbf{X}_i'$ es una matriz de diseño de tamaño $n_i \times p$, la cual caracteriza la parte sistemática (efectos fijos) de la respuesta que depende de las covariables y del tiempo. $\boldsymbol{\beta}$ es un vector de parámetros de los efectos fijos de tamaño p, $\mathbf{W}_i$ es otra matriz de diseño que caracteriza la parte estocástica (efectos aleatorios) en la respuesta, que se puede atribuir a las fuentes de variación dentro de las unidades experimentales. $\boldsymbol{\lambda}_i$ corresponde a un vector de parámetros de efectos aleatorios de tamaño R (p y R no necesariamente iguales). Por último, $\boldsymbol{\varepsilon}_i$ es un vector de errores de tamaño n_i, el cual caracteriza la variación propia de la forma de medición de la unidad i-ésima.

(**?**) afirma que los supuestos generales acerca del modelo clásico anterior se dan a continuación:

- $\boldsymbol{\varepsilon}_i \sim Normal(\mathbf{0}, \boldsymbol{\Sigma}_i)$, donde $\boldsymbol{\Sigma}_i$ es una matriz de covarianzas de tamaño

$n_i \times n_i$ que caracteriza la variación y la correlación dentro de las unidades. Esta variación incluye el error de medición en la respuesta y la posible correlación inducida por la naturaleza serial de la recolección de los datos. La escogencia más simple y común para $\mathbf{W}_i$ es el modelo que afirma que la varianza es la misma en todos los puntos del tiempo y que las mediciones están suficientemente apartadas en el tiempo; de tal manera que no hay ninguna correlación inducida por la recolección de los datos. El anterior caso se escribe como $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$

- $\boldsymbol{\lambda}_i \sim Normal(\mathbf{0}, \mathbf{L})$, donde $\mathbf{L}$ es una matriz de covarianzas que caracteriza la variación entre individuos. Es posible que $\mathbf{L}$ tenga una forma particular o que no esté estructurada. También, es posible tener diferentes matrices $\mathbf{L}$ para diferentes grupos.
- $\boldsymbol{\varepsilon}_i$ es independiente de $\boldsymbol{\lambda}_i$ para todo $i = 1, \ldots, n_i$.

Como consecuencia de los anteriores supuestos, se tiene que, el modelo queda completamente especificado como se escribe a continuación:

$$\begin{aligned} E(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{W}_i'\boldsymbol{\lambda}_i \\ Var(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \sigma^2 \boldsymbol{\Sigma}_i \end{aligned}$$

Nota: Como se destaca en la introducción de esta sección, dado que un modelo mixto cubre muchos otros modelos estadísticos, como los modelos de intercepto aleatorio, intercepto y pendiente aleatoria, efectos aleatorios cruzados, efectos aleatorios anidados, suavizamiento generalizado, modelos aditivos, modelos aditivos uniparamétricos, entre otros, se debe notar que (19) no es la única acepción del modelo lineal mixto y que el contexto del análisis de datos longitudinales y medidas repetidas no es el único en donde se utiliza este tipo de modelos. Por lo tanto, dada la generalidad de este trabajo de investigación, de aquí en adelante, a menos que se diga lo contrario, la escritura del modelo lineal mixto toma la siguiente forma general

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} + \boldsymbol{\varepsilon} \tag{14.1.33}$$

En donde la esperanza y varianza del vector de variables repuesta está dada por

$$\begin{aligned} E(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} \\ Var(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \boldsymbol{\Sigma} \end{aligned}$$

Nótese que, bajo diferentes contextos, el modelo se acomoda de acuerdo a la estructura que tomen las matrices $\mathbf{X}$, $\mathbf{W}$, $\boldsymbol{\Sigma}$ y los vectores $\mathbf{y}$ y $\boldsymbol{\varepsilon}$. En particular, el modelo de (**?**) toma su lugar cuando se escriben los anteriores componentes de

la siguiente manera:

$$\begin{aligned}
\mathbf{X}' &= (\mathbf{X}_1' \ldots, \mathbf{X}_n')' \\
\mathbf{W} &= \bigoplus_{i=1}^{n} \mathbf{W}_i = diag(\mathbf{W}_1 \ldots, \mathbf{W}_n) \\
\boldsymbol{\Sigma} &= diag(\boldsymbol{\Sigma}_1 \ldots, \boldsymbol{\Sigma}_n) \\
\mathbf{y} &= (\mathbf{y}_1' \ldots, \mathbf{y}_n')' \\
\boldsymbol{\varepsilon} &= (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_n')
\end{aligned}$$

Con base en lo anteriormente expuesto, se nota que es posible utilizar diferentes enfoque estadísticos para realizar la estimación de los parámetros de interés, la predicción de los efectos fijos y la estimación de los componentes de varianza entre y dentro individuos. Por ejemplo, (**?**) utilizan el enfoque de máxima-verosimilitud y la descomposición ortogonal de las matrices de covarianzas del modelo para optimizar el problema de optimización de la función de verosimilitud y reduce la dimensión de la optimización. Otros aspectos importantes en la estimación frecuentista de los componentes de varianza pueden ser encontrados en (**?**, section 4.5).

Modelo mixto no lineal

En las últimas décadas, los modelos mixtos no lineales han sido bastante estudiados y se han desarrollado metodologías de estimación, tanto clásicas como no paramétricas y bayesianas, en contextos como datos longitudinales y medidas repetidas (**???**). Asimismo, las técnicas de inferencia utilizadas han sido distintas y varios enfoques de estimación se han desarrollado (**????**) en distintas áreas del saber (**??**). Como lo afirma (**?**), los modelos no lineales de efectos mixtos se han convertido en un marco de referencia estándar en muchas áreas de aplicación.

De esta manera, la definición general del modelo no lineal mixto está dada por la relación del vector de variables respuesta $\mathbf{y} = (y_1, \ldots, y_n)'$ con un vector de medias $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ de la siguiente manera

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{14.1.34}$$

donde, $\boldsymbol{\varepsilon}$ es un vector de errores y las componentes de $\boldsymbol{\mu}$ son funciones no lineales de los parámetros fijos y aleatorios ligados con variables de información auxiliar en $\mathbf{X}$ y en $\mathbf{W}$, definidas en la sección anterior. De esta forma, se define la componente media del modelo para el i-ésimo individuo como

$$\mu_i = f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) \tag{14.1.35}$$

donde f es una función no lineal y diferenciable. Nótese que la esperanza y la varianza del vector de vriables respuesta es En donde la esperanza y varianza del vector de variables repuesta está dada por

$$\begin{aligned}
E(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \boldsymbol{\mu} \\
Var(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}) &= \boldsymbol{\Sigma}
\end{aligned}$$

De la misma manera, se asume que los componentes del vector de errores tienen distribución normal con media nula y varianza constante. Además, se tiene que el vector de parámetros de los efectos aleatorios tiene distribución normal y es independiente del vector de errores.

Modelo mixto generalizado

Tal como lo afirma (**?**), los modelos mixtos son una herramienta que permite lidiar con un gran rango de complicaciones en análisis de regresión. Cuando son tratados desde un punto de vista general, este tipo de modelos permite la inferencia de otros modelos avanzados; por ejemplo, modelos tipo *spline*, modelos de suavizamiento, modelos aditivos, e incluso modelos espaciales como el *kriging*. Existen varios autores que han trabajado este modelo general; sin embargo, el lector puede referirse a (**?**), (**?**), (**?**) y (**?**). Un puñado de las muy variadas aplicaciones se pueden encontrar en (**?**), (**?**), (**?**), (**?**), (**?**) y (**?**).

Cuando la variable respuesta pertenece a la familia exponencial bi-paramétrica, una forma general, aunque no única, de un modelo lineal generalizado mixto está dado por las siguientes expresiones

$$p(y_i \mid \theta_i, \tau_i) = a(y) \exp\{d_1(\theta_i, \tau_i)T_1(y_i) + d_2(\theta_i, \tau_i)T_2(y_i) + b(\theta_i, \tau_i)\} \quad (14.1.36)$$

$$E(y_i \mid \theta_i, \tau_i) = \mu_i \quad (14.1.37)$$

$$h(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} \quad (14.1.38)$$

Donde, si y es una variable continua, entonces p se asume como una densidad con respecto a la medida de Lebesgue; mientras que, si y es una variable discrete, p se asume como una densidad con respecto a la medida de conteo. También se supone que p es medida-integrable. Bajo esta formulación, se asume que que el espacio del parámetro natural contiene un rectángulo bidimensional que, por propiedades de traslación, contiene el punto $\tau = 0$. Luego, la familia exponencial uniparamétrica surge cuando $\tau = 0$ y la verosimilitud de las observaciones toma la siguiente forma

$$p(y_i \mid \theta_i) = a(y_i) \exp\{d(\theta_i)T(y_i) + c(\theta_i)\}$$

De manera tradicional, la función de vínculo h se considera diferenciable y está relacionada con la modelación de la media de la variable aleatoria relacionada con parámetros fijo y aleatorios. Una vez más, se deben realizar los supuesto de que los componentes del vector de errores tienen distribución normal con media nula y varianza constante y que el vector de parámetros de interés de los efectos aleatorios tiene distribución normal y es independiente del vector de errores.

Tratamiento bayesiano para los efectos aleatorios y efectos fijos

Al respecto de la definición estadística de los efectos aleatorios y los efectos fijos, el autor de esta propuesta doctoral suscribe completamente lo afirmado por (**?**, p. 391), quienes afirman que los términos ≪efectos fijos≫ y ≪efectos aleatorios≫ vienen de la tradición frecuentista clásica (no Bayesiana) y pueden generar cierta

confusión dado que bajo el contexto Bayesiano, todos los parámetros se consideran aleatorios. Luego, los no bayesianos consideran que los efectos fijos son cantidades fijas pero desconocidas, aunque sus métodos de estimación arrojan resultados similares a los encontrados con un análisis bayesiano bajo una distribución previa no informativa.

Es común que el tratamiento bayesiano para los componentes de $\boldsymbol{\beta}$ - llamados efectos fijos en la literatura estadística clásica - esté enfocado en asignar una distribución previa con varianza muy grande, sin importar en donde estén centrados. Sin embargo, respetando la definición del modelo mixto, los componentes de $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r, \ldots, \lambda_R)'$ - efectos aleatorios - deben ser modelados con media cero y varianza común. De esta manera, y a menos que se indique lo contrario, el tratamiento previa de estos vectores estará dado por las siguientes expresiones

$$\begin{aligned} \boldsymbol{\beta} &\sim Normal(\mathbf{b}, \mathbf{B}) \\ \lambda_r \mid \sigma_r^2 &\sim Normal(0, \sigma_r^2) \qquad r = 1, \ldots, R \end{aligned}$$

En donde $\mathbf{B}$ es una matriz de varianzas diagonal, con sus entradas tendiendo a infinito y σ_r^2 es el componente de variación previa de los efectos aleatorios. La mayoría de las veces, este parámetro es desconocido, así que se debe utilizar un acercamiento jerárquico en la definición de las distribuciones previa.

14.1.6 Propuesta para modelos lineales mixtos

(**?**) afirma que este tipo de modelos surgen de la necesidad que se genera para conocer, no sólo estimaciones de los efectos en los modelos, sino también de modelar la varianza que estos generan en el modelo. De esta forma, un modelo lineal puede contemplar tanto efectos fijos como aleatorios que inducen variabilidad en la respuesta; en este caso se habla de un modelo lineal mixto. De esta manera, el modelo lineal mixto tiene la siguiente formulación:

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} + \boldsymbol{\varepsilon} \tag{14.1.39}$$

En este caso $\boldsymbol{\beta}$ es un vector P-dimensional de coeficientes de regresión, $\boldsymbol{\lambda}$ es un vector de efectos aleatorios tal que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r, \ldots, \lambda_R)'$. Asociados con estos parámetros, se encuentran las covariables en $\mathbf{X}$ y en $\mathbf{W}$ que explican la variación de los parámetros fijos y aleatorios, respectivamente. Este modelo general se complementa con la asignación de las distribuciones previa para los parámetros de interés. Así,

$$\begin{aligned} \boldsymbol{\beta} &\sim Normal(\mathbf{b}, \mathbf{B}) \\ \lambda_r \mid \sigma_r^2 &\sim Normal(0, \sigma_r^2) \\ \sigma_r^2 &\sim Inversa - Gama(a, h) \end{aligned}$$

Para todo $r = 1, \ldots, R$. Por último, cuando existe heteroscedasticidad en la respuesta, es necesario modelar la varianza de las observaciones, de los contrario como

lo afirma (**?**), la ignorancia del cambio en la varianza para cada individuo genera estimaciones y predicciones sesgadas. Utilizando el enfoque propuesto en (**?**), se propone una transformación de $Var(\varepsilon_i) = \sigma_i^2$ mediante una función conveniente tal que

$$g(\sigma_i^2) = \mathbf{z}_i'\boldsymbol{\gamma} \tag{14.1.40}$$

En modelos lineales mixtos, es justificable la escogencia de $g(\cdot) = \log(\cdot)$ (**????**). Por otro lado, véase que $\mathbf{z}_i$ es un vector conteniendo las covariables pertinentes para la modelación de la varianza en el i-ésimo individuo. Luego, también es necesario asignar una distribución a $\boldsymbol{\gamma}$; de esta manera se tiene que

$$\boldsymbol{\gamma} \sim Normal(\mathbf{g}, \mathbf{G})$$

Nótese que en la asignación de las distribuciones previa se ha seguido el enfoque de (**?**, p. 99) en donde la varianza de los parámetros de efectos aleatorios y la varianza de la variable respuesta se consideran independientes. De esta manera $p(\sigma_i^2, \sigma_r^2) = p(\sigma_i^2)p(\sigma_r^2)$ para todo $i = 1, \ldots, n$ y $r = 1 \ldots, R$; de la misma forma, los parámetros de efectos fijos, aleatorios y los que modelan la varianza se consideran independientes (aunque la extensión es inmediata cuando se deba considerar dependencia).

Con la expuesta anteriormente, la verosimilitud de las observaciones queda definida por

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ \frac{-1}{2}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta} - \mathbf{W}'\boldsymbol{\lambda})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta} - \mathbf{W}'\boldsymbol{\lambda}) \right\} \tag{14.1.41}$$

En donde $\boldsymbol{\Sigma} = diag(\sigma_i^2)$ y $\sigma_i^2 = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$. Bajo las anteriores formulaciones, se tiene el siguiente resultado que establece la distribución condicional de algunos de los parámetros de interés

Resultado 14.1.1. *Las distribuciones condicionales posterior de los parámetros de interés $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ y σ_r^2 están dadas por*

$$\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y} \sim Normal(\mathbf{b}^*, \mathbf{B}^*) \tag{14.1.42}$$

$$\boldsymbol{\lambda} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_r^2, \mathbf{y} \sim Normal(\mathbf{l}^*, \mathbf{L}^*) \tag{14.1.43}$$

$$\sigma_r^2 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y} \sim Inversa - Gama(-1/2 - a, \lambda_r^2/2 + h) \tag{14.1.44}$$

respectivamente; haciendo $\mathbf{y}_b = \mathbf{y} - \mathbf{W}'\boldsymbol{\lambda}$, se tiene que

$$\mathbf{B}^* = \left(\mathbf{B}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}$$
$$\mathbf{b}^* = \mathbf{B}^*\left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}_b\right)$$

Por otro lado, definiendo $\mathbf{y}_l = \mathbf{y} - \mathbf{X}'\boldsymbol{\beta}$ y $\mathbf{L} = diag(\sigma_r^2)$, se tiene que

$$\mathbf{L}^* = \left(\mathbf{L}^{-1} + \mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{W}\right)^{-1}$$
$$\mathbf{l}^* = \mathbf{L}^*\left(\mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{y}_l\right)$$

Prueba. Al suponer independencia previa entre los parámetros de interés, se tiene que la distribución previa conjunta está dada por

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2) = p(\boldsymbol{\beta})p(\boldsymbol{\gamma}) \prod_{r=1}^{R} p(\lambda_r \mid \sigma_r^2)p(\sigma_r^2) \tag{14.1.45}$$

De esta manera, siguiendo el teorema de bayes, la distribución posterior conjunta de los parámetros de interés toma la siguiente forma

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \times p(\boldsymbol{\beta})p(\boldsymbol{\gamma}) \prod_{r=1}^{R} p(\lambda_r \mid \sigma_r^2)p(\sigma_r^2) \tag{14.1.46}$$

Recurriendo al condicionamiento sucesivo, y llevando todos los términos que no dependen de $\boldsymbol{\beta}$ a la constante de proporcionalidad, la distribución condicional posterior para $\boldsymbol{\beta}$, dados valores específicos para $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$ y σ_r^2 $(r = 1, \ldots, R)$, está dada por

$$\begin{aligned}
p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y}) &\propto p(\boldsymbol{\beta}, \underbrace{\boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2}_{fijos} \mid \mathbf{y}) \\
&\propto p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \times p(\boldsymbol{\beta}) \\
&\propto \exp\left\{ \frac{-1}{2} (\mathbf{y}_b - \mathbf{X}'\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_b - \mathbf{X}'\boldsymbol{\beta}) + (\boldsymbol{\beta} - \mathbf{b})' \mathbf{B}^{-1} (\boldsymbol{\beta} - \mathbf{b}) \right\} \\
&\propto \exp\left\{ \frac{-1}{2} (\boldsymbol{\beta} - \mathbf{b}^*)' \mathbf{B}^{*-1} (\boldsymbol{\beta} - \mathbf{b}^*) \right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de densidad de una variable aleatoria con distribución $Normal(\mathbf{b}^*, \mathbf{B}^*)$.

Análogamente, para el tratamiento de los parámetros de efectos aleatorios, la distribución condicional posterior para $\boldsymbol{\lambda}$, dados valores específicos para $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ y σ_r^2, está dada por

$$\begin{aligned}
p(\boldsymbol{\lambda} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_r^2, \mathbf{y}) &\propto p(\boldsymbol{\lambda}, \underbrace{\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_r^2}_{fijos} \mid \mathbf{y}) \\
&\propto p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \times \prod_{r=1}^{R} p(\lambda_r \mid \sigma_r^2) \\
&\propto \exp\left\{ \frac{-1}{2} (\mathbf{y}_l - \mathbf{W}'\boldsymbol{\lambda})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_l - \mathbf{W}'\boldsymbol{\lambda}) + \sum_{r=1}^{R} \frac{\lambda_r^2}{\sigma_r^2} \right\} \\
&= \exp\left\{ \frac{-1}{2} (\mathbf{y}_l - \mathbf{W}'\boldsymbol{\lambda})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_l - \mathbf{W}'\boldsymbol{\lambda}) + \boldsymbol{\lambda}' \mathbf{L}^{-1} \boldsymbol{\lambda} \right\} \\
&\propto \exp\left\{ \frac{-1}{2} (\boldsymbol{\lambda} - \mathbf{l}^*)' \mathbf{L}^{*-1} (\boldsymbol{\lambda} - \mathbf{l}^*) \right\}
\end{aligned}$$

Factorizando convenientemente, se encuentra una expresión idéntica a la función de densidad de una variable aleatoria con distribución $Normal(\mathbf{l}^*, \mathbf{L}^*)$. Por último, para la distribución posterior condicional para cada σ_r^2 $(r = 1, \ldots, R)$ se encuentra utilizando el condicionamiento sucesivo

$$\begin{aligned} p(\sigma_r^2 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y}) &\propto p(\sigma_r^2 \underbrace{\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}}_{fijos} \mid \mathbf{y}) \\ &\propto p(\lambda_r \mid \sigma_r^2) \times p(\sigma_r^2) \\ &\propto (\sigma_r^2)^{-1/2} \exp\left\{\frac{-1}{2\sigma_r^2}\lambda_r^2\right\} (\sigma_r^2)^{-a-1} \exp\left\{\frac{-1}{\sigma_r^2}h\right\} \\ &= (\sigma_r^2)^{-1/2-a-1} \exp\left\{\frac{-1}{\sigma_r^2}(\lambda_r^2/2 + h)\right\} \end{aligned}$$

Factorizando convenientemente, se encuentra una expresión idéntica a la función de densidad de una variable aleatoria con distribución $Inversa - Gama(-1/2 - a, \lambda_r^2/2 + h)$. ■

Nótese que del anterior razonamiento, y siguiendo el espíritu de los modelos mixtos en la práctica, es posible plantear el escenario en donde los efectos aleatorios estén correlacionados previa. Para esto se define la distribución previa para el vector de parámetros aleatorios como

$$\boldsymbol{\lambda} \mid \mathbf{L} \sim Normal(\mathbf{0}, \mathbf{L})$$

en donde $\mathbf{L}$ es una matriz de varianzas no diagonal que tiene distribución previa

$$\mathbf{L} \sim Inversa - Whishart\left(\frac{n}{2}, \frac{n\mathbf{S}}{2}\right)$$

con $\mathbf{S}$ una matriz cuadrada y conocida, del mismo orden que la matriz $\mathbf{L}$. Bajo este escenario, es fácil probar que la distribución posterior del vector de efectos aleatorios es normal multivariante, mientras que la distribución posterior de la matriz de varianzas de estos efectos aleatorios es inversa- Whishart.

Por otro lado, nótese también que una consecuencia directa del anterior resultado es que, para poder realizar inferencias acerca de los parámetros de interés, se deben utilizar técnicas MCMC para la simulación de valores provenientes de las distribuciones condicionales posterior. En particular el método de Gibbs es fácilmente aplicable y conveniente para obtener inferencias posterior para los parámetros de interés. Sin embargo, nótese que la distribución posterior condicional para el parámetro $\boldsymbol{\gamma}$ está dada por la siguiente expresión

$$\begin{aligned} p(\boldsymbol{\gamma} \mid \boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma_r^2, \mathbf{y}) &\propto p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \times p(\boldsymbol{\gamma}) \\ &\propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{\frac{-1}{2}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta} - \mathbf{W}'\boldsymbol{\lambda})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta} - \mathbf{W}'\boldsymbol{\lambda})\right\} \\ &\quad \times \exp\left\{\frac{-1}{2}(\boldsymbol{\gamma} - \mathbf{g})'\mathbf{G}^{-1}(\boldsymbol{\gamma} - \mathbf{g})\right\} \end{aligned}$$

con $\boldsymbol{\Sigma} = diag(\sigma_i^2)$. La anterior expresión no tiene una forma cerrada y además no es log-concava; por lo tanto utilizar métodos de simulación como el algoritmo de Gibbs no es conveniente puesto que las cadenas generadas pueden no converger. En este punto se propone la utilización de la metodología de (**?**), que utiliza el algoritmo de Metropolis-Hastings para la simulación de valores provenientes de la distribución posterior conjunta. Para esto, es necesario definir nuevas variables de trabajo aplicando una adaptación del algoritmo IRSL (Iterative Reweighted Least Squares) (**?**, p. 86) que fue utilizado por (**?**) en el contexto de los modelos lineales generalizados.

Resultado 14.1.2 (Algoritmo IRLS adaptado)**.** *Suponiendo que* $g(\cdot) = \log(\cdot)$ *y combinando la distribución previa del vector de parámetros en la modelación de la varianza con la siguiente variable de trabajo, resultante de la aplicación del método de Fisher-Scoring,*

$$\tilde{y}_i = \mathbf{z}_i'\boldsymbol{\gamma} + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\lambda})^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1 \qquad i = 1, \ldots, n. \tag{14.1.47}$$

cuya distribución es normal con media $\mathbf{z}_i'\boldsymbol{\gamma}$ *y varianza 2, siendo* $\boldsymbol{\beta}^{(c)}$, $\boldsymbol{\lambda}^{(c)}$, $\sigma_r^{2(c)}$ $(r = 1, \ldots, R)$, $\boldsymbol{\gamma}^{(c)}$ *los valores actuales de los parámetros de interés, una distribución de salto apropiada con kernel gausiano* $q_{\boldsymbol{\gamma}}$ *se obtiene como*

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(c)}) = Normal(\mathbf{g}^*, \mathbf{G}^*) \tag{14.1.48}$$

En donde,

$$\mathbf{G}^* = \left(\mathbf{G}^{-1} + \frac{1}{2}\mathbf{Z}'\mathbf{Z}\right)^{-1}$$
$$\mathbf{g}^* = \mathbf{G}^*\left(\mathbf{G}^{-1}\mathbf{g} + \frac{1}{2}\mathbf{Z}'\tilde{\mathbf{y}}\right)$$

Con $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$.

Prueba. El algoritmo IRLS requiere la aproximación de la transformación de las observaciones alrededor de los valores actuales de las estimaciones para los parámetros. Luego, se define

$$t_i = \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(c)} - \mathbf{w}_i'\boldsymbol{\lambda}^{(c)}\right)^2 \qquad i = 1, \ldots, n.$$

Claramente, t_i/σ^2 tiene distribución chi-cuadrado con un grado de libertad. Luego, $E(t_i) = \sigma^2$ y $Var(t_i) = 2\sigma^4$. Por otro lado, se define la variable de trabajo, acudiendo a la diferenciabilidad de $g(\cdot)$, como la aproximación de Taylor de primer orden de $g(t_i)$ evaluada en el punto $E(t_i)$. Así, dado que $\sigma_i^2 = g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)})$, se tiene que

$$\begin{aligned}\tilde{y}_i &:= g(E(t_i)) + g'(E(t_i))(t_i - E(t_i)) \\ &= g(\sigma_i^2) + g'(\sigma_i^2)(t_i - \sigma_i^2) \\ &= \mathbf{z}_i'\boldsymbol{\gamma}^{(c)} + g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))(t_i - g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))\end{aligned}$$

Ahora, utilizando los resultados de (**?**), se sigue que de la aplicación del método de Fisher-Scoring en alguna vecindad de $E(t_i) = \sigma^2$, la variable de trabajo $\tilde{y}_i$ tiene distribución normal con media dada por

$$\begin{aligned} E[\tilde{y}_i] &= E[\mathbf{z}_i'\boldsymbol{\gamma}^{(c)} + g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))(t_i - g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))] \\ &= \mathbf{z}_i'\boldsymbol{\gamma}^{(c)} + g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))E[(t_i - g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))] \\ &= \mathbf{z}_i'\boldsymbol{\gamma}^{(c)} + g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}) - g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)})) \\ &= \mathbf{z}_i'\boldsymbol{\gamma}^{(c)} \end{aligned}$$

y varianza dada por

$$\begin{aligned} Var[\tilde{y}_i] &= Var[\mathbf{z}_i'\boldsymbol{\gamma}^{(c)} + g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))(t_i - g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))] \\ &= (g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)})))^2 Var[t_i] \\ &= 2(g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)})))^2 (g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))^2 \end{aligned}$$

Luego, si $g(\cdot) = \log(\cdot)$, entonces la variable de trabajo quedaría definida por

$$\tilde{y}_i = \mathbf{z}_i'\boldsymbol{\gamma} + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\lambda})^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1$$

cuya distribución es normal con media $\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}$ y varianza 2. Luego, asumiendo independencia y definiendo a $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_i, \ldots, \tilde{y}_n)'$, se tiene que la pseudo-verosimilitud del vector de variables de trabajo está dada por

$$p(\tilde{\mathbf{y}} \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}, \boldsymbol{\gamma}^{(c)}, \sigma_r^2) \propto \exp\left\{ \frac{-1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}'\boldsymbol{\gamma}^{(c)})' \frac{1}{2}\mathbf{I}(\tilde{\mathbf{y}} - \mathbf{Z}'\boldsymbol{\gamma}^{(c)}) \right\} \qquad (14.1.49)$$

con $\mathbf{I}$, la matriz identidad. Por tanto, combinando la anterior distribución normal con la distribución previa normal del vector de parámetros $\boldsymbol{\gamma}$, siguiendo la regla de bayes como en la demostración del resultado 4.1, se tiene fácilmente que una distribución de salto apropiada con kernel gausiano $q_{\boldsymbol{\gamma}}$ se obtiene como (28). ■

Con base en lo expuesto anteriormente, se crea un algoritmo de Metropolis-Hastings modificado que permite la simulación de valores provenientes de las distribución posterior (26) mediante la escogencia de una distribución de salto apropiada para el modelamiento en las varianzas. A continuación, se describe este algoritmo que permite la realización de la debida inferencia posterior para el vector de parámetros $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_1^2, \ldots, \sigma_R^2)'$.

1. Iniciar el contador de iteraciones en $j = 1$.
2. Fijar valores iniciales para la cadena dados por $\boldsymbol{\beta}^{(j-1)}$, $\boldsymbol{\lambda}^{(j-1)}$, $\boldsymbol{\gamma}^{(j-1)}$, $\sigma_1^{2(j-1)}, \ldots, \sigma_R^{2(j-1)}$.
3. Actualizar el vector $\boldsymbol{\beta}$ a un nuevo valor $\boldsymbol{\beta}^{(j)}$, generado desde (22).
4. Actualizar el vector $\boldsymbol{\lambda}$ a un nuevo valor $\boldsymbol{\lambda}^{(j)}$, generado desde (23).
5. Para $r = 1, \ldots, R$ actualizar σ_r^2 a un nuevo valor $\sigma^{2(j)}$, generado desde (24).

6. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\gamma}}$, generado desde la distribución de salto (28).
7. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\gamma}}$, de otra manera $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$.
8. Actualizar el contador de la cadena de j a $j+1$.
9. Volver al paso 3 y repetir el procedimiento hasta que la cadena alcance la convergencia deseada.

El anterior algoritmo híbrido representa una estrategia eficiente para la simulación de observaciones de la distribución posterior conjunta. Este procedimiento hace parte de las estrategias llamadas *Metropolis-whitin-Gibbs* (**?**, p. 230) que son válidas puesto que la distribución estacionaria de las cadenas generadas es la distribución posterior de interés.

Para este algoritmo propuesto en este proyecto de tesis, se realizó un ejercicio empírico (ver apéndice B) y se observó que, aunque la tasa de aceptación[2] es baja y varía entre el 50 y 70 por ciento, las cadenas convergen rápidamente y las estimaciones son cercanas a los valores de verdaderos de los parámetros. Dado que esta metodología tiene en cuenta la heteroscedasticidad en las respuestas, se espera que las estimaciones sean mejores que las propuestas por la metodología clásica que no las tiene en cuenta. Por supuesto, en el desarrollo de la tesis se complementará este análisis básico con la metodología dada en la sección 5.

14.1.7 Propuesta para modelos no-lineales mixtos

La definición general del modelo no lineal mixto está dada por la relación del vector de variables respuestas $\mathbf{y} = (y_1, \ldots, y_n)'$ con un vector de medias $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ de la siguiente manera

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{14.1.50}$$

donde, $\boldsymbol{\varepsilon}$ es un vector de errores y las componentes de $\boldsymbol{\mu}$ son funciones no lineales de los parámetros fijos y aleatorios ligados con variables de información auxiliar en $\mathbf{X}$ y en $\mathbf{W}$, definidas en la sección anterior. De esta forma, se define la componente media del modelo para el i-ésimo individuo como

$$\mu_i = f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) \tag{14.1.51}$$

donde f es una función no lineal y diferenciable. De la misma manera, se asume que los componentes del vector de errores tienen distribución normal con media nula y varianza no constante $Var(\varepsilon_i) = \sigma_i^2$. En la misma vía que en la sección anterior, se supone que la heteroscedasticidad en las respuestas puede ser modelada como en (20).

Con base en lo expuesto anteriormente, en esta sección se propone una metodología para la estimación bayesiana de los parámetros de interés utilizando como

[2]En un ejercicio empírico, (**?**) demostraron que una tasa de aceptación entre el 15 y el 50 por ciento guía a una eficiencia cercana al ochenta por ciento.

línea base la linealización de las funciones no lineales involucradas en el modelo $f()$, para la media y $g()$, para la varianza. Una vez que el modelo se pueda aproximar por la linealización de las funciones pertinentes, se utilizará una adaptación del algoritmo IRLS para obtener un kernel de las distribuciones de salto en un algoritmo híbrido de Metropolis-Hastings. Esta idea surge del trabajo desarrollado por (**?**) y (**?**) en modelos de regresión lineales y no lineales.

En primer, y siguiendo las ideas de la sección pasada, las distribuciones previa para los parámetros de interés en el modelo están dadas por

$$\begin{aligned}
\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y} &\sim Normal(\mathbf{b}, \mathbf{B}) \\
\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y} &\sim Normal(\mathbf{g}, \mathbf{G}) \\
\lambda_r \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_r^2, \mathbf{y} &\sim Normal(0, \sigma_r^2) \\
\sigma_r^2 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y} &\sim Inversa - Gama(a, h)
\end{aligned}$$

para todo $r = 1, \ldots, R$. Al igual que en la sección pasada, la varianza de los efectos aleatorios y la varianza de la respuesta se consideran independientes previa. Este mismo supuesto de independencia se realiza con los parámetros de los efectos fijos, aleatorios y los que modelan la varianza. Luego, la distribución previa de todos los parámetros involucrados en el modelo está dada por la siguiente expresión:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2) = p(\boldsymbol{\beta}) p(\boldsymbol{\gamma}) \prod_{r=1}^{R} p(\lambda_r \mid \sigma_r^2) p(\sigma_r^2)$$

Por otro lado la función de verosimilitud de las observaciones se escribe de la siguiente manera

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ \frac{-1}{2} (\mathbf{y} - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}))' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})) \right\} \tag{14.1.52}$$

Entonces, siguiendo la regla de bayes, la expresión que se presenta a continuación da cuenta de la distribución posterior completa para todos los parámetros del modelo:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \times p(\boldsymbol{\beta}) p(\boldsymbol{\gamma}) \prod_{r=1}^{R} p(\lambda_r \mid \sigma_r^2) p(\sigma_r^2)$$

Ahora, esta distribución no tiene una forma cerrada conocida, además de ser intratable analíticamente. Lo propio sucede también con las distribuciones posterior condicionales: $p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y})$, $p(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_r^2, \mathbf{y})$ y $p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y})$. Lo anterior garantiza que el muestreo para $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ y $\boldsymbol{\gamma}$ no sea sencillo de implementar.

Por el espíritu jerárquico del modelo, los parámetros de la varianza de los efectos aleatorios son los únicos cuya distribución condicional posterior es conjugada con respecto a su distribución previa. Luego, dados valores de los parámetros de interés, es posible utilizar el algoritmo de Gibbs para seleccionar muestras provenientes de esta distribución posterior y por consiguiente realizar inferencias acerca de los parámetros σ_r^2 $(r = 1, \ldots, R)$.

Resultado 14.1.3. *La distribución condicional posterior de los parámetros de interés σ_r^2 está dada por*

$$\sigma_r^2 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y} \sim Inversa - Gama(-1/2 - a, \lambda_r^2/2 + h) \tag{14.1.53}$$

Prueba. La demostración es inmediata siguiendo el razonamiento de la última parte de la verificación del resultado 4.1. ■

En primer lugar, nos enfocaremos en los parámetros que incumben en la función de media $f()$, es decir $\boldsymbol{\beta}$ y $\boldsymbol{\lambda}$. (**?**) han generalizado, para el caso frecuentista, los supuestos del modelo lineal mixto para estructura generales de varianza y función de medias no lineal. Su enfoque se basa en la linealización de Taylor de primer orden de la media. De esta manera, acudiendo a esta idea, se propone aproximar la verosimilitud (32) a una normal mediante la creación de una variable de trabajo. Luego, una vez más adaptando el algoritmo IRLS, se propone un kernel de las distribuciones de salto para ser implementadas en un algoritmo de Metropolis-Hastings.

Resultado 14.1.4. *Siendo $\boldsymbol{\beta}^{(c)}$ y $\boldsymbol{\lambda}^{(c)}$ lo valores actuales de los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\lambda}$, respectivamente, la función $f()$ se aproxima mediante la siguiente expresión*

$$f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) \cong f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)}) + \tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)}) \tag{14.1.54}$$

Donde

$$\tilde{\mathbf{x}}_i = \nabla_{\boldsymbol{\beta}} f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) = \left[\frac{\partial f}{\partial \beta_1}, \ldots, \frac{\partial f_i}{\partial \beta_P}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}=\boldsymbol{\lambda}^{(c)}} \tag{14.1.55}$$

y

$$\tilde{\mathbf{w}}_i = \nabla_{\boldsymbol{\lambda}} f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) = \left[\frac{\partial f}{\partial \lambda_1}, \ldots, \frac{\partial f_i}{\partial \lambda_R}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}=\boldsymbol{\lambda}^{(c)}} \tag{14.1.56}$$

Prueba. La demostración es inmediata aplicando la bien conocida linealización de Taylor de primer orden para una función multivariante en el punto $(\boldsymbol{\beta} = \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda} = \boldsymbol{\lambda}^{(c)})$ y notando que

$$\begin{aligned}\nabla f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) \begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}^{(c)} \\ \boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)} \end{pmatrix} &= \nabla_{\boldsymbol{\beta}} f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)}) \\ &\quad + \nabla_{\boldsymbol{\lambda}} f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)})(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)})\end{aligned}$$

■

Una vez se ha encontrado la aproximación de la función de media, se procede a proponer unas variables de trabajo convenientes de tal manera que su distribución tenga la forma de una normal para combinarla con la distribución previa de los parámetros de interés. El siguiente resultado provee estas variables de trabajo y las distribuciones de salto generadas por ellas.

Resultado 14.1.5. *Suponiendo que $f()$ se aproxima mediante (33) y combinando la distribución previa del vector de parámetros $\boldsymbol{\beta}$ con la aproximación de la verosimilitud generada con la siguiente variable de trabajo*

$$\tilde{y}_i^b = y_i - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i \boldsymbol{\beta}^{(c)} - \tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)}) \tag{14.1.57}$$

cuya distribución condicional es normal con media $\tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)}$ y varianza σ_i^2, siendo $\boldsymbol{\beta}^{(c)}$, $\boldsymbol{\lambda}^{(c)}$, σ_r^2 $(r = 1, \ldots, R)$ y $\boldsymbol{\gamma}^{(c)}$ los valores actuales de los parámetros de interés, una distribución de salto apropiada con kernel gausiano $q_{\boldsymbol{\beta}}$ se obtiene como

$$q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(c)}) = Normal(\mathbf{b}^*, \mathbf{B}^*) \tag{14.1.58}$$

En donde,

$$\mathbf{B}^* = \left(\mathbf{B}^{-1} + \tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}$$
$$\mathbf{b}^* = \mathbf{B}^*\left(\mathbf{B}^{-1}\mathbf{b} + \tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}}_b\right)$$

Con $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1', \ldots, \tilde{\mathbf{x}}_n')$ e $\tilde{\mathbf{y}}_b = (\tilde{y}_1^b, \ldots, \tilde{y}_n^b)'$.

Análogamente para el vector de efectos aleatorios $\boldsymbol{\lambda}$, se define la la siguiente variable de trabajo

$$\tilde{y}_i^l = y_i - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{w}}_i \boldsymbol{\lambda}^{(c)} - \tilde{\mathbf{x}}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)}) \tag{14.1.59}$$

cuya distribución condicional es normal con media $\tilde{\mathbf{w}}_i\boldsymbol{\lambda}^{(c)}$ y varianza σ_i^2. Luego, una distribución de salto apropiada con kernel gausiano $q_{\boldsymbol{\lambda}}$ se obtiene como

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^{(c)}) = Normal(\mathbf{l}^*, \mathbf{L}^*) \tag{14.1.60}$$

En donde,

$$\mathbf{L}^* = \left(\mathbf{L}^{-1} + \tilde{\mathbf{W}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{W}}\right)^{-1}$$
$$\mathbf{l}^* = \mathbf{L}^*\left(\tilde{\mathbf{W}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{y}}_l\right)$$

Con $\mathbf{L} = diag(\sigma_r^2)$, $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1', \ldots, \tilde{\mathbf{w}}_n')$ e $\tilde{\mathbf{y}}_l = (\tilde{y}_1^l, \ldots, \tilde{y}_n^l)'$.

Prueba. En primer lugar, como consecuencia del Resultado 4.4, se tiene que la aproximación de la verosimilitud (32) está dada por el reemplazo de $f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})$ por $f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)})\tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)})$. En el primer término de la forma cuadrática dentro de la verosimilitud se tiene que

$$\begin{aligned} y - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) &\cong y - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)})\tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)}) \\ &= y - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} - \tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)}) - \tilde{\mathbf{x}}_i\boldsymbol{\beta} \\ &= \tilde{y}_i^b - \tilde{\mathbf{x}}_i\boldsymbol{\beta} \end{aligned}$$

Luego, condicionando en el punto $\boldsymbol{\beta} = \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda} = \boldsymbol{\lambda}^{(c)}$, la variable de trabajo $\tilde{y}_i^b$ tiene distribución normal con media

$$\begin{aligned} E(\tilde{y}_i^b) &= E(y - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} - \tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)})) \\ &= f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} - \tilde{\mathbf{w}}_i(\boldsymbol{\lambda}^{(c)} - \boldsymbol{\lambda}^{(c)}) \\ &= \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} \end{aligned}$$

y varianza

$$Var(\tilde{y}_i^b) = Var(y - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i\boldsymbol{\beta}^{(c)} - \tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)})) = Var(y) = \sigma_i^2$$

Por tanto la pseudo-verosimilitud de las variables de trabajo está dada por

$$p(\tilde{\mathbf{y}}_b \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ \frac{-1}{2} (\tilde{\mathbf{y}}_b - \tilde{\mathbf{X}}'\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\tilde{\mathbf{y}}_b - \tilde{\mathbf{X}}'\boldsymbol{\beta}) \right\}$$

Luego, combinando la anterior distribución normal con la distribución previa normal del vector de parámetros $\boldsymbol{\beta}$, siguiendo la regla de bayes, se tiene que una distribución de salto apropiada con kernel gaussiano se obtiene como (38).

La demostración para el vector de efectos aleatorios $\boldsymbol{\lambda}$ es similar, notando que

$$\begin{aligned} y - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) &\cong y - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}) + \tilde{\mathbf{x}}_i(\boldsymbol{\beta} - \boldsymbol{\beta}^{(c)}) \tilde{\mathbf{w}}_i(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(c)}) \\ &= \tilde{y}_i^l - \tilde{\mathbf{w}}_i\boldsymbol{\lambda} \end{aligned}$$

Por lo tanto, condicionando en el punto $\boldsymbol{\beta} = \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda} = \boldsymbol{\lambda}^{(c)}$, la variable de trabajo $\tilde{y}_i^l$ tiene distribución normal con media $E(\tilde{y}_i^l) = \tilde{\mathbf{w}}_i\boldsymbol{\lambda}^{(c)}$ y varianza $Var(\tilde{y}_i^b) = \sigma_i^2$. Entonces, la pseudo-verosimilitud de las variables de trabajo está dada por

$$p(\tilde{\mathbf{y}}_l \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \sigma_r^2) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ \frac{-1}{2} (\tilde{\mathbf{y}}_l - \tilde{\mathbf{W}}'\boldsymbol{\lambda})' \boldsymbol{\Sigma}^{-1} (\tilde{\mathbf{y}}_l - \tilde{\mathbf{W}}'\boldsymbol{\lambda}) \right\}$$

Combinando la anterior distribución normal con la distribución previa normal del vector de parámetros $\boldsymbol{\lambda}$, se obtiene una distribución de salto apropiada con kernel gaussiano como en (40). ■

Con el anterior procedimiento es posible proponer valores para los parámetros de interés de la función de media desde las distribuciones de salto. A continuación, nos enfocaremos en la inferencia para el vector de parámetros de interés en la función de varianza $g()$. Se utilizará una propuesta similar a la de la anterior sección, mediante la implementación de Fisher-Scoring y la creación de una variable de trabajo que permita usar la adaptación del algoritmo IRLS. Dado que $p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y})$, la distribución condicional posterior del parámetro $\boldsymbol{\gamma}$, no tiene una forma cerrada y no es log-concava, se debe aproximar la verosimilitud de los datos y combinarla con la distribución previa de los parámetros de interés para conseguir una distribución de salto conveniente que permita la generación de valores de $p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma_r^2, \mathbf{y})$. El siguiente resultado muestra el procedimiento para conseguirlo.

Resultado 14.1.6 (Algoritmo IRLS adaptado)**.** *Suponiendo que* $g(\cdot) = \log(\cdot)$ *y combinando la distribución previa del vector de parámetros en la modelación de la varianza con la siguiente variable de trabajo, resultante de la aplicación del método de Fisher-Scoring,*

$$\tilde{y}_i = \mathbf{z}_i'\boldsymbol{\gamma} + \frac{(y_i - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}))^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1 \qquad i = 1, \ldots, n. \tag{14.1.61}$$

para todo $i = 1, \ldots, n$ *cuya distribución es normal con media* $\mathbf{z}_i'\boldsymbol{\gamma}$ *y varianza 2, siendo* $\boldsymbol{\beta}^{(c)}$, $\boldsymbol{\lambda}^{(c)}$, $\sigma_r^{2(c)}$ $(r = 1, \ldots, R)$, $\boldsymbol{\gamma}^{(c)}$ *los valores actuales de los parámetros de interés, una distribución de salto apropiada con kernel gausiano* $q_{\boldsymbol{\gamma}}$ *se obtiene como*

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(c)}) = Normal(\mathbf{g}^*, \mathbf{G}^*) \tag{14.1.62}$$

En donde,

$$\mathbf{G}^* = \left(\mathbf{G}^{-1} + \frac{1}{2}\mathbf{Z}'\mathbf{Z}\right)^{-1}$$

$$\mathbf{g}^* = \mathbf{G}^*\left(\mathbf{G}^{-1}\mathbf{g} + \frac{1}{2}\mathbf{Z}'\tilde{\mathbf{y}}\right)$$

Con $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$.

Prueba. Se define

$$t_i = \left(y_i - f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)})\right)^2 \qquad i = 1, \ldots, n.$$

Luego, t_i/σ^2 tiene distribución chi-cuadrado con un grado de libertad; por lo tanto $E(t_i) = \sigma^2$ y $Var(t_i) = 2\sigma^4$. La variable de trabajo es la aproximación de Taylor de primer orden de $g(t_i)$ evaluada en el punto $E(t_i)$. Así, dado que $\sigma_i^2 = g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)})$, se tiene que

$$\tilde{y}_i = \mathbf{z}_i'\boldsymbol{\gamma}^{(c)} + g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))(t_i - g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))$$

Utilizando los resultados de (**?**), se sigue que de la aplicación del método de Fisher-Scoring en alguna vecindad de $E(t_i) = \sigma^2$, la variable de trabajo $\tilde{y}_i$ tiene distribución normal con media dada por

$$E[\tilde{y}_i] = \mathbf{z}_i'\boldsymbol{\gamma}^{(c)}$$

y varianza dada por

$$Var[\tilde{y}_i] = 2(g'(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)})))^2(g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}))^2$$

Cuando $g(\cdot) = \log(\cdot)$, entonces la variable de trabajo quedaría definida por (40) cuya distribución es normal con media $\mathbf{z}_i'\boldsymbol{\gamma}^{(c)}$ y varianza 2. Luego, se tiene que la pseudo-verosimilitud del vector de variables de trabajo está dada por

$$p(\tilde{\mathbf{y}} \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}, \boldsymbol{\gamma}^{(c)}, \sigma_r^2) \propto \exp\left\{\frac{-1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}'\boldsymbol{\gamma}^{(c)})'\frac{1}{2}\mathbf{I}(\tilde{\mathbf{y}} - \mathbf{Z}'\boldsymbol{\gamma}^{(c)})\right\} \tag{14.1.63}$$

Combinando la anterior distribución normal con la distribución previa normal del vector de parámetros $\boldsymbol{\gamma}$, se obtiene que una distribución de salto apropiada con kernel gausiano $q_{\boldsymbol{\gamma}}$ se obtiene como (41). ■

Con base en las propuestas anteriores, se propone un algoritmo de Metropolis-Hastings modificado que permite la simulación de valores provenientes de la distribución posterior mediante la escogencia de tres distribuciones de salto apropiadas para el modelamiento en la función de media y en la función de varianza. Este algoritmo se describe a continuación:

1. Iniciar el contador de iteraciones en $j = 1$.
2. Fijar valores iniciales para la cadena dados por $\boldsymbol{\beta}^{(j-1)}$, $\boldsymbol{\lambda}^{(j-1)}$, $\boldsymbol{\gamma}^{(j-1)}$, $\sigma_1^{2(j-1)}, \ldots, \sigma_R^{2(j-1)}$.
3. Para $r = 1, \ldots, R$ actualizar σ_r^2 a un nuevo valor $\sigma^{2(j)}$, generado desde (33).
4. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\beta}}$, generado desde la distribución de salto (38).
5. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\beta}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\beta}}$, de otra manera $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$.
6. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\lambda}}$, generado desde la distribución de salto (40).
7. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\lambda}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\lambda}}$, de otra manera $\boldsymbol{\lambda}^{(j)} = \boldsymbol{\lambda}^{(j-1)}$.
8. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\gamma}}$, generado desde la distribución de salto (42).
9. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\gamma}}$, de otra manera $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$.
10. Actualizar el contador de la cadena de j a $j + 1$.
11. Volver al paso 3 y repetir el procedimiento hasta que la cadena alcance la convergencia deseada.

Nótese que, en particular, cuando $f(\mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}$, entonces el algoritmo anterior se convierte en una réplica del algoritmo de la sección 4.1 puesto que, en el punto $\boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}, \boldsymbol{\gamma}^{(c)}$, se tiene lo siguiente:

- $\tilde{\mathbf{X}} = \mathbf{X}$ y $\tilde{\mathbf{W}} = \mathbf{W}$.
- El vector de variables de trabajo $\tilde{\mathbf{y}}_b$ se convierte en el vector $\mathbf{y}_b$ definido en el resultado 4.1 y la distribución de salto $q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(c)})$ coincide absolutamente con la distribución condicional posterior para $\boldsymbol{\beta}$. Luego, la probabilidad de aceptación es uno en todos los pasos del algoritmo.
- El vector de variables de trabajo $\tilde{\mathbf{y}}_l$ se convierte en el vector $\mathbf{y}_l$ definido en el resultado 4.1 y la distribución de salto $q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^{(c)})$ coincide absolutamente con la distribución condicional posterior para $\boldsymbol{\lambda}$. Luego, la probabilidad de aceptación es uno en todos los pasos del algoritmo.
- Para la inferencia de $\boldsymbol{\gamma}$, la variable de trabajo $\tilde{y}_i$ coincide con la variable de trabajo propuesta en el resultado 4.2. Las distribuciones de salto propuestas para ambas variables de trabajo también coinciden.

14.1.8 Propuesta para modelos lineales generalizados mixtos

Para datos de naturaleza continua y cuando la respuesta puede ser modelada mediante una verosimilitud gaussiana, se proponen en este trabajo doctoral los modelos de las dos secciones pasadas; sin embargo, cuando la respuesta se desvía de la normalidad o incluso cuando la respuesta es discreta se necesita un marco de trabajo que permita el ajuste de este tipo de modelos. La alternativa que se propone es encasillar estas diferentes verosimilitudes en una familia de distribuciones exponencial. En esta sección se amplia el desarrollo de la parametrización propuesta en (**?**) para modelos lineales generalizados en donde se puede escribir la verosimilitud de las observaciones de forma que pertenezcan a la familia exponencial bi-paramétrica (**?**). Una de las características de esta notación es que los parámetros de la verosimilitud resultan ser ortogonales. Sin embargo, teniendo en cuenta que distribuciones importantes como la Beta no pertenecen a esta familia, se decide trabajar bajo el contexto general; es decir, cuando no necesariamente se presenta ortogonalidad también es posible, mediante la propuesta de este trabajo doctoral y la adaptación del trabajo de (**?**), obtener estimaciones y realizar el proceso de inferencia de los parámetros de interés.

De esta forma se pretende abarcar una gran cantidad de modelos que son bastante usados en la vida práctica del estadístico; entre otros, se encuentran los modelos con respuesta gamma, beta y normal. Gracias a la parametrización utilizada es fácil demostrar que esta familia cubre modelos que llegan hasta la regresión Poisson y los modelos logísticos con respuesta dicotómica, generalizando el desarrollo bayesiano de (**?**).

En un trabajo reciente, (**?**) han reparametrizado la formulación del modelo lineal generalizado en la misma vía que (**?**). Bajo esta reparametrización se demuestra que, suponiendo condiciones de regularidad, μ resulta ortogonal a τ en el sentido de (**?**) y (**?**). Lo anterior conlleva a que de un modelo lineal mixto generalizado clásico sea posible pasar fácilmente a un modelo lineal mixto generalizado de sobredispersión (**?**) cuando se supone además que asociadas con las respuestas y_i existe un vector de covariables $\mathbf{z}_i$ tales que

$$g(\tau_i) = \mathbf{z}_i'\boldsymbol{\gamma} \tag{14.1.64}$$

Donde g es una función estrictamente monótona y diferenciable. Por tanto para forzar la sobredispersión en el modelo ($\tau_i > 0$) es posible tomar, por ejemplo, $\tau_i = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$. En este trabajo doctoral se propone utilizar el mismo enfoque de modelación de (**?**), en términos de los parámetros de interés, pero sin tener en cuenta la ortogonalidad. Es decir, sin importar que le modelo de respuesta induzca o no parámetros ortogonales, se pretende modelar $h(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}$ y $g(\tau_i) = \mathbf{z}_i'\boldsymbol{\gamma}$, de forma conjunta. De manera general, la no ortogonalidad se puede incluir en la distribución conjunta previa de los parámetros de interés. La figura 1 muestra una representación simbólica del contexto general de este modelo.

En este caso, la formulación de las distribuciones previa se debe hacer en forma general y conjunta para involucrar la posible no ortogonalidad de los parámetros. Sin embargo, se sigue suponiendo que los parámetros de los efectos aleatorios son

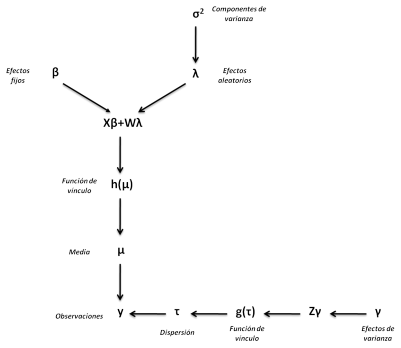

Figura 14.1: *Representación simbólica de un modelo lineal generalizado mixto con estructura de sobredispersión.*

independientes previa de los parámetros de efectos fijo; esto no ocurre entre los parámetros de efectos fijo con los parámetros de dispersión ni entre los parámetros de efectos aleatorios con los parámetros de dispersión. De esta manera, se tiene el siguiente sistema de distribuciones previa

$$\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma} \mid \mathbf{L} \sim Normal \left[\begin{pmatrix} \mathbf{b}_0 \\ \mathbf{0} \\ \mathbf{g}_0 \end{pmatrix}, \begin{pmatrix} \mathbf{B}_0 & \mathbf{0} & \mathbf{C}_1 \\ \mathbf{0} & \mathbf{L} & \mathbf{C}_2 \\ \mathbf{C}_1' & \mathbf{C}_2' & \mathbf{G}_0 \end{pmatrix} \right] \tag{14.1.65}$$

Nótese que es el vector de efectos aleatorios está supeditado a los componentes de varianza en la matriz aleatoria $\mathbf{L}$. Por tanto, es necesario proponer una distribución previa para modelar la variabilidad del vector $\boldsymbol{\lambda}$. Para esto, un acercamiento general es la asignación de la siguiente distribución previa

$$\mathbf{L} \sim Inversa - Whishart_v(\mathbf{S}^{-1}) \tag{14.1.66}$$

Con base en lo anterior, la distribución previa conjunta toma la siguiente forma

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{L}) = p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma} \mid \mathbf{L}) p(\mathbf{L})$$

Utilizando los resultados de la teoría estadística para distribuciones multivariantes, y después de un rápido desarrollo algebraico, se tiene que las distribuciones condicionales previa para los vectores involucrados en el modelamiento de la media, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ y para el vectores involucrados en el modelamiento de la varianza, $\boldsymbol{\gamma}$, son

$$\begin{aligned}\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{L} &\sim Normal(\mathbf{b}, \mathbf{B}) \\ \boldsymbol{\lambda} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L} &\sim Normal(\mathbf{l}, \mathbf{L}) \\ \boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L} &\sim Normal(\mathbf{g}, \mathbf{G})\end{aligned}$$

respectivamente. En donde $\mathbf{b}$, $\mathbf{l}$, $\mathbf{g}$ y $\mathbf{B}$, $\mathbf{L}$, $\mathbf{G}$ se derivan de la distribución previa conjunta. Nótese que cuando $\mathbf{C}_1 = \mathbf{0}$ y $\mathbf{C}_2 = \mathbf{0}$, es decir cuando el supuesto de ortogonalidad parezca plausible de asumir en el modelo, entonces $\mathbf{b} = \mathbf{b}_0$, $\mathbf{l} = \mathbf{0}$ y $\mathbf{g} = \mathbf{g}_0$.

Por otro lado, Teniendo en cuenta que la función de verosimilitud del vector de observaciones $\mathbf{y} = (y_1, \ldots, y_n)'$ es

$$p(\mathbf{y} \mid \theta_i, \tau_i) \propto \exp\left\{\sum_{i=1}^{n} [d_1(\theta_i, \tau_i)T_1(y_i) + d_2(\theta_i, \tau_i)T_2(y_i) + b(\theta_i, \tau_i)]\right\} \quad (14.1.67)$$

Entonces, siguiendo la regla de bayes, la distribución posterior completa para todos los parámetros está dada entonces por la siguiente expresión

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{L} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta_i, \tau_i)p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma} \mid \mathbf{L})p(\mathbf{L})$$

Dado que la anterior distribución no tiene una forma cerrada, es difícil obtener directamente de ésta, muestras de los parámetros de interés. Por lo tanto, utilizar el enfoque del condicionamiento sucesivo pondría arrojar mejores resultados. De esta manera, se obtiene la distribución posterior condicional para el vector de parámetros $\boldsymbol{\beta}$ como

$$\begin{aligned}&p(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{L}, \mathbf{y}) \\ &\propto p(\mathbf{y} \mid \theta_i, \tau_i)p(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \boldsymbol{\gamma}\mathbf{L}) \\ &\propto \exp\left\{\frac{-1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{B}^{-1}(\boldsymbol{\beta} - \mathbf{b}) + \sum_{i=1}^{n} [d_1(\theta_i, \tau_i)T_1(y_i) + d_2(\theta_i, \tau_i)T_2(y_i) + b(\theta_i, \tau_i)]\right\}\end{aligned} \quad (14.1.68)$$

La distribución posterior condicional para el vector de parámetros $\boldsymbol{\lambda}$ es

$$\begin{aligned}&p(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L}, \mathbf{y}) \\ &\propto p(\mathbf{y} \mid \theta_i, \tau_i)p(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L}) \\ &\propto \exp\left\{\frac{-1}{2}(\boldsymbol{\lambda} - \mathbf{l})'\mathbf{L}^{-1}(\boldsymbol{\lambda} - \mathbf{l}) + \sum_{i=1}^{n} [d_1(\theta_i, \tau_i)T_1(y_i) + d_2(\theta_i, \tau_i)T_2(y_i) + b(\theta_i, \tau_i)]\right\}\end{aligned} \quad (14.1.69)$$

Por último, la distribución posterior condicional para el vector de parámetros $\boldsymbol{\gamma}$ es

$$\begin{aligned} & p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{L}, \mathbf{y}) \\ & \propto p(\mathbf{y} \mid \theta_i, \tau_i) p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{L}) \\ & \propto \exp\left\{ \frac{-1}{2}(\boldsymbol{\gamma} - \mathbf{g})' \mathbf{G}^{-1} (\boldsymbol{\gamma} - \mathbf{g}) + \sum_{i=1}^{n} [d_1(\theta_i, \tau_i) T_1(y_i) + d_2(\theta_i, \tau_i) T_2(y_i) + b(\theta_i, \tau_i)] \right\} \end{aligned} \tag{14.1.70}$$

Fácilmente se puede corroborar que ninguna de las distribuciones en (51) - (53) tiene una forma conocida o cerrada. Por tanto no es posible obtener muestras, utilizando el condicionamiento sucesivo, a menos que la verosimilitud de las observaciones fuese norma y las funciones de media h y varianza h fueran la identidad. En este caso es posible utilizar una variante del algoritmo de Gibbs para seleccionar muestras de las distribuciones condicionales. El único parámetro que sí tiene una distribución previa conjugada es $\mathbf{L}$, tal como lo afirma el siguiente resultado.

Resultado 14.1.7. *La distribución condicional posterior de* $\mathbf{L}$ *está dada por*

$$\mathbf{L} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{L}, \mathbf{y} \sim Inversa - Whishart_{v+1+R}(\mathbf{S} + \mathbf{S}_{\boldsymbol{\lambda}}) \tag{14.1.71}$$

Donde R *es el número de componentes en el vector de los efectos aleatorios y*

$$\mathbf{S}_{\boldsymbol{\lambda}} = (\boldsymbol{\lambda} - \mathbf{l})(\boldsymbol{\lambda} - \mathbf{l})'.$$

Prueba. En primer lugar, se ha de notar que la distribución condicional previa para $\boldsymbol{\lambda}$ es normal multivariante y puede ser escrita como

$$p(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L}) \propto |\mathbf{L}|^{-1/2} \exp\left\{ \frac{-1}{2} traza(\mathbf{L}^{-1} \mathbf{S}_{\boldsymbol{\lambda}}) \right\}$$

Por otro lado, utilizando el condicionamiento sucesivo, se encuentra que

$$\begin{aligned} p(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L}, \mathbf{y}) &\propto p(\boldsymbol{\lambda}, \underbrace{\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L}}_{fijos} \mid \mathbf{y}) \\ &\propto p(\boldsymbol{\lambda} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{L}) p(\mathbf{L}) \\ &\propto |\mathbf{S}|^{v/2} |\mathbf{L}|^{-(v+R+1+1)/2} \exp\left\{ \frac{-1}{2} traza(\mathbf{S}\mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{S}_{\boldsymbol{\lambda}}) \right\} \\ &= |\mathbf{S}|^{v/2} |\mathbf{L}|^{-(v+1+R+1)/2} \exp\left\{ \frac{-1}{2} traza(\mathbf{L}^{-1}[\mathbf{S} + \mathbf{S}_{\boldsymbol{\lambda}}]) \right\} \end{aligned}$$

Factorizando convenientemente, se encuentra una expresión idéntica a la función de densidad inversa-Whishart para una matriz aleatoria con los parámetros mencionados. ■

Al igual que en la sección pasada, el tratamiento del desconocimiento de una forma cerrada por parte de las funciones de distribución involucradas en los parámetros μ y τ será tratado en dos instancias. En principio, se propondrá una distribución de salto general para los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$, generada mediante la implementación y modificación del algoritmo IRLS. El tratamiento será similar para proponer una distribución de salto general para el parámetro $\boldsymbol{\gamma}$. Al final se tendrán tres distribuciones de salto adecuadas para implementar un algoritmo híbrido y así poder generar muestras de las distribuciones posterior condicionales. Esta metodología sigue los lineamientos del trabajo desarrollado por (**?**). El siguiente resultado provee las variables de trabajo y las distribuciones de salto generadas para la inferencia condicional de $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$.

Resultado 14.1.8 (Algoritmo IRLS adaptado)**.** *Combinando la distribución previa del vector de parámetros $\boldsymbol{\beta}$ con la siguiente variable de trabajo, resultante de la aplicación del método de Fisher-Scoring,*

$$\tilde{y}_i^b = \mathbf{x}_i'\boldsymbol{\beta} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))(y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})) \tag{14.1.72}$$

cuya distribución condicional es normal con media $\mathbf{x}_i'\boldsymbol{\beta}$ y varianza $\tilde{V}_i = (h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})))^2 Var(y_i)$, siendo $\boldsymbol{\beta}^{(c)}$, $\boldsymbol{\lambda}^{(c)}$, $\boldsymbol{\gamma}^{(c)}$ los valores actuales de los parámetros de interés, una distribución de salto apropiada con kernel gaussiano $q_{\boldsymbol{\beta}}$ se obtiene como

$$q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(c)}) = Normal(\mathbf{b}^*, \mathbf{B}^*) \tag{14.1.73}$$

En donde,

$$\mathbf{B}^* = \left(\mathbf{B}^{-1} + \mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}$$
$$\mathbf{b}^* = \mathbf{B}^*\left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{y}}_b\right)$$

con $\tilde{\mathbf{V}} = diag(\tilde{V}_i)$ e $\tilde{\mathbf{y}}_b = (\tilde{y}_1^b, \ldots, \tilde{y}_n^b)'$.

Análogamente para el vector de efectos aleatorios $\boldsymbol{\lambda}$, se define la la siguiente variable de trabajo

$$\tilde{y}_i^l = \mathbf{w}_i'\boldsymbol{\lambda} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))(y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})) \tag{14.1.74}$$

cuya distribución condicional es normal con media $\mathbf{w}_i'\boldsymbol{\lambda}$ y varianza $\tilde{V}_i$. Luego, una distribución de salto apropiada con kernel gaussiano $q_{\boldsymbol{\lambda}}$ se obtiene como

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^{(c)}) = Normal(\mathbf{l}^*, \mathbf{L}^*) \tag{14.1.75}$$

En donde,

$$\mathbf{L}^* = \left(\mathbf{L}^{-1} + \mathbf{W}'\tilde{\mathbf{V}}^{-1}\mathbf{W}\right)^{-1}$$
$$\mathbf{l}^* = \mathbf{L}^*\left(\mathbf{L}^{-1}\mathbf{l} + \mathbf{W}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{y}}_l\right)$$

con $\tilde{\mathbf{y}}_l = (\tilde{y}_1^l, \ldots, \tilde{y}_n^l)'$.

Prueba. Se define $t_i = y_i$, la cual tiene distribución en la familia exponencial bi-paramétrica con $E(t_i) = \mu_i$ y $Var(t_i) = v_i$. De lo anterior se define la variable de trabajo general para la media, $\tilde{y}_i$, acudiendo a las propiedades de la función de vínculo $h()$, como la aproximación de Taylor de primer orden de $h(t_i)$ evaluada en el punto $E(t_i)$. De esta manera, dado que $\mu_i = h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})$, se tiene lo siguiente

$$\begin{aligned}\tilde{y}_i &:= h(E(t_i)) + h'(E(t_i))(t_i - E(t_i)) \\ &= h(\mu_i) + h'(\mu_i)(t_i - \mu_i) \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))(y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))\end{aligned}$$

Luego, de la aplicación del método de Fisher-Scoring en alguna vecindad de $E(t_i) = \mu_i$, la variable de trabajo $\tilde{y}_i$ tiene distribución normal con media dada por

$$\begin{aligned}E[\tilde{y}_i] &= E[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))(y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))E[y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}\end{aligned}$$

y varianza $\tilde{V}_i$ dada por

$$\begin{aligned}\tilde{V}_i := Var[\tilde{y}_i] &= Var[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))(y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))] \\ &= (h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})))^2 Var[y_i] \\ &= (h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})))^2 v_i\end{aligned}$$

Con base en lo anterior, para el vector de parámetros de efectos fijo $\boldsymbol{\beta}$, se define la variable de trabajo $\tilde{y}_i^b$ dada por

$$\tilde{y}_i^b := \tilde{y}_i - \mathbf{w}_i'\boldsymbol{\lambda} = \mathbf{x}_i'\boldsymbol{\beta} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))(y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))$$

cuya distribución es normal con esperanza $E(\tilde{y}_i^b) = \mathbf{x}_i'\boldsymbol{\beta}$ y varianza $Var(\tilde{y}_i^b) = \tilde{V}_i$. Luego, la pseudo-verosimilitud del vector de variables de trabajo es

$$p(\tilde{\mathbf{y}}_b \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}, \boldsymbol{\gamma}^{(c)}, \mathbf{L}^{(c)}) \propto \exp\left\{\frac{-1}{2}(\tilde{\mathbf{y}}_b - \mathbf{X}'\boldsymbol{\beta}^{(c)})'\tilde{\mathbf{V}}^{-1}(\tilde{\mathbf{y}}_b - \mathbf{X}'\boldsymbol{\beta}^{(c)})\right\}$$

Combinando la anterior distribución normal con la distribución previa normal del vector de parámetros $\boldsymbol{\beta}$, se obtiene que una distribución de salto apropiada con kernel gaussiano $q_{\boldsymbol{\beta}}$ dado por (56).

Un enfoque similar se utiliza para la construcción de la propuesta para el vector de parámetros de los efectos aleatorios $\boldsymbol{\lambda}$. Se define la variable de trabajo $\tilde{y}_i^l$ dada por

$$\tilde{y}_i^b := \tilde{y}_i - \mathbf{x}_i'\boldsymbol{\beta} = \mathbf{w}_i'\boldsymbol{\lambda} + h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))(y_i - h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))$$

cuya distribución es normal con esperanza $E(\tilde{y}_i^l) = \mathbf{w}_i'\boldsymbol{\lambda}$ y varianza $Var(\tilde{y}_i^b) = \tilde{V}_i$. Luego, la pseudo-verosimilitud del vector de variables de trabajo es

$$p(\tilde{\mathbf{y}}_l \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}, \boldsymbol{\gamma}^{(c)}, \mathbf{L}^{(c)}) \propto \exp\left\{\frac{-1}{2}(\tilde{\mathbf{y}}_l - \mathbf{W}'\boldsymbol{\lambda}^{(c)})'\tilde{\mathbf{V}}^{-1}(\tilde{\mathbf{y}}_l - \mathbf{W}'\boldsymbol{\lambda}^{(c)})\right\}$$

Combinando la anterior distribución normal con la distribución condicional previa normal del vector de parámetros $\boldsymbol{\lambda}$, se obtiene que una distribución de salto apropiada con kernel gaussiano $q_{\boldsymbol{\lambda}}$ dada por (58). ■

Con el anterior procedimiento es posible modelar la media μ de la distribución al proponer valores para los parámetros de interés desde las distribuciones de salto. Por otro lado, es necesario realizar una metodología similar para la inferencia del vector de parámetros τ. Una vez más sea pelará a la implementación del Fisher-Scoring y a la creación de una variable de trabajo. Como se analizó antes, dado que (53), la distribución condicional posterior de $\boldsymbol{\gamma}$, no tiene una forma cerrada y no es log-concava, se debe aproximar la verosimilitud de los datos a una normal y combinarla con la distribución previa de los parámetros. El siguiente resultado muestra el procedimiento para conseguirlo.

Resultado 14.1.9 (Algoritmo IRLS adaptado)**.** *Asumiendo que existe una variable t_i tal que $E(t_i) = \tau_i = g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma})$, se define la variable de trabajo dada por la aproximación de Taylor de $g(\cdot)$ en el punto $E(t_i)$ como*

$$\tilde{y}_i := \mathbf{z}_i'\boldsymbol{\gamma} + g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))(t_i - g^{-1}(\mathbf{z}'\boldsymbol{\gamma})) \tag{14.1.76}$$

cuya distribución es normal con media $E(\tilde{y}_i) = \mathbf{z}_i'\boldsymbol{\gamma}$ y varianza $\tilde{V}_i = Var(\tilde{y}_i) = [g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))]^2 Var(t_i)$, siendo $\boldsymbol{\beta}^{(c)}$, $\boldsymbol{\lambda}^{(c)}$, $\boldsymbol{\gamma}^{(c)}$ y $\mathbf{L}$ los valores actuales de los parámetros de interés, una distribución de salto apropiada con kernel gaussiano $q_{\boldsymbol{\gamma}}$ se obtiene como

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}^{(c)}) = Normal(\mathbf{g}^*, \mathbf{G}^*) \tag{14.1.77}$$

En donde,

$$\mathbf{G}^* = \left(\mathbf{G}^{-1} + \mathbf{Z}'\tilde{\mathbf{V}}^{-1}\mathbf{Z}\right)^{-1}$$
$$\mathbf{g}^* = \mathbf{G}^*\left(\mathbf{G}^{-1}\mathbf{g} + \mathbf{Z}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{y}}\right)$$

con $\tilde{\mathbf{V}} = diag(\tilde{V}_i)$ e $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$.

Prueba. El algoritmo IRLS requiere la aproximación de la transformación de las observaciones alrededor de los valores actuales de las estimaciones para los parámetros. Luego, suponiendo la existencia de una variable t_i cuya esperanza sea $E(t_i) = \tau_i$, se define la variable de trabajo, acudiendo a la diferenciabilidad

de $g(\cdot)$, como la aproximación de Taylor de primer orden de $g(t_i)$ evaluada en el punto $E(t_i)$, de la siguiente forma

$$\begin{aligned}\tilde{y}_i &:= g(E(t_i)) + g'(E(t_i))(t_i - E(t_i)) \\ &= \mathbf{z}_i'\boldsymbol{\gamma} + g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))(t_i - g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))\end{aligned}$$

Luego, de la aplicación del método de Fisher-Scoring en alguna vecindad de $E(t_i) = \tau_i$, la variable de trabajo $\tilde{y}_i$ tiene distribución normal con media $E(\tilde{y}_i) = \mathbf{z}_i'\boldsymbol{\gamma}$ y varianza $\tilde{V}_i = Var(\tilde{y}_i) = [g'(g^{-1}(\mathbf{z}'\boldsymbol{\gamma}))]^2 Var(t_i)$.

Asumiendo independencia y definiendo a $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_i, \ldots, \tilde{y}_n)'$, se tiene que la pseudo-verosimilitud del vector de variables de trabajo está dada por

$$p(\tilde{\mathbf{y}} \mid \boldsymbol{\beta}^{(c)}, \boldsymbol{\lambda}^{(c)}, \boldsymbol{\gamma}^{(c)}, \mathbf{L}^{(c)}) \propto \exp\left\{ \frac{-1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}'\boldsymbol{\gamma}^{(c)})'\tilde{\mathbf{V}}^{-1}(\tilde{\mathbf{y}} - \mathbf{Z}'\boldsymbol{\gamma}^{(c)}) \right\} \quad (14.1.78)$$

con $\tilde{\mathbf{V}} = diag(\tilde{V}_i)$. Por tanto, combinando la anterior distribución normal con la distribución previa normal del vector de parámetros $\boldsymbol{\gamma}$, se tiene fácilmente que una distribución de salto apropiada con kernel gausiano $q_{\boldsymbol{\gamma}}$ se obtiene como (60). ■

Utilizando los anteriores resultados, se propone el siguiente algoritmo híbrido para la selección de muestras provenientes de las distribuciones condicionales posterior dadas por (50) - (53).

1. Iniciar el contador de iteraciones en $j = 1$.
2. Fijar valores iniciales para la cadena dados por $\boldsymbol{\beta}^{(j-1)}$, $\boldsymbol{\lambda}^{(j-1)}$, $\boldsymbol{\gamma}^{(j-1)}$, $\mathbf{L}^{(j-1)}$.
3. Actualizar $\mathbf{L}$ a un nuevo valor $\mathbf{L}^{(j)}$, generado desde (54).
4. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\beta}}$, generado desde la distribución de salto (56).
5. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\beta}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\beta}}$, de otra manera $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$.
6. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\lambda}}$, generado desde la distribución de salto (58).
7. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\lambda}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\lambda}}$, de otra manera $\boldsymbol{\lambda}^{(j)} = \boldsymbol{\lambda}^{(j-1)}$.
8. Proponer un nuevo valor $\boldsymbol{\phi}_{\boldsymbol{\gamma}}$, generado desde la distribución de salto (60).
9. Calcular la probabilidad de aceptación del movimiento. Si el movimiento es aceptado, entonces $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\phi}_{\boldsymbol{\gamma}}$, de otra manera $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$.
10. Actualizar el contador de la cadena de j a $j + 1$.
11. Volver al paso 2 y repetir el procedimiento hasta que la cadena alcance la convergencia deseada.

Para concluir esta sección de propuesta doctoral, se darán algunos ejemplos de la implementación del anterior algoritmo en variables que pertenecen a la familia exponencial bi-paramétrica y uniparamétrica.

Implementación en un modelo Normal

Considérese el modelo de regresión mixto $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} + \varepsilon_i$ con $i = 1, \ldots, n$ y $\varepsilon_i \sim Normal(0)$. Suponga que $h(\mu_i) = \mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}$. Bajo este contexto, se tiene que la variable de trabajo para el vector de efectos fijos es

$$\tilde{y}_i^b = \mathbf{x}_i'\boldsymbol{\beta} + (y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\lambda}) = y_i - \mathbf{w}_i'\boldsymbol{\lambda}$$

La variable de trabajo para el vector de efectos aleatorios es

$$\tilde{y}_i^l = \mathbf{w}_i'\boldsymbol{\lambda} + (y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\lambda}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$$

Nótese entonces que, bajo este contexto, la distribución de salto (56) coincide con la distribución posterior condicional (51). De la misma manera la distribución de salto (58) coincide con la distribución posterior condicional (52). Por lo tanto, la tasa de aceptación en el algoritmo propuesto será del cien por ciento.

Por otro lado, siendo $\tau_i = \sigma_i^2$ y $g(\sigma_i^2) = \log(\sigma_i^2) = \mathbf{z}_i'\boldsymbol{\gamma}$. Entonces se define $t_i = \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(c)} - \mathbf{w}_i'\boldsymbol{\lambda}^{(c)}\right)^2$ puesto que $E(t_i) = \sigma^2$. La variable de trabajo para el vector de efectos de varianza es

$$\tilde{y}_i = \mathbf{z}_i'\boldsymbol{\gamma} + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\lambda})^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} - 1$$

Teniendo en cuenta que $\tilde{V}_i = Var(y_i) = \sigma_i^2$, entonces se tiene que, bajo este marco de referencia, la metodología propuesta en esta sección es un caso particular de la propuesta presentada en la sección 4.1.

Implementación en un modelo Gamma

Considérese el siguiente modelo mixto en donde $y_i \sim Gamma(a_i, b_i)$. Nótese que $\mu_i = a_i/b_i$ y suponga que se utiliza la función de vínculo canónico[3] $h(\mu_i) = 1/\mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}$. Bajo este contexto, se tiene que la variable de trabajo para el vector de efectos fijos es

$$\tilde{y}_i^b = 2\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda} - y_i(\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\lambda})^2 \tag{14.1.79}$$

La variable de trabajo para el vector de efectos aleatorios es

$$\tilde{y}_i^l = 2\mathbf{w}_i'\boldsymbol{\lambda} + \mathbf{x}_i'\boldsymbol{\beta} - y_i(\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\lambda})^2 \tag{14.1.80}$$

Por otro lado, siendo $\tau_i = a_i$ y $g(a_i) = \log(a_i) = \mathbf{z}_i'\boldsymbol{\gamma}$. Entonces se define $t_i = b_i y_i$ puesto que $E(t_i) = a_i$. Ahora, teniendo en cuenta que

$$b_i = \frac{a_i}{\mu_i} = \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{\mu_i} = \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}{h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} = \exp(\mathbf{z}_i'\boldsymbol{\gamma})(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})$$

[3] (**?**) afirma que dado que los parámetros de la distribución gamma son positivos, entonces se debe asegurar que, utilizando los datos de la respuesta y las covariables, la función de vínculo canónico arroje valores estrictamente positivos. Se sugiere utilizar funciones de vínculo como la logarítmica para forzar resultado estrictamente positivos.

La variable de trabajo para el vector de efectos de varianza es

$$\begin{aligned}\tilde{y}_i &= \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}(b_i y_i - \exp(\mathbf{z}_i'\boldsymbol{\gamma})) \\ &= \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}(\exp(\mathbf{z}_i'\boldsymbol{\gamma})(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})y_i - \exp(\mathbf{z}_i'\boldsymbol{\gamma})) \\ &= \mathbf{z}_i'\boldsymbol{\gamma} + (\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})y_i - 1 \end{aligned} \tag{14.1.81}$$

Por lo tanto, notando que $a_i = \mu_i b_i = \tau_i$, $b_i = a_i/\mu_i = \tau_i/\mu_i$, $Var(y_i) = a_i/b_i^2 = \mu_i^2/\tau_i$, $\mu_i = 1/(\mathbf{x}_i'\boldsymbol{\beta}+\mathbf{w}_i'\boldsymbol{\lambda})$ y $\tau_i = \exp(\mathbf{z}_i'\boldsymbol{\gamma})$, entonces el algoritmo puede ser llevado a cabo mediante la realización de la siguiente expresión

$$\begin{aligned}\tilde{V}_i &= [h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))]^2 Var(y_i) \\ &= (\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})^4 \mu_i^2/\tau_i \\ &= \frac{(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})^2}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})} \end{aligned} \tag{14.1.82}$$

Por lo tanto el modelo queda completamente especificado como

$$\tilde{y}_i \sim Gamma(\tau_i, \tau_i/\mu_i)$$

y es posible utilizar las expresiones (54), (56), (58) y (60) para seleccionar muestras de las distribuciones condicionales posterior de los parámetros de interés.

Implementación en un modelo Beta

Asuma el siguiente modelo mixto en donde $y_i \sim Beta(a_i, b_i)$. Nótese que $\mu_i = a_i/(a_i + b_i)$ y, por la naturaleza de esta distribución, suponga que se utiliza la función de vínculo *logit* para asegurar que el resultado de la inversa de esta función pertenezca al intervalo $(0, 1)$. Por lo tanto, $h(\mu_i) = logit(\mu_i) = \log(\mu_i/(1 - \mu_i))$. Luego, teniendo en cuenta que $h^{-1}(x) = e^x/(1 + e^x)$ y $h'(x) = 1/(x(1 - x))$, se tiene que la variable de trabajo para el vector de efectos fijos es

$$\tilde{y}_i^b = \mathbf{x}_i'\boldsymbol{\beta} + \frac{(1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))^2}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}) - 1 \tag{14.1.83}$$

La variable de trabajo para el vector de efectos aleatorios es

$$\tilde{y}_i^l = \mathbf{w}_i'\boldsymbol{\lambda} + \frac{(1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))^2}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}) - 1 \tag{14.1.84}$$

Por otro lado, siendo $\tau_i = a_i + b_i$ y $g(a_i + b_i) = \log(a_i + b_i) = \mathbf{z}_i'\boldsymbol{\gamma}$. Entonces se define $t_i = \frac{(a_i+b_i)^2}{a_i} y_i$ puesto que $E(t_i) = a_i + b_i = \tau_i$. Ahora, teniendo en cuenta

que

$$\begin{aligned} t_i = \frac{(a_i + b_i)^2}{a_i} y_i &= \frac{\tau_i}{\mu_i} y_i \\ &= \frac{g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma})}{h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} y_i \\ &= \frac{\exp(\mathbf{z}_i'\boldsymbol{\gamma})(1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} \end{aligned}$$

La variable de trabajo para el vector de efectos de varianza es

$$\begin{aligned} \tilde{y}_i &= \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1}{\exp(\mathbf{z}_i'\boldsymbol{\gamma})}(t_i - \exp(\mathbf{z}_i'\boldsymbol{\gamma})) \\ &= \mathbf{z}_i'\boldsymbol{\gamma} + \frac{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} y_i - 1 \end{aligned}$$

Por lo tanto, notando que $a_i = \mu_i\tau_i$, $b_i = \tau_i - \mu_i\tau_i$, $Var(y_i) = a_i b_i/[(a_i + b_i)^2(a_i + b_i + 1)] = \mu_i(1 - \mu_i)/(\tau_i + 1)$, entonces el algoritmo puede ser llevado a cabo mediante la realización de la siguiente expresión

$$\begin{aligned} \tilde{V}_i &= [h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))]^2 Var(y_i) \\ &= \frac{1}{\mu_i(1 - \mu_i)(\tau_i + 1)} \\ &= \frac{(1 + \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))^2}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})(\exp(\mathbf{z}_i'\boldsymbol{\gamma}) + 1)} \end{aligned}$$

Por lo tanto el modelo queda completamente especificado como

$$\tilde{y}_i \sim Beta(\mu_i\tau_i, \tau_i - \mu_i\tau_i)$$

y es posible utilizar las expresiones (54), (56), (58) y (60) para seleccionar muestras de las distribuciones condicionales posterior de los parámetros de interés.

Implementación en un modelo Poisson

Claramente cuando $\tau_i = 0$ para todo $i = 1, \ldots, n$, entonces la variable respuesta hace parte de la familia exponencial uniparamétrica. En esta familia es posible encontrar importantes modelos como el modelo Poisson y el binomial. En este apartado y en el próximo, se motivará la modelación bayesiana por medio del algoritmo propuesto para estos modelos.

Asuma el siguiente modelo mixto en donde $y_i \sim Poisson(a_i)$. Es claro que $\mu_i = a_i$) y suponga que se utiliza la función de vínculo natural $h(\mu_i) = \log(\mu_i) = \log(\mu_i/(1-\mu_i))$. Luego, se tiene que la variable de trabajo para el vector de efectos fijos es

$$\tilde{y}_i^b = \mathbf{x}_i'\boldsymbol{\beta} + \frac{y_i}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} - 1 \tag{14.1.85}$$

La variable de trabajo para el vector de efectos aleatorios es

$$\tilde{y}_i^l = \mathbf{w}_i'\boldsymbol{\lambda} + \frac{y_i}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})} - 1 \tag{14.1.86}$$

Por lo tanto, notando que $Var(y_i) = a_i = \mu_i = h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})$, $b_i = \tau_i - \mu_i\tau_i$, entonces el algoritmo puede ser llevado a cabo mediante la realización de la siguiente expresión

$$\begin{aligned}\tilde{V}_i &= [h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))]^2 Var(y_i) \\ &= \frac{1}{(\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))^2}\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}) = \frac{1}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})}\end{aligned}$$

Por lo tanto el modelo queda completamente especificado como

$$\tilde{y}_i \sim Poisson(\mu_i)$$

Implementación en un modelo Binomial

Cuando la variable respuesta es tal que $y_i \sim Binomial(n_i, \pi_i)$. Sin embargo, se desea modelar la proporción de éxitos $p_i = y_i/n_i$. Por lo tanto, se tiene que $\mu_i = \pi_i$ y se asume que esta media está relacionada con los efectos fijos y aleatorios a través del vínculo natural $h(\mu_i) = logit(\mu_i) = logit(\pi_i)$. Luego, se tiene que la variable de trabajo $\tilde{y}_i^b$ para el vector de efectos fijos está dada por (66), mientras que la variable de trabajo $\tilde{y}_i^l$ para el vector de efectos aleatorios está dada por (67).

De otro lado, acudiendo a que $Var(p_i) = \pi_i(1-\pi_i)/n_i = \mu_i(1-\mu_i)/n_i$, entonces el algoritmo puede ser llevado a cabo mediante la realización de la siguiente expresión

$$\begin{aligned}\tilde{V}_i &= [h'(h^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))]^2 Var(y_i) \\ &= [h'(\mu_i)]^2 \frac{\mu_i(1-\mu_i)}{n_i} \\ &= \frac{1}{n_i\mu_i(1-\mu_i)} \\ &= \frac{(1+\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda}))^2}{n_i\exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\lambda})}\end{aligned}$$

Nota: Por último, cabe resaltar que la metodología presentada en esta sección puede ser vista como una generalización de los trabajos de (**?**), (**?**) y (**?**). De esta manera, cuando τ es cero, el contexto del modelo se encuentra en la familia exponencial uniparamétrica y la metodología utilizada se torna un caso particular en donde hay equivalencia absoluta con el trabajo desarrollado por (**?**). De otra parte, cuando $\tau \neq 0$ y el modelo permite obviar los efectos aleatorios, entonces los parámetros $\sigma_r^2 = 0$ y la metodología sería equivalente a la utilizada por (**?**). Con base en lo anterior y suponiendo que $h(\mu_i) = \mu_i$ y que la verosimilitud de las observaciones es normal, entonces esta metodología coincide con el desarrollo de (**?**).

14.1.9 Aplicación a modelos jerárquicos y multinivel

Los modelos multinivel nacen cono un intento de definición de modelos estadísticos que encajan naturalmente con la naturaleza de datos agrupados en conglomerados o en grupos conocidos. Por ejemplo, en estudios agrícolas, las cosechas están agrupados por parcelas o secciones agrícolas (**???**); en estudios educacionales, los alumnos están agrupados en escuelas (**??**); en estudios de crianza animal, las crías están agrupadas por semental (**?**); en estudios longitudinales se tienen medidas repetidas sobre un mismo individuo (**??**).

En términos de notación, se establece que las observaciones individuales, llamadas *casos* se observan dentro de los conglomerados o grupos, llamados *unidades*. Luego, los casos son para las unidades como los estudiantes a las escuelas o los hogares a las ciudades. De esta manera, siempre y cuando se tenga acceso a información auxiliar en el nivel de las unidades y también de los casos, es posible plantear un modelo que explique el fenómeno de estudio en los casos y al mismo tiempo tener en cuenta un submodelo que explique la presencia de los casos dentro de las unidades.

Teniendo en cuenta lo anterior, en principio, es posible definir a los parámetros del modelo que están directamente ligados a los casos como efectos fijos, mientras que los parámetros que están ligados a las unidades se definen como efectos aleatorios, puesto que son modelados e incluidos en el modelo de las observaciones individuales. De esta manera se tiene un modelo mixto y dado que los parámetros fijos y aleatorios vienen de casos y unidades, se dice que el modelo mixto también es un modelo multinivel.

En esta propuesta doctoral, optamos por la definición de modelo multinivel como una generalización de un modelo jerárquico. Tal como lo afirma (**?**), en un sentido estricto no es necesario que los observaciones individuales estén anidadas en forma natural a las unidades . Por ejemplo, en estudios sociales, es posible modelar el comportamiento de individuos y modelar sus parámetros dependiendo de el año de observación o de la ubicación geográfica a la cuál pertenece. Es claro que ni año ni ubicación geográfica conforman una jerarquía natural para un individuo. De esta manera, luego de definir un modelo multinivel, se dará paso a la aplicación de la metodología expuesta en las anteriores secciones para la inferencia bayesiana de todas la componentes, fijas y aleatorias, del modelo formulado.

Existe muy variada literatura sobre el tratamiento frecuentista y bayesiano de los modelos multinivel. Nosotros referimos al lector a (**?**), (**?**), (**?**) , quienes dan una excelente introducción basada en modelos lineales. (**?**) y (**?**), proveen una discusión acerca de la modelación no lineal bajo el contexto de aplicaciones bioestadísticas. En el caso de modelos lineales generalizados (**?**), (**?**), (**?**) y (**?**) son un buen punto de partida en términos de inferencia estadística y aplicaciones prácticas cuando existe una estructura multinivel.

Relación entre un modelo multinivel y un modelo mixto

En este apartado, se considera un modelo jerárquico con dos niveles. En primer lugar, se considera que existen n individuos agrupados (no necesariamente mediante una jerarquía directa) dentro de J unidades; en cada grupo existen n_j casos, tal que $\sum_{j=1}^{J} n_j = n$, y se supone que para las J unidades, es posible modelar el vector de variables respuesta como

$$\mathbf{y}_j = \mathbf{X}_j'\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \tag{14.1.87}$$

donde cada matriz $\mathbf{X}_j$ tiene dimensiones $n_j \times P$ y $\boldsymbol{\varepsilon}_j$ se distribuye normal con media $\mathbf{0}$ y matriz de covarianzas $\boldsymbol{\Sigma}_j$. En el siguiente nivel, se quiere modelar la variación del vector $\boldsymbol{\beta}_j$ de una unidad a otra. En otras palabras, se quiere un modelo entre unidades que vincula a los casos del modelo anterior. De esta manera, suponiendo que existe información auxiliar $\mathbf{u}_{jp}$ para el p-ésimo parámetro $(p = 1, \ldots, P)$ perteneciente a la j-ésima unidad $(p = 1, \ldots, P)$, entonces se tiene el siguiente modelo

$$\beta_{jk} = \mathbf{u}_{jk}'\boldsymbol{\alpha}_k + \delta_{jk} \tag{14.1.88}$$

donde δ_{ij} tiene distribución normal con media cera y varianza σ_p^2. Nótese que, en forma más general, el anterior modelo puede escribirse de la siguiente manera, llegando a un modelo entre unidades

$$\boldsymbol{\beta}_j = \mathbf{U}_j'\boldsymbol{\alpha} + \boldsymbol{\delta}_j \tag{14.1.89}$$

Donde $\boldsymbol{\delta}_j$ tiene distribución normal con media $\mathbf{0}$ y varianza $\mathbf{L}_j$, además se tiene que

$$\begin{aligned}
\mathbf{U}_j &= \begin{pmatrix} \mathbf{u}_{j1}' & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_{j2}' & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{u}_{jP}' \end{pmatrix} \\
&= \bigoplus_{p=1}^{P} \mathbf{u}_{jp}' \\
\boldsymbol{\alpha} &= (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_P)' \\
\boldsymbol{\delta}_j &= (\delta_{j1}, \ldots, \delta_{jP})'
\end{aligned}$$

Por lo tanto, al combinar (77) y (79), se tiene la siguiente expresión

$$\mathbf{y}_j = \mathbf{X}_j'\mathbf{U}_j'\boldsymbol{\alpha} + \mathbf{X}_j'\boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j \tag{14.1.90}$$

La anterior formulación indica que un modelo multinivel es un caso particular de un modelo mixto puesto que (80) puede escribirse también como

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} + \boldsymbol{\varepsilon} \tag{14.1.91}$$

En donde

$$\begin{aligned}\mathbf{y} &= (\mathbf{y}_1, \ldots, \mathbf{y}_J)' \\ \mathbf{X} &= (\mathbf{X}_1'\mathbf{U}_1', \ldots, \mathbf{X}_J'\mathbf{U}_J')' \\ \mathbf{W} &= \bigoplus_{j=1}^{J} \mathbf{X}_j' \\ \boldsymbol{\lambda} &= (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_J)' \\ \boldsymbol{\varepsilon} &= (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_J)'\end{aligned}$$

Luego,

$$\begin{aligned}\boldsymbol{\Sigma} &= \bigoplus_{j=1}^{J} \boldsymbol{\Sigma}_j \\ \mathbf{L} &= \bigoplus_{j=1}^{J} \mathbf{L}_j\end{aligned}$$

Modelos multinivel generalizados

Los modelos multinivel pueden ser generalizados de la misma forma que (**?**) generalizaron el modelo lineal. Lo anterior permite que el investigador pueda modelar la variable respuesta como binomial, poisson, logístico, entre otras al mismo tiempo que se utilizan los efectos aleatorios para modelar el anidadamiento de las observaciones o sobredispersión. (**?**) introducen algunas técnicas bayesianas para encontrar la densidad posterior de los parámetros de interés mediante la técnica de Gibbs. El anterior acercamiento generaliza el trabajo de (**?**), (**?**), (**?**).

Cuando la variable respuesta pertenece a la familia exponencial bi-paramétrica, un modelo lineal generalizado multinivel se puede escribir de la siguiente manera

$$\begin{aligned}p(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= a(y_{ij}) \exp\left\{d_1(\theta_{ij}, \tau_{ij})T_1(y_{ij}) + d_2(\theta_{ij}, \tau_{ij})T_2(y_{ij}) + b(\theta_{ij}, \tau_{ij})\right\} \\ E(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= \mu_{ij} \\ h(\mu_{ij}) &= \eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_j\end{aligned}$$

Nótese que el vector $\boldsymbol{\eta}$ en la j-ésima unidad está dado por

$$\boldsymbol{\eta}_j = \mathbf{X}_j'\boldsymbol{\beta}_j$$

En el siguiente nivel, se quiere modelar la variación del vector $\boldsymbol{\beta}_j$ de una unidad a otra. Luego, suponiendo que existe información auxiliar $\mathbf{U}_j$ para la j-ésima unidad, entonces se tiene que

$$\boldsymbol{\beta}_j = \mathbf{U}_j'\boldsymbol{\alpha} + \boldsymbol{\delta}_j$$

Donde $\boldsymbol{\delta}_j$ tiene distribución normal con media $\mathbf{0}$ y varianza $\mathbf{L}_j$. Al combinar los modelos en los dos niveles, se tiene la siguiente expresión

$$\boldsymbol{\eta}_j = \mathbf{X}_j'\mathbf{U}_j'\boldsymbol{\alpha} + \mathbf{X}_j'\boldsymbol{\delta}_j$$

La anterior formulación indica que un modelo multinivel generalizado es un caso particular de un modelo mixto generalizado puesto que el modelo global puede escribirse también como

$$\boldsymbol{\eta} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda}$$

En donde las anteriores cantidades se definen análogamente al anterior apartado.

Nota: La formulación tradicional de los modelos multinivel (lineales o generalizados) contempla que la variación de los casos es constante. Sin embargo, si se presenta el caso en donde haya sospecha de que exista heteroscedasticidad dentro de las unidades, es necesario revaluar este supuesto. Es aquí en donde la alternativa que se propone en este trabajo doctoral puede contemplarse como una solución a esta situación.

Ejemplo: modelo lineal de intercepto aleatorio

Por supuesto, la anterior escritura de un modelo lineal mixto como caso general de un modelo multinivel es muy amplia y de allí se desprenden una gran variedad de modelos particulares. Nótese, por ejemplo, que en la formulación del modelo multinivel anterior, todos los efectos se pueden considerar aleatorios puesto que se encuentran anidados por otro modelo. Sin embargo, existen modelos que consideran la estructura andanada para un subconjunto del vector de parámetros del modelo del primer nivel. Un ejemplo claro de este tipo de modelos es el bien conocido modelo de intercepto aleatorio que incluye indicadoras para los grupos y puede ser interpretado como un modelo con un intercepto distinto para cada grupo aunque la pendiente es invariante. Por lo tanto si existen J grupos, el modelo tendrá $J+1$ parámetros para estimar. La forma funcional de modelo se da a continuación:

$$y_{ij} \sim Normal(\beta_{0j} + \beta_1 x_{ij}, \sigma_y^2), \qquad i = 1, \ldots, n_j \tag{14.1.92}$$

$$\beta_{0j} \sim Normal(\alpha_0 + \alpha_1 u_j, \sigma_\beta^2), \qquad j = 1, \ldots, J \tag{14.1.93}$$

El anterior modelo puede ser escrito en forma de modelo mixto de la siguiente manera

$$y_{ij} = \beta_1 x_{ij} + \alpha_0 + \alpha_1 u_j + \delta_j + \varepsilon_{ij} \tag{14.1.94}$$

En donde ε_{ij} tiene distribución normal con media nula y varianza σ_y^2 y δ_j tiene distribución normal con media nula y varianza σ_β^2. en términos vectoriales, es posible escribir el mismo modelo como

$$\mathbf{y}_j = \mathbf{X}_j'\boldsymbol{\beta} + \mathbf{W}_j'\boldsymbol{\lambda}_j + \boldsymbol{\varepsilon}_j$$

En donde

$$
\begin{aligned}
\mathbf{y}_j &= (y_{1j}, \ldots, y_{n_j j})' \\
\boldsymbol{\beta} &= (\beta_1, \alpha_0, \alpha_1)' \\
\mathbf{W}_j &= \mathbf{1}'_{n_j} \\
\boldsymbol{\lambda}_j &= \delta_j \\
\mathbf{X}'_j &= \begin{pmatrix} x_{1j} & 1 & u_j \\ x_{2j} & 1 & u_j \\ \vdots & \vdots & \vdots \\ x_{n_j j} & 1 & u_j \end{pmatrix}
\end{aligned}
$$

Para este caso particular, (**?**) abordan un ejemplo en donde la variable respuesta y_{ij} es el nivel del gas radon en una vivienda i de un determinado condado j. A su vez, ésta es explicada por la variable de información x_{ij} que es una variable indicadora que toma el valor 1 si la vivienda tiene dos pisos. El intercepto de este modelo de primer nivel es explicado por una variable de información auxiliar u_j describiendo el nivel de uranio en el condado j.

Algoritmo propuesto para modelos multinivel

Para el caso particular del modelo multinivel con intercepto aleatorio, $\boldsymbol{\beta}$ es un vector de coeficientes de regresión de tamaño 3, $\boldsymbol{\lambda}$ resulta ser un vector de efectos aleatorios de J componentes tal que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)'$ y se supone que la variación entre los efectos es idéntica. La asignación de las distribuciones previa para los parámetros de interés resulta ser de la siguiente forma

$$
\begin{aligned}
\boldsymbol{\beta} &\sim Normal(\mathbf{b}, \mathbf{B}) \\
\boldsymbol{\lambda} \mid \sigma^2_\lambda &\sim Normal(\mathbf{0}, \sigma^2_\lambda \mathbf{I}) \\
\sigma^2_\lambda &\sim Inversa - Gama(a, h)
\end{aligned}
$$

Nótese que

$$
\mathbf{b} = \begin{pmatrix} b \\ a_0 \\ a_1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} B & 0 & 0 \\ 0 & A_0 & 0 \\ 0 & 0 & A_1 \end{pmatrix}
$$

Por supuesto que, para implementar nuestra propuesta, se supone que existe heteroscedasticidad en la respuesta y que es necesario modelar la varianza de las observaciones. Utilizando el enfoque propuesto, se supone que

$$
g(\sigma^2_i) = \mathbf{z}'_i \boldsymbol{\gamma}
$$

De esta manera, adecuando el resultado 4.1, las distribuciones condicionales

posterior de los parámetros de interés $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ y σ^2_λ están dadas por

$$\begin{aligned}
&\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma^2_\lambda, \mathbf{y} \sim Normal(\mathbf{b}^*, \mathbf{B}^*) \\
&\boldsymbol{\lambda} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2_\lambda, \mathbf{y} \sim Normal(\mathbf{l}^*, \mathbf{L}^*) \\
&\sigma^2_\lambda \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y} \sim Inversa - Gama\left(-1/2 - a, \sum_{j=1}^{J} \lambda_j^2/2 + h\right)
\end{aligned}$$

En donde las cantidades anteriores son equivalentes a las del resultado 4.1. Con base en lo expuesto anteriormente, se utiliza el algoritmo modificado de Metropolis-Hastings que permite la simulación de valores provenientes de las distribuciones posterior involucradas en el análisis de los datos.

Una vez se tenga la estimación de los parámetros de interés $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ y σ^2_λ entonces el modelo multinivel queda totalmente identificado y es posible realizar todo tipo de inferencias en ambos niveles. Además de los anterior, utilizando las distribuciones del vector de variables respuesta y el vector de efectos aleatorios también es posible acceder a nuevas predicciones para observaciones que no fueron incluidas en el análisis previo. Luego, siguiendo el enfoque de (**?**, pp. 272 - 275) fácilmente se obtienen predicciones para grupos existentes o para nuevos grupos.

Por otro lado, cuando la variable respuesta no tiene naturaleza gaussiana, se utiliza el acercamiento del modelo lineal generalizado anidado y con base en la reparametrización que se tuvo al principio de esta sección, se obtiene un modelo mixto generalizado. De esta forma, recurriendo a la propuesta de la sección 4.3, se tiene que los parámetros de la media μ_i y de τ_i quedan completamente especificados y es posible realizar inferencia en ambos niveles. Por ejemplo, considere el siguiente escenario

$$\begin{aligned}
&y_{ij} \sim Bernoulli(\pi_{ij}) \\
&\pi_{ij} = logit^{-1}(\beta_{0j} + \beta_1 x_{ij}) \\
&\beta_{0j} \sim Normal(\alpha_1 u_j, \sigma^2_\beta), \quad j = 1, \ldots, J
\end{aligned}$$

El anterior modelo puede ser escrito en forma de modelo mixto generalizado así:

$$\eta_{ij} = \beta_1 x_{ij} + \alpha_1 u_j + \delta_j$$

En donde δ_j tiene distribución normal con media nula y varianza σ^2_β. En términos vectoriales, es posible escribir el mismo modelo como

$$\boldsymbol{\eta}_j = \mathbf{X}'_j \boldsymbol{\beta} + \mathbf{W}'_j \boldsymbol{\lambda}_j$$

Es fácil definir cada una de las anteriores cantidades convenientemente, de tal forma que los resultados propuestos en este trabajo doctoral en las secciones anteriores puedan ser utilizadas. Bajo este contexto de modelos lineales generalizados multinivel, se debe ser más cuidadoso en términos predictivos puesto que se debe tener en cuenta el carácter secuencial y anidado del modelo.

14.1.10 Aplicación a modelos de predicción en áreas pequeñas

Las técnicas de estimación en áreas pequeñas son una parte muy importante de estudios por muestreo debido a la creciente demanda pública y privada de estadísticas confiables acerca del parámetro de interés en subpoblaciones específicas. Las técnicas de muestreo e inferencia en poblaciones finitas que utilizan estimadores directos, como el estimador de Horvitz-Thompson, Hansen-Hurvitz, entre otros (**?**), tienen la particularidad de arrojar altos e inaceptables errores estándar; aún más, si la muestra seleccionada no contiene algún elemento perteneciente al dominio de interés, entonces no es posible contar con tal estimación.

Con base en lo anterior, la inferencia en poblaciones finitas basada en el diseño muestral no alcanza a resolver los problemas de estimación en este tipo de situaciones. Por lo tanto, es necesario recurrir a técnicas de inferencia basadas en modelos poblacionales con el fin de obtener estimaciones satisfactorias en cada dominio de interés, independientemente de que se haya seleccionado o no elementos en todos los dominios de interés.

El término área pequeña es utilizado para denotar un área geográfica determinada. Por ejemplo, un área pequeña puede ser un condado, una municipalidad, un departamento o cualquier división censal. Sin embargo esta definición no está superditada únicamente a la acepción geográfica. Un dominio o un área pequeña puede estar relacionado con una subpoblación pequeña conformada por elementos dentro de un grupo específico como edad-género-nivel socioeconómico. En esta propuesta doctoral se utilizará el término área pequeña indistintamente de las anteriores consideraciones tal como lo sugiere (**?**).

La referencia principal de la inferencia frecuentista y bayesiana en problemas de áreas pequeñas es el trabajo reciente de (**?**). Otras referencias de importancia en el tema son (**?**), (**?**), (**?**), (**?**), (**?**), (**?**) y (**?**).

En esta sección, se utilizarán las propuestas de inferencia para modelos mixtos en el contexto de estimación de áreas pequeñas en dos niveles: ajustando un modelo poblacional general a los elementos pertenecientes a los dominios de interés y ajustando un modelo a las estimaciones de los parámetros de interés (totales, promedios o razones) en cada dominio. A lo largo de este apartado se justificará que estos modelos son un caso particular del modelo lineal mixto. Además, es natural considerar que la variación de las variables aleatorias que constituyen la población finita no sea constante y por lo tanto se quiera modelar.

Teniendo en cuenta estas razones, es plausible utilizar la propuesta inferencial bayesiana como un camino intermedio en la estimación de los parámetros de interés en cada uno de los dominios. Vale aclarar que bajo este contexto el objetivo principal está relacionado con la predicción de las unidades no seleccionadas en la muestra. Por lo tanto, el enfoque primordial estará basado en la adecuación de algoritmos orientados a este fin.

Estimadores directos y sintéticos

Bajo la teoría de inferencia en poblaciones finitas, se considera un parámetro de interés a una función conocida de los valores que toma la variable respuesta en cada uno de los elementos de una población $U = \{1, \ldots, k, \ldots, N\}$. Siendo y_k el valor de la variable respuesta en el individuo i, algunos ejemplos de parámetros de interés son el total poblacional dado por

$$Y = \sum_{k \in U} y_k$$

el promedio poblacional dado por

$$\bar{Y} = \frac{Y}{N}$$

entre otros. Suponga que la población finita está particionada en J subgrupos poblacionales llamados dominios U_j tales que

$$U = \bigcup_{j=1}^{J} U_j \qquad U_j \bigcap U_{j'} = \emptyset$$

Es de interés para el investigador obtener estimaciones de los parámetros de interés (totales, promedios, razones) en cada uno de los dominios. La teoría de muestreo provee estimadores de estas cantidades con el supuesto de que al menos un elemento de cada uno de los dominios fue seleccionado en una muestra seleccionada s. El lector puede referirse a los textos de (**?**), (**?**) y (**?**) para una introducción al tema. Además de lo anterior, se denota por $\hat{Y}_j$ el estimador del total en el dominio U_j dado por Y_j y se supone que este estimador es insesgado con respecto al diseño de muestreo p utilizado en la selección de s; en otras palabras $E_p(\hat{Y}_j) = Y_j$ para todo $j = 1, \ldots, J$. Cuando el tamaño de muestra n_j en el dominio U_j es muy pequeño, los estimadores directos pueden arrojar resultados incoherentes influenciados por una gran variación.

Una solución al anterior inconveniente está dada por los estimadores sintéticos que, tal como lo define (**?**), son usados para derivar estimativos de los parámetros de interés en los subgrupos poblacionales bajo el supuesto de que estos dominios se comportan de igual forma que la población finita. De esta manera, y en particular suponiendo que se quieran estimar los totales en los subgrupos, se asume que existe un estimador directo $\hat{Y}$ para el total poblacional Y que puede ser calculado mediante los datos obtenidos en la muestra s y la estimación tendrá un error estándar pequeño puesto que emplea toda la información recolectada en el estudio. Al mismo tiempo, se asume que existe información auxiliar de los totales en los dominios para alguna característica de información auxiliar x. Luego $X_j = \sum_{k \in U_j} x_k$ se asume conocida para todo U_j. Un estimador sintético para Y_j está dado por

$$\hat{Y}_j = \frac{X_j}{X} \hat{Y}$$

donde X es el total poblacional de la característica de la información auxiliar. Nótese que $\sum_{j=1}^{J} \hat{Y}_j = \hat{Y}$ y además que incluso si la muestra seleccionada s no contiene elementos de un dominio determinado, siempre es posible calcular una estimación para tal dominio. En el caso particular en que $X_j = N_j$, el tamaño del dominio U_j, este estimador asumiría que la media global $\bar{Y}$ es equivalente a las medias $\bar{Y}_j$. Es claro que el anterior supuesto es muy fuerte y por tanto este tipo de estimadores pueden ser ampliamente sesgados para algunos subgrupos de interés.

En resumen, los estimadores directos y sintéticos carecen de consistencia y es necesario mejorarlos mediante la inclusión de un modelo que palee los anteriores inconvenientes. Estos modelos pueden estar dados en dos vías, la primera referente al nivel de dominios y la segunda referente a los elementos de la población U.

Modelos tipo A

Como lo afirma (**?**, p. 75), los métodos tradicionales de estimación están basados en modelos implícitos que vinculan los dominios con información auxiliar disponible. El enfoque que se adaptará en adelante hace explícitos los modelos en los dominios de interés y su variación específica entre diferentes dominios. Los modelos tipo A hacen referencia a modelos especificados para la información de un dominio per se; es decir, totales o medias en cada dominio. En primer lugar, se asume que $\theta_j = v(\bar{Y}_j)$ hace referencia al parámetro poblacional de interés en el dominio U_j; por ejemplo, si el parámetro de interés es el promedio, entones v será la función idéntica. Si el parámetro de interés es el total, entonces $v(\bar{Y}_j) = N_j \bar{Y}_j$. Asuma que θ_j está relacionado con un vector de variables auxiliares $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})'$ para el dominio U_j a través de un modelo lineal

$$\theta_j = \mathbf{x}_j' \boldsymbol{\beta} + u_j \delta_j \tag{14.1.95}$$

donde los u_j son constantes positivas conocidas, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ es un vector de coeficientes de regresión, δ_j son efectos aleatorios especificados para cada subgrupo poblacional U_j que se asumen independientes e idénticamente distribuidos con esperanza $E_m(\delta_j) = 0$ y varianza $Var_m(\delta_j) = \sigma_\delta^2 \geq 0$. El sufijo m denota que la esperanza o la varianza se toma al respecto del modelo. El parámetro σ_δ^2 es una medida de homogeneidad de los dominios una vez se ha contemplado el efecto de las covariables en $\mathbf{x}_j$. (**?**) afirma que en algunos estudios no es posible seleccionar a todos los dominios en la muestra seleccionada, sin embargo, se asume que el modelo (85) obedece a un modelo poblacional para todo $j = 1, \ldots, J$ y que los dominios seleccionados en la muestra siguen esa estructura. En algunas palabras, el sesgo de selección de los dominios es nulo.

Para realizar inferencias acerca del parámetro de interés θ_j, es necesario asumir que existe un estimador directo $\hat{\theta}_j$ que se comporta de la siguiente manera:

$$\hat{\theta}_j = \theta_j + e_j \tag{14.1.96}$$

donde e_j representan los errores de muestreo, distribuidos normalmente con

$E_p(e_j|\theta_j) = 0$ y $Var_p(e_j|\theta_j) = \sigma_j^2$. Combinando (85) y (86) se tiene que

$$\hat{\theta}_j = \mathbf{x}_j'\boldsymbol{\beta} + u_j\delta_j + e_j \tag{14.1.97}$$

Nótese que si asumimos independencia entre e_j y δ_j, la anterior formulación puede ser vista como un caso especial de un modelo lineal mixto, puesto que

$$\hat{\boldsymbol{\theta}} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{W}'\boldsymbol{\lambda} + \mathbf{e} \tag{14.1.98}$$

con $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$, $\mathbf{W} = diag(u_1, \ldots, u_J)$, $\boldsymbol{\lambda} = (\delta_1, \ldots, \delta_J)'$ y $\mathbf{e} = (e_1, \ldots, e_J)'$. Si además suponemos que σ_j^2 puede ser modelada de la siguiente manera

$$g(\sigma_j^2) = \mathbf{z}_j'\boldsymbol{\gamma}$$

entonces el modelo se convierte en un caso particular de la propuesta para modelos lineales mixtos tal como se expuso en la sección 4.1. Nótese que bajo el contexto de estimación en áreas pequeñas es muy justificable asignar una variación distinta en cada dominio. De esta forma, $\mathbf{z}_j$ debe corresponder a un vector de variables auxiliares relacionadas directamente con el dominio U_j.

Cabe aclarar que, para llevar a cabo procedimientos de estimación de los parámetros $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ y σ_λ^2, es necesario recurrir al algoritmo propuesto en la sección 4.1, siendo este caso una extensión directa de esta propuesta doctoral. Finalmente, la riqueza de este método se encuentra en la implementación predictiva para cada parámetro θ_j de los dominios de interés. De esta manera, un estimador natural para estos parámetros nace de la esperanza condicional de la variable respuesta, que en este caso particular son los estimadores directos en cada dominio, dado los parámetros del modelo como se indica a continuación.

$$\begin{aligned}\hat{\theta}_j^* &= E(\hat{\theta}_j|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\lambda^2) \\ &= \mathbf{x}_j'\boldsymbol{\beta} + u_j\delta_j\end{aligned} \tag{14.1.99}$$

Para obtener una medida de la dispersión de las estimaciones se utilizan los valores generados por el algoritmo de la sección 4.1 para cada parámetro en cada iteración, y con estos valores se crea una nueva cadena cuya distribución estacionaria esté dada por una distribución normal con media $\mathbf{x}_j'\boldsymbol{\beta} + u_j\delta_j$ y varianza $\sigma_j^2 = g^{-1}(\mathbf{z}_j'\boldsymbol{\gamma})$. Luego, la medida de dispersión será la variación de la cadena.

Ahora suponga que la muestra seleccionada s no contiene ningún elemento para algún dominio $U_{j'}$. Una consecuencia directa de lo anterior es que no existe una estimación directa $\hat{\theta_{j'}}$ y por lo tanto, se debe predecir su valor usando (87). Para esto debemos generar un nuevo término de error $\delta_{j'}$ el cual proviene de la distribución común $Normal(0, \sigma_\delta^2)$, donde σ_δ^2 fue estimada usando el procedimiento descrito en la sección 4.1 con las observaciones de los dominios que fueron seleccionadas. Utilizando la información auxiliar del dominio en cuestión dada por $\mathbf{x}_{j'}$, $u_{j'}$ y $\mathbf{z}_{j'}$ y recurriendo a los valores estimados de $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, y $\delta_{j'}$ entonces podemos simular valores para $\hat{\theta}_{j'}$. Luego, para cada $\delta_{j'}$ generado, se simulan valores de $\hat{\theta}_{j'}$. Al hacer esto repetidas veces, se garantiza que la incertidumbre del dominio $U_{j'}$ se incorpore dentro de la incertidumbre del modelo.

Modelos tipo B

A diferencia de los modelos tipo A, este tipo de modelos tienen en cuenta la información individual de los elementos pertenecientes a la población U y por lo tanto la información auxiliar está directamente relacionada con estos individuos, mas no con el dominio como ocurre en los modelos tipo A. De esta forma, para el individuo i-ésimo perteneciente al j-ésimo dominio $(i = 1, \ldots, N_j)$, $(j = 1, \ldots, J)$ se supone conocido un vector de información auxiliar $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijP})'$ el cual se relaciona con la variable respuesta y_{ij} a través de un modelo de regresión dado por

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \delta_j + e_{ij} \tag{14.1.100}$$

con δ_j denotando el efecto aleatorio del dominio U_j cuya distribución es $Normal(0, \sigma^2_\delta)$ y e_{ij} son variables aleatorias independientes y normalmente distribuidas con media $E_m(e_{ij}) = 0$ y $Var_m(e_{ij}) = \sigma^2_{ij}$. Asumiendo que el diseño de muestreo garantiza que para cada dominio U_j de tamaño N_j exista una submuestra s_j de tamaño n_j y que los valores muestrales obedezcan al modelo (90). Nótese entonces que para la población finita (90) puede ser reescrito como

$$\mathbf{y}_j = \mathbf{X}'_j\boldsymbol{\beta} + \delta_j\mathbf{1}_j + \mathbf{e}_j \tag{14.1.101}$$

$$= \begin{pmatrix} \mathbf{y}^s_j \\ \mathbf{y}^r_j \end{pmatrix} = \begin{pmatrix} \mathbf{X}^s_j \\ \mathbf{X}^r_j \end{pmatrix}\boldsymbol{\beta} + \delta_j \begin{pmatrix} \mathbf{1}^s_j \\ \mathbf{1}^r_j \end{pmatrix} + \begin{pmatrix} \mathbf{e}^s_j \\ \mathbf{e}^r_j \end{pmatrix} \tag{14.1.102}$$

donde el superíndice s denota las unidades que fueron seleccionadas en la muestra y r denota las unidades que no fueron seleccionadas en la muestra. Cabe aclarar que el modelo anterior es válido siempre y cuando el diseño de muestreo se pueda considerar no informativo y por lo tanto ignorable (**??**). Por ejemplo, el modelo (90) no es apropiado si el diseño de muestreo utilizado es por conglomerados o en varias etapas puesto que el efecto de los conglomerados no está incorporado. Por otra parte, nótese que si el parámetro de interés es el total Y_j, entonces

$$Y_j = \sum_{k \in U_j} y_k = \sum_{k \in s_j} y_k + \sum_{k \in r_j} y_k$$

donde s_j denota el conjunto de elementos de U_j que fueron seleccionados en la muestra y $r_j = U_j - s_j$. Por lo tanto, el problema de estimar el total Y_j se convierte en un problema de predicción de y_k para todo $k \in r_j$. De la misma manera, si el parámetro de interés es la media $\bar{Y}_j$, entonces

$$\bar{Y}_j = f_j\bar{y}_{s_j} + (1 - f_j)\bar{y}_{r_j}$$

En este caso, la estimación del promedio de la variable respuesta en el dominio U_j es equivalente a la predicción de $\bar{y}_{r_j}$ dados los datos de la muestra $\mathbf{y}^s_j$ y el vector de variables auxiliares $\mathbf{X}_j = (\mathbf{X}^s_j, \mathbf{X}^r_j)'$. Nótese que el modelo (92) es un caso particular de un modelo lineal mixto de la forma $\mathbf{y}_j = \mathbf{x}'_j\boldsymbol{\beta} + \mathbf{W}'_j\boldsymbol{\lambda}_j + \mathbf{e}_j$. Si las varianzas $Var_m(e_{ij}) = \sigma^2_{ij}$ se puede modelar por medio de variables $\mathbf{z}_{ij}$ de la forma $g(\sigma^2_{ij}) = \mathbf{z}'_{ij}\boldsymbol{\gamma}$, entonces para los elementos del dominio U_j que no fueron

seleccionados en la muestra, es necesario predecir el valor de la variable respuesta y_{ij}. Para esto es necesario generar un nuevo término de error δ_j mediante una distribución $Normal(0, \sigma_\delta^2)$, donde σ_δ^2 fue estimada usando el procedimiento descrito en la sección 4.1 con las observaciones de los elementos que fueron seleccionados en cada dominio. Utilizando la información auxiliar de los elementos $\mathbf{x}_{ij}$, y $\mathbf{z}_{ij}$ y recurriendo a los valores estimados de $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, y δ_j entonces es posible obtener valores y_{ij} para $i \in r_j$ mediante (90). Luego, para cada δ_j generado, se simula un vector $\mathbf{y}_j^r$. El proceso anterior debe ser repetido una gran cantidad de veces y la predicción de $\mathbf{y}_j^r$ está dada por la esperanza condicional que es estima promediando estos vectores simulados.

Modelos en la familia exponencial bi-paramétrica

Siguiendo el enfoque de (**?**), es posible proponer modelos lineales generalizados con efectos aleatorios concernientes a los dominios de la población U condicionado a μ_{ij}, τ_{ij}, los parámetros de la familia exponencial bi-paramétrica, la variable respuesta y_{ij} se asume independientes con función de densidad de probabilidad perteneciendo a la familia exponencial de la siguiente manera

$$\begin{aligned} p(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= a(y_{ij}) \exp\left\{d_1(\theta_{ij}, \tau_{ij}) T_1(y_{ij}) + d_2(\theta_{ij}, \tau_{ij}) T_2(y_{ij}) + b(\theta_{ij}, \tau_{ij})\right\} \\ E(y_{ij} \mid \theta_{ij}, \tau_{ij}) &= \mu_{ij} \\ h(\mu_{ij}) &= \eta_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta}_j + \delta_j \end{aligned}$$

La anterior formulación incluye distribuciones bien conocidas como la binomial, normal, Poisson, Gamma, entre otras. Así mismo, recurriendo a los resultados encontrados en la sección 4.3, es posible modelar el parámetro τ_{ij} mediante información auxiliar $\mathbf{z}_{ij}$. Por ejemplo, si $\mu_{ij} = logit(\pi_{ij})$, entones π_{ij} puede denotar una proporción asociada a una variable binaria en el i-ésimo barrio de la j-ésima ciudad. De esta forma, volvemos a los resultados de la sección 4.3 para estimar los parámetros concernientes en el modelo, y más aún, predecir los parámetros poblacionales (totales, medias, razones) de interés en cada uno de los dominios de la población.

14.2 Programación y convergencia del modelo lineal simple

En este apéndice se propone una simulación empírica para un modelo determinado con el objetivo de comprobar el comportamiento de la metodología bayesiana en un modelo lineal normal. Suponga entonces que la variable respuesta tiene está regida por la siguiente distribución de probabilidad.

$$Y_i \sim Normal(\mu_i, \sigma_i^2) \qquad i = 1, \ldots, n.$$

Además existe una relación entre las variables auxiliares tanto para la media

como para la varianza dada por las siguientes expresiones.

$$\mu_i = -35 + 0.35x_{i1} - 1.7x_{i2}$$
$$\sigma_i^2 = \exp(-8 + 0.026x_{i1} - 0.4x_{i3})$$

De esta manera, se asume que $\mathbf{x}_i = (1, x_{i1}, x_{i2})'$ y que $\mathbf{z}_i = (1, x_{i1}, x_{i3})'$. La inferencia bayesiana posterior se realiza mediante la implementación del siguiente código computacional escrito en R.

```
> library(mvtnorm)
> n <- 400
> x1 <- runif(n,300,400)
> x2 <- runif(n,10,23)
> x3 <- runif(n,0,10)

> # Verosimilitudes
> mu <- -35+0.35*x1-1.7*x2
> sigma2 <- exp(-8+0.026*x1-0.4*x3)
> Sigma <- diag(sigma2)
> y <- rnorm(n, mu, sqrt(sigma2))
> x <- cbind(1,x1,x2)
> z <- cbind(1,x1,x3)

> # Apriori para beta
> B_pri <- diag(rep(10000,3))
> b_pri <- rep(0,3)
> # Apriori para gama
> G_pri <- diag(rep(10000,3))
> g_pri <- rep(0,3)

> # Posteriori condicional para betas
> B_pos <- solve(solve(B_pri)+t(x)%*%solve(Sigma)%*%x)
> b_pos <- B_pos%*%(solve(B_pri)%*%b_pri+t(x)%*%solve(Sigma)%*%y)

> # Genera valores propuestos de la distribución de salto
> r.proposal <- function(gamas.ini){
+ a.now <- z%*%gamas.ini
+ b.now <- y-x%*%betas.now
+ y.now <- a.now+(b.now^2/exp(a.now))-1
+ G_pos <- solve(solve(G_pri)+(0.5)*t(z)%*%z)
+ g_pos <- G_pos%*%(solve(G_pri)%*%g_pri+(0.5)*t(z)%*%y.now)
+ gamas.prop <- rmvnorm(1,g_pos,G_pos)
+ gamas.prop
+ }

> # Densidad para un nuevo valor en la distribución de salto con parámetros viejos
> # Densidad para un viejo valor en la distribución de salto con parámetros nuevos
> dproposal <- function(gamas.now, gamas.old){
+ a.now <- z%*%gamas.old
+ b.now <- y-x%*%betas.now
+ y.now <- a.now+(b.now^2/exp(a.now))-1
+ G_pos <- solve(solve(G_pri)+(0.5)*t(z)%*%z)
+ g_pos <- G_pos%*%(solve(G_pri)%*%g_pri+(0.5)*t(z)%*%y.now)
+ dmvnorm(gamas.now,g_pos,G_pos)
+ }

> # Calcula la densidad de la distribución posterior conjunta
```

```
> dpost <- function(betas, gamas){
+ fc.y <- t(y-x%*%betas)%*%solve(Sigma)%*%(y-x%*%betas)
+ fc.beta <- t(betas-b_pri)%*%solve(B_pri)%*%(betas-b_pri)
+ fc.gama <- t(gamas-g_pri)%*%solve(G_pri)%*%(gamas-g_pri)
+ dp <- (1/sqrt(det(Sigma)))*exp(-0.5*(fc.y+fc.beta+fc.gama))
+ dp
+ }

> # Algoritmo MH modificado Cepeda(2004)
> nsim=5000
> beta.mcmc <- matrix(NA, nrow=nsim, 3)
> gama.mcmc <- matrix(NA, nrow=nsim, 3)
> gamas.ini <- c(-8,0.026,-0.4)
> ind <-rep(0,nsim)
> for(i in 1:nsim){
+ # Valores a posteirori condicional para beta
+ betas.now <- c(rmvnorm(1,b_pos,B_pos))
+ gamas.now <- c(r.proposal(gamas.ini))
+ # Distribución de salto
+ q1 <- dproposal(gamas.now, gamas.ini)
+ q2 <- dproposal(gamas.ini, gamas.now)
+ p1 <- dpost(betas.now, gamas.now)
+ p2 <- dpost(betas.now, gamas.ini)
+ T.val <- min(1,((p1/p2)*(q1/q2)))
+ u <- runif(1)
+ if (u<= T.val){
+ gamas.ini <- gamas.now
+ ind[i]<-1
+ }
+ beta.mcmc[i,]<-betas.now
+ gama.mcmc[i,]<-gamas.ini
+ }

# Calcula la tasa de aceptación del algoritmo
> sum(ind)/nsim
```

Para 5000 iteraciones del algoritmo, se tuvo una tasa de aceptación cercana al 60 %. En la figura 1 se puede observar la rápida convergencia de las cadenas generadas para los tres parámetros del modelamiento de media y para los tres parámetros del modelamiento de la varianza.

14.3 Programación y convergencia de la propuesta para modelos lineales mixtos

En este apéndice se propone una simulación empírica para un modelo mixto básico con el objetivo de comprobar el buen comportamiento de la metodología bayesiana propuesta en esta investigación para un modelo lineal con efectos fijos y aleatorios. Se asume que la variable respuesta está regida por la siguiente distribución de probabilidad.

$$Y_i \sim Normal(\mu_i + v_i, \sigma_i^2) \qquad i = 1, \ldots, n.$$

Además, se supone que existe una relación entre las variables auxiliares tanto

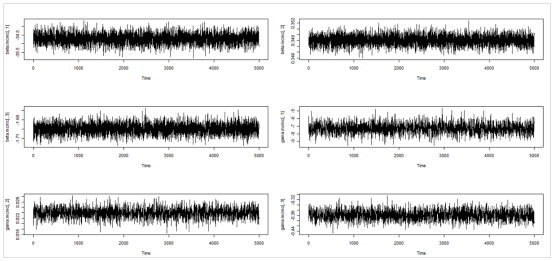

Figura 14.2: *Convergencia de las cadenas para los parámetros del modelamiento conjunto de media y varianza.*

para las medias como para la varianza dada por las siguientes expresiones.

$$
\begin{aligned}
v_i &\sim Normal(\lambda_1 + \lambda_2 * x_{i3}, 0.2) \\
\mu_i &= 30 + 2.5 * x_{i1} \\
\sigma_i^2 &= \exp(-8 + 0.2 * x_{i2})
\end{aligned}
$$

De esta manera, se asume que $\mathbf{x}_i = (1, x_{i1})'$, $\mathbf{z}_i = (1, x_{i3})'$, que $\mathbf{w}_i = (1, x_{i2})'$ y que $\sigma_\lambda = 0.2$. La inferencia bayesiana posterior se realiza mediante la implementación del siguiente código computacional escrito en R.

```
library(mvtnorm)
n <- 100

x1 <- runif(n,10,30)
x2 <- runif(n,10,20)
x3 <- runif(n,20,40)

x <- cbind(1,x1)
z <- cbind(1,x3)
w <- cbind(1,x2)

# Verosimilitudes
mu <- 30+2.5*x1
v <- rnorm(n,0+0*x3,0.2)

sigma2 <- exp(-8+0.2*x2)
Sigma <- diag(sigma2)
```

```
y <- rnorm(n, mu, sqrt(sigma2)) + v

# Apriori para beta
B_pri <- diag(rep(10000,2))
b_pri <- rep(0,2)

# Apriori para gama
G_pri <- diag(rep(0.2,2))
g_pri <- rep(0,2)

# Apriori para lambda
L_pri <- diag(rep(10000,2))
l_pri <- rep(0,2)

# Generates a value for beta from the posterior conditional distribution
rb.proposal <- function(gamas.ini, lambdas.ini){
sigma2.now <- exp(c(w%*%lambdas.ini))
Sigma.now <- diag(sigma2.now)
v.now <- z%*%gamas.ini
B_pos <- solve(solve(B_pri)+t(x)%*%solve(Sigma.now)%*%x)
b_pos <- B_pos%*%(solve(B_pri)%*%b_pri+t(x)%*%solve(Sigma.now)%*%(y-v.now))
# Posteriori condicional para betas
betas.now <- c(rmvnorm(1,b_pos,B_pos))
betas.now
}

# Generates a value for beta from the posterior conditional distribution
rg.proposal <- function(betas.ini, lambdas.ini){
sigma2.now <- exp(c(w%*%lambdas.ini))
Sigma.now <- diag(sigma2.now)
mu.now <- x%*%betas.ini
G_pos <- solve(solve(G_pri)+t(z)%*%solve(Sigma.now)%*%z)
g_pos <- G_pos%*%(solve(G_pri)%*%g_pri+t(z)%*%solve(Sigma.now)%*%(y-mu.now))
# Posteriori condicional para betas
gamas.now <- c(rmvnorm(1,g_pos,G_pos))
gamas.now
}

# Generates a value from the Proposal distribution
rl.proposal <- function(lambdas.ini){
al.now <- w%*%lambdas.ini
bl.now <- y-mu-v
yl.now <- al.now+(bl.now^2/exp(al.now))-1
L_pos <- solve(solve(L_pri)+(0.5)*t(w)%*%w)
l_pos <- L_pos%*%(solve(L_pri)%*%l_pri+(0.5)*t(w)%*%yl.now)
rmvnorm(1,l_pos,L_pos)
}

# Computes the probability of the new value from the old Proposal distribution
# Computes the probability of the old value from the new Proposal distribution
dl.proposal <- function(lambdas.now, lambdas.old){
al.now <- w%*%lambdas.ini
bl.now <- y-(x%*%betas.now)-(z%*%gamas.now)
yl.now <- al.now+(bl.now^2/exp(al.now))-1
L_pos <- solve(solve(L_pri)+(0.5)*t(w)%*%w)
l_pos <- L_pos%*%(solve(L_pri)%*%l_pri+(0.5)*t(w)%*%yl.now)
dmvnorm(lambdas.now,l_pos,L_pos)
}
```

```
# Computes the probability of the posterior distribution
d.post <- function(betas, gamas, lambdas){
fc.y <- t(y-(x%*%betas.now)-(z%*%gamas.now))%*%solve(Sigma)
        %*%(y-(x%*%betas.now)-(z%*%gamas.now))
fc.beta <- t(betas-b_pri)%*%solve(B_pri)%*%(betas-b_pri)
fc.gama <- t(gamas-g_pri)%*%solve(G_pri)%*%(gamas-g_pri)
fc.lambda <- t(lambdas-l_pri)%*%solve(L_pri)%*%(lambdas-l_pri)
dp <- (1/sqrt(det(Sigma)))*exp(-0.5*(fc.y+fc.beta+fc.gama+fc.lambda))
dp
}

# Algoritmo de MH Gutiérrez - Cepeda(2010)
nsim=2000
beta.mcmc <- matrix(NA, nrow=nsim, 2)
gama.mcmc <- matrix(NA, nrow=nsim, 2)
lambda.mcmc <- matrix(NA, nrow=nsim, 2)
betas.ini <- c(30, 2.5)
gamas.ini <- c(6, 4)
lambdas.ini <- c(-8, 0.2)
ind <-rep(0,nsim)

for(i in 1:nsim){
# Valores a posteirori condicional para beta
betas.now <- c(rb.proposal(gamas.ini, lambdas.ini))
gamas.now <- c(rg.proposal(betas.ini, lambdas.ini))
lambdas.now <- c(rl.proposal(lambdas.ini))
# Jumping distribution para MH
q1.l <- dl.proposal(lambdas.now, lambdas.ini)
q2.l <- dl.proposal(lambdas.ini, lambdas.now)
p1 <- d.post(betas.now, gamas.now, lambdas.now)
p2 <- d.post(betas.now, gamas.now, lambdas.ini)
Tl.val <- min(1,((p1/p2)*(q1.l/q2.l)))
u <- runif(1)
if (u<= Tl.val){
lambdas.ini <- lambdas.now
ind[i]<-1
}

beta.mcmc[i,]<-betas.now
gama.mcmc[i,]<-gamas.now
lambda.mcmc[i,]<-lambdas.ini
}

sum(ind)/nsim
0.6565
```

En la figura 2 se puede observar la rápida convergencia de las cadenas generadas para los parámetros de la media y una aceptable tasa de aceptación para el algoritmo modificado de Metropolis-Hastings propuesto para los parámetros de la media. Nótese que estas cadenas generadas no han pasado por la etapas de *burning* ni *thining*. La figura 3 muestra la convergencia de las cadenas para los parámetros del modelamiento conjunto de media y varianza, mientras que la figura 4 describe el histograma de las distribuciones posterior condicionales para estos parámetros. En términos de estimación, las cadenas generadas permiten obtener estimaciones crudas que en este caso particular son muy cercanas a los verdaderos parámetros

de medias y varianza

```
> colMeans(beta.mcmc)
[1] 30.037374  2.498153
> colMeans(gama.mcmc)
[1]  0.046623198 -0.001548626
> colMeans(lambda.mcmc)
[1] -7.4989295  0.1969167
```

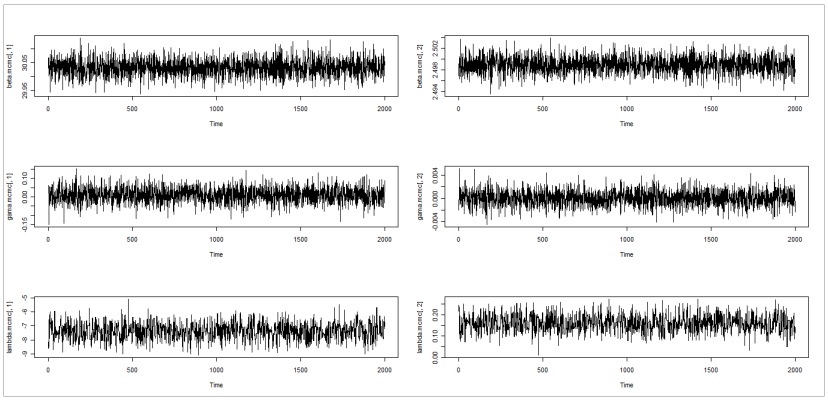

Figura 14.3: *Convergencia de las cadenas para los parámetros del modelamiento conjunto de media y varianza.*

Según (**?**, cap. 4), el enfoque clásico para los modelos lineales mixtos está basado en la teoría de la predicción; de esta manera, el mejor predictor lineal insesgado para los parámetros de la media fijos y aleatorios está dado por.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{14.3.1}$$

$$\hat{\boldsymbol{\gamma}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{14.3.2}$$

Donde $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \boldsymbol{\Sigma}$ y $\mathbf{G}$, la matriz de covarianzas de los parámetros de efectos aleatorios. En el siguiente código, se observa que la variación de los efectos aleatorios afecta las predicciones finales del intercepto fijo. De una manera empírica, es posible corroborar que el enfoque propuesto arroja mejores resultados para este ejercicio particular, además de considerar y modelar la heteroscedasticidad en la variable respuesta.

```
> G <- diag(0.2,nrow=2)
> R <- Sigma
```

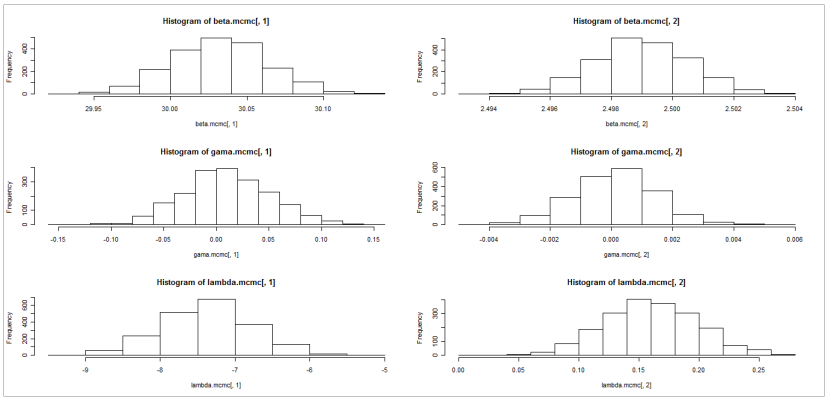

Figura 14.4: *Histograma de las distribuciones posterior condicionales para para los parámetros del modelamiento conjunto de media y varianza.*

```
> V <- z%*%G%*%t(z)+R

> b.est <- solve(t(x)%*%solve(V)%*%x)%*%(t(x)%*%solve(V)%*%y)
> b.est
[1,] 36.047737      [2,] 2.499836
> u.est <- G%*%t(z)%*%solve(V)%*%(y-mu)
> u.est
[1,] 0.998864       [2,] 0.399828
```

15 Tópicos avanzados

15.1 Modelos no-paramétricos

15.2 Ensayos clínicos

15.3 Geoestadística

Parte I

Apéndices

A Introducción a R

B Introducción a WINBUGS

C Defensa del enfoque bayesiano

En la página web del autor del éxito en ventas ≪Bayesian Data Analysis≫ se encuentra un punto de vista acerca de la inferencia realizada por los estadístico Bayesianos.

La inferencia Bayesiana es una teoría matemática coherente pero no brinda la suficiente confianza en usos científicos. Las distribuciones previa subjetivas no inspiran confianza porque ni siquiera existe algún principio objetivo para elegir una a distribución previa no informativa (incluso si ese concepto estuviera definido matemáticamente, pues no lo está). ¿De dónde vienen las distribuciones previa? No confío en ellas y no veo ninguna razón para recomendarlas a otra gente, apenas me siento cómodo acerca de su coherencia filosófica.

La teoría Bayesiana requiere un pensamiento mucho más profundo sobre la situación y recomendar el teorema de Bayes para el uso de los científicos es como darle al hijo del vecino la llave de un F-16. De veras que, yo comenzaría con algo de métodos probados y confiables, y entonces generalizaría la situación utilizando los principios estadísticos y la teoría del minimax, que no dependen de ninguna creencia subjetiva. Especialmente cuando las distribuciones previa que veo en la práctica toman formas conjugadas. ¡Qué coincidencia!

Dejando de lado las preocupaciones matemáticas: Me gustan las estimaciones insesgadas, los intervalos de confianza con un nivel real de cobertura. Pienso que la manera correcta de inferir es acercarse al parámetro tanto como sea posible y desarrollar métodos robustos que trabajen con supuestos mínimos. El acercamiento Bayesiano intenta aproximar el insesgamiento, mientras asume supuestos más y más fuertes. En los viejos tiempos, los métodos Bayesianos por lo menos tenían la virtud de estar matemáticamente limpios. Hoy en día, cualquier inferencia se realiza mediante el uso de las cadenas de Markov mediante métodos de Monte Carlo (MCMC). Lo anterior significa que, no sólo no se pueden evaluar las características estadísticas del método, sino que tampoco se puede asegurar su convergencia.

La gente tiende a creer los resultados que apoyan sus preconceptos y descreen los resultados que los sorprenden, ésta es una forma errada y sesgada de pensar. Pues bien, los métodos Bayesianos animan este modo indisciplinado de pensamiento. Estoy seguro que muchos estadísticos Bayesianos están actuando de buena fe. Sin embargo; al mismo tiempo, están proporcionando estímulo a científicos descuidados y poco éticos por todas partes, porque el investigador queda estancado al momento de escoger una distribución previa.

Y para no pasar a temas más críticos, termino la discusión con lo que los Bayesianos piensan acerca de la recolección de los datos. Los cálculos de la teoría Bayesiana de la decisión guían a la idea de que el muestreo probabilístico y la

asignación aleatoria de tratamientos son ineficaces, de que los mejores diseños y muestras son los deterministas. No tengo ninguna conflictos con estos cálculos matemáticos - el conflicto es más profundo, en los fundamentos filosóficos, en la idea de que el objetivo de la estadística consiste en tomar una decisión óptima. Un estimador Bayesiano es un estimador estadístico que reduce al mínimo el riesgo promedio. Sin embargo, cuando hacemos estadística, no estamos intentando «reducir al mínimo el riesgo promedio», estamos intentando hacer estimación y juzgamiento de hipótesis.

No puedo estar al tanto de lo que están haciendo todos esos Bayesianos hoy en día -desafortunadamente, toda clase de personas están siendo seducidas por las promesas de la inferencia automática con la «magia de MCMC»- pero desearía que todos paráramos de una vez y por todas y empezáramos, de nuevo, a hacer estadística de la forma en que debe ser hecha, volviendo a los viejos tiempos en que un p-valor era utilizado para algo, cuando un intervalo de confianza tenía significado, y el sesgo estadístico era algo que se quería eliminar y no algo que se debiera abrazar.

El autor de este blog, comparte algunas ideas de la anterior disertación. Sin embargo, reconoce la magnitud y el impacto que los Bayesianos han tenido no sólo en el desarrollo de la teoría estadística sino también en el pensamiento estadístico del autor. La estadística Bayesiana debe ser utilizada con expertise. Al ser utilizada por investigadores neófitos puede ser tergiversada. Sin embargo, el mal uso que se le dé a un método no involucra su credibilidad, sino la ignorancia del investigador.

Esos cuestionamientos los hizo Andrew Gelman, uno de los autores Bayesianos más leídos de la última década.

Esos cuestionamientos hicieron que el editor de una importante publicación internacional convenciera a Andrew Gelman para que escribiera un manuscrito al respecto. En esta entrada se dan las respuestas a esos cuestionamientos.

Antes que nada Andrew Gelman hace la siguiente aclaración: La estadística Bayesiana se trata de hacer afirmaciones de probabilidad, mientras que la estadística frecuentista se trata de evaluar afirmaciones de probabilidad... De esta forma un estadístico (entendido como la persona que ejecuta métodos estadísticos) puede ser frecuentista y Bayesiano en diferentes ocasiones. Aún más un simple método de análisis puede ser frecuentista y Bayesiano al mismo tiempo.

«Los métodos Bayesianos son presentados como un motor de inferencia automática»: La inferencia Bayesiana tiene tres etapas: formulación del modelo, ajuste del modelo a los datos, comprobar el ajuste. Así que el procedimiento no es automático.

«Como científicos debemos tratar con el conocimiento objetivo y dejar a un lado las creencias subjetivas»: Las distribuciones previa que maneja la inferencia Bayesiana son objetivas de la misma forma que lo son los métodos frecuentistas. El resultado final sólo depende del modelo asumido y de los datos recolectados.

«Los métodos Bayesianos parecen moverse rápidamente hacia la computación elaborada»: Para bien o para mal, la computación se está convirtiendo en una plataforma central para el desarrollo estadístico.

≪No existe un principio objetivo para la escogencia de una distribución previa no informativa... De todas formas, ¿de dónde vienen esas distribuciones previa?≫: Nótese que tampoco existe un principio objetivo para escoger una verosimilitud... ¿de dónde vienen las regresiones logísticas? ¿quién dijo que los datos eran normales? Bernardo dice que como toda ciencia, la estadística se basa en procedimientos subjetivos que guían a resultados que se pueden probar de una manera objetiva.

≪¿Por qué debería creer en una previa subjetiva?≫: Si hay una seria diferenciación entre las creencias subjetivas y los resultados posterior, debería ser un indicador de reevaluar el modelo usado.

≪Es preferible la inferencia insesgada y los intervalos de confianza que permiten tener un nivel real de cobertura≫: Acerca de este tópico Andrew Gelman recomienda dar un vistazo al ejemplo de las páginas 248-249 de su libro Bayesian Data Analysis.

≪La gente tiende a creer en resultados que apoyan sus preconceptos y son incrédulos ante los resultados que los logran sorprender≫: como Bayesiano uno puede utilizar una distribución previa débil y añadir más información si se necesita.

≪Un estimador Bayesiano es un estimador estadístico que minimiza el riesgo promedio. Sin embargo, cuando utilizamos estadística no tratamos de minimizar el riesgo promedio,; tratamos de hacer estimación y juzgamiento de hipótesis≫: Es cierto, sin embargo, el lector puede referirse al capítulo 22 de Bayesian Data Analysis donde se habla de funciones de perdida y análisis de decisión como herramientas fundamentales en decisión y no en inferencia estadística.

Termino esta entrada reiterando mi punto de vista acerca de la inferencia Bayesiana: ≪Reconozco la magnitud y el impacto que los Bayesianos han tenido no sólo en el desarrollo de la teoría estadística sino también en el pensamiento estadístico del autor. La estadística Bayesiana debe ser utilizada con expertise. cuando es usada por investigadores neófitos puede ser tergiversada. Sin embargo, el mal uso que se le dé a un método no involucra su credibilidad, sino la ignorancia del investigador.≫

D Algunas distribuciones de probabilidad

Antes de repasar las distribuciones de probabilidad, se define los conceptos de **parámetro de distribución** y **espacio paramétrico**. Un parámetro de distribución es aquella cantidad que define la forma funcional de una distribución de probabilidad, es decir, cuando el parámetro cambia de valor, la función de densidad y la función de distribución cambian. Las distribuciones de probabilidad pueden tener más de un parámetro. Cuando una distribución tiene solo un parámetro, éste se denota usualmente por θ, cuando se presenta más de un parámetro, la notación se cambia a $\boldsymbol{\theta}$, representando el vector de parámetros. El espacio paramétrico, Θ, es el conjunto que contiene todos los posibles valores que puede tomar el parámetro o el vector de parámetros. Para distribuciones con un solo parámetro, Θ será un subconjunto de $\mathbb{R}$, mientras que para distribuciones con dos parámetros, Θ será un subconjunto de $\mathbb{R} \times \mathbb{R}$.

D.1 Distribuciones discretas

D.1.1 Distribución uniforme discreta

DefiniciÃ³n D.1.1. *Una variable aleatoria Y tiene distribución uniforme discreta sobre el conjunto $\{1, 2, \cdots, N\}$ si su función de densidad está dada por:*

$$f_Y(y) = \frac{1}{N} I_{\{1,2,\cdots,N\}}(y) \tag{D.1.1}$$

Esta distribución describe situaciones donde los resultados de un experimento aleatorio tienen la misma probabilidad de ocurrencia. Entre los ejemplos de la distribución uniforme discreta en la vida práctica están lanzamiento de una moneda corriente, lanzamiento de un dado corriente, extracción de una urna que contiene bolas enumeradas de 1 a N.

Resultado D.1.1. *Si Y es una variable aleatoria con distribución uniforme discreta sobre el conjunto $\{1, 2, \cdots, N\}$, entonces*

1. $E(Y) = \frac{N+1}{2}$.
2. $Var(Y) = \frac{N^2-1}{12}$.
3. $m_Y(t) = \sum_{i=1}^{N} \frac{e^{ti}}{N}$.

D.1.2 Distribución hipergeométrica

DefiniciÃ³n D.1.2. *Una variable aleatoria Y tiene distribución hipergeométrica con parámetros n, R y N si su función de densidad está dada por:*

$$f_Y(y) = \frac{\binom{R}{y}\binom{N-R}{n-y}}{\binom{N}{n}} I_{\{0,1,\cdots,n\}}(y), \tag{D.1.2}$$

y se nota como $Y \sim Hg(n, R, N)$.

Suponga que en una urna hay N bolas en total, donde R de ellas son del color negro y los $N - R$ son del color blanco, se extrae aleatoriamente n bolas de la urna ($n < N$), entonces la variable "número de bolas negras extraídas"tiene distribución hipergeométrica con parámetros n, R y N. Otro uso de la distribución hipergeométrica es el problema de captura-recaptura (ver ejercicio 1).

Resultado D.1.2. *Si Y es una variable aleatoria con distribución hipergeométrica con parámetros n, R y N, entonces*

1. $E(Y) = \frac{nR}{N}$.
2. $Var(Y) = \frac{nR(N-R)(N-n)}{N^2(N-1)}$.

El anterior resultado no incluye la función generadora de momentos, pues éste no ha resultado ser útil en la teoría relacionada con la distribución hipergeométrica.

D.1.3 Distribución Bernoulli

La distribución Bernoulli debe su nombre al matemático suizo Jacob Bernoulli (1654-1705) que describe el éxito o fracaso de un

DefiniciÃ³n D.1.3. *Una variable aleatoria Y tiene distribución Bernoulli con parámetro $p \in (0, 1)$ si su función de densidad está dada por:*

$$f_Y(y) = p^y(1-p)^{1-y} I_{\{0,1\}}(y), \tag{D.1.3}$$

y se nota como $Y \sim Ber(p)$.

Nota: La distribución Bernoulli es un caso particular de la distribución binomial cuando $n = 1$.

Resultado D.1.3. *Si Y es una variable aleatoria con distribución Bernoulli con parámetro p, entonces*

1. $E(Y) = p$.
2. $Var(Y) = p(1-p)$.
3. $m_Y(t) = pe^t + 1 - p$.

Resultado D.1.4. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes e idénticamente distribuidas con distribución Bernoulli con parámetro p, entonces la variable $\sum_{i=1}^n Y_i$ tiene distribución $Bin(n, p)$.*

Prueba. La demostración radica en el hecho de que la función generadora de momentos caracteriza la distribución probabilística, entonces basta demostrar que la función generadora de momentos de $\sum_{i=1}^{n} X_i$ es la de una distribución $Bin(n, p)$. Tenemos lo siguiente:

$$\begin{aligned} m_{\sum Y_i}(t) = E(e^{\sum tY_i}) &= E(\prod_{i=1}^{n} e^{tY_i}) \\ &= \prod_{i=1}^{n} E(e^{tY_i}) \quad \text{(por independencia)} \\ &= \prod_{i=1}^{n} (pe^t + 1 - p) \quad \text{(definición de } m_{Y_i}(t)) \\ &= (pe^t + 1 - p)^n \end{aligned}$$

■

D.1.4 Distribución binomial

DefiniciÃ³n D.1.4. *Una variable aleatoria Y tiene distribución binomial con los parámetros $n \in \mathbb{N}$ y $p \in (0, 1)$ si su función de densidad está dada por:*

$$f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} I_{\{0,1,\cdots,n\}}(y), \tag{D.1.4}$$

y se nota como $Y \sim Bin(n, p)$.

Una aplicación de esta distribución es cuando tenemos un número n de repeticiones independientes de un experimento donde cada uno tiene dos posibles resultados que se podrían llamarse como éxito o fracaso, donde la probabilidad de obtener el éxito p es constante en cada una de las repeticiones, entonces la variable número de éxitos obtenidos en las n repeticiones tiene distribución $Bin(n, p)$. La distribución binomial tiene dos parámetros: n y p, sin embargo, cuando n es conocido, la distribución dependerá sólo del valor p que sería el único parámetro con espacio paramétrico $\Theta = (0, 1)$.

Resultado D.1.5. *Si Y es una variable aleatoria con distribución binomial con parámetros n y p, entonces*

1. $E(Y) = np$.
2. $Var(Y) = np(1-p)$.
3. $m_Y(t) = (pe^t + 1 - p)^n$.

D.1.5 Distribución Binomial negativa

DefiniciÃ³n D.1.5. *Una variable aleatoria Y tiene distribución Binomial negativa con parámetros (θ, r) si su función de densidad está dada por:*

$$P(y \mid \theta, r) = \frac{\Gamma(r + y_i)}{y_i!\Gamma(r)}\theta^r(1-\theta)^{1-y_i}I_{(0,1,2,\ldots)}(y) \tag{D.1.5}$$

Nota: Esta distribución siempre ha tenido lugar al resolver el problema del número de ensayos necesarios para lograr tantos éxitos. Por supuesto, si r es el número de éxitos necesarios y se conoce que la probabilidad de éxito es θ, entonces la distribución binomial negativa corresponde a un modelo probabilístico, afianzado durante siglos, que permite la resolución de este tipo de situaciones. Sin embargo, es posible asignar al parámetro r valores que sean reales; en este caso no hay ninguna interpretación práctica en el contexto del número de ensayos necesarios para determinados éxitos. Sin embargo, en términos de distribución r es un parámetro más. Esto nos lleva a uno de los verdaderos usos prácticos de esta distribución: la sobredispersión. Dado que la forma funcional de arriba corresponde a una generalización de la función de distribución Poisson, entonces es posible suponer que los datos de conteo vienen de una distribución binomial negativa. Lo anterior tiene sus ventajas puesto que si la media de los datos recolectados no corresponde con la varianza (característica esencial de la Poisson) entonces cualquier modelo que de allí surgiese sería altamente cuestionable. Si lo anterior se presenta es mejor acudir a la distribución binomial negativa dando valores reales al parámetro r.

Resultado D.1.6. *Si Y es una variable aleatoria con distribución binomial-negativa con parámetros (θ, r), entonces*

1. $E(Y) = \frac{r\theta}{1-\theta}$.
2. $Var(Y) = \frac{r\theta}{(1-\theta)^2}$.
3. $m_Y(t) = \left(\frac{1-\theta}{1-\theta e^t}\right)^r$.

D.1.6 Distribución Poisson

La distribución Poisson debe su nombre al francés Siméon-Denis Poisson (1781-1840) quien descubrió esta distribución en el año 1838, cuando la usó para describir el número de ocurrencias de algún evento durante un intervalo de tiempo de longitud dada.

DefiniciÃ³n D.1.6. *Una variable aleatoria Y tiene distribución Poisson con parámetros $\lambda > 0$ si su función de densidad está dada por:*

$$f_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}I_{\{0,1,\cdots\}}(y) \tag{D.1.6}$$

y se nota como $Y \sim P(\lambda)$.

Nótese que la distribución Poisson tiene solo un parámetro $\theta = \lambda$, y el espacio paramétrico es $\Theta = (0, \infty)$.

Resultado D.1.7. *Si Y es una variable aleatoria con distribución Poisson con parámetro λ, entonces*

1. $E(Y) = \lambda$.
2. $Var(Y) = \lambda$.
3. $m_Y(t) = \exp\{\lambda(e^t - 1)\}$.

Resultado D.1.8. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes con distribución $P(\lambda_i)$ para $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n X_i$ tiene distribución $P(\sum_{i=1}^n \lambda_i)$.*

Prueba. Análogo a la demostración del Resultado 1.1.5. ■

D.2 Distribuciones continuas

D.2.1 Distribución Uniforme Continua

DefiniciÃ³n D.2.1. *Una variable aleatoria Y tiene distribución uniforme continua sobre el intervalo $[a, b]$ con $a < b$ si su función de densidad está dada por:*

$$f_Y(y) = \frac{1}{b-a} I_{[a,b]}(y) \tag{D.2.1}$$

Resultado D.2.1. *Si Y es una variable aleatoria con distribución uniforme continua sobre $[a, b]$, entonces*

1. $E(Y) = \frac{a+b}{2}$.
2. $Var(Y) = \frac{(b-a)^2}{12}$.
3. $m_Y(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$.

D.2.2 Distribución Weibull

DefiniciÃ³n D.2.2. *Una variable aleatoria Y tiene distribución uniforme continua sobre los reales positivos si su función de densidad está dada por:*

$$p(Y \mid \theta, \gamma) = \frac{\theta}{\gamma^\theta} y^{\theta-1} exp\left\{-\frac{y^\theta}{\gamma^\theta}\right\} I_{[0,\infty)}(y) \tag{D.2.2}$$

Resultado D.2.2. *Si Y es una variable aleatoria con distribución Weibull, entonces*

1. $E(Y) = \gamma\Gamma\left(1 + \frac{1}{\theta}\right)$.
2. $Var(Y) = \gamma^2\left[\Gamma\left(1 + \frac{2}{\theta}\right) + \Gamma^2\left(1 + \frac{1}{\theta}\right)\right]$.
3. $m_Y(t) = \sum_{n=0}^{\infty} \frac{t^n \gamma^n}{n!} \Gamma\left(1 + \frac{n}{\theta}\right),\ \theta \geq 1$.

D.2.3 Distribución valor-extremo

DefiniciÃ³n D.2.3. *Una variable aleatoria Y tiene distribución valor-extremo si su función de densidad está dada por:*

$$p(y \mid \theta, \lambda) = \theta \exp(\theta y) \exp\{\lambda - \exp(\lambda + \theta y)\} \tag{D.2.3}$$

Resultado D.2.3. *Si Y es una variable aleatoria con distribución valor-extremo, entonces*

1. $E(Y) = -\frac{\lambda}{\theta} - \frac{\epsilon}{\theta}$.
2. $Var(Y) = \frac{\pi^2}{6\theta^2}$.

Donde $\pi \approx 3.1416$ es el número Pi y $\epsilon = 0.5772$ es la constante de Euler.

D.2.4 Distribución Gamma

DefiniciÃ³n D.2.4. *Una variable aleatoria Y tiene distribución Gamma con parámetro de forma $\alpha > 0$ y parámetro de escala $\theta > 0$ si su función de densidad está dada por:*

$$p(\theta \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} I_{(0,\infty)}(\theta). \tag{D.2.4}$$

donde $\Gamma(k) = \int_0^\infty u^{k-1} \exp(-u)\, du$.

La distribución Gamma tiene dos parámetros: α y β, en este caso, el vector de hiper-parámetros es $\boldsymbol{\eta} = (\alpha, \theta)'$ donde el espacio paramétrico está dado por $\boldsymbol{\eta} = (0, \infty) \times (0, \infty)$.

Resultado D.2.4. *Si Y es una variable aleatoria con distribución Gamma con parámetro de forma α y parámetro de escala θ, entonces*

1. $E(Y) = \alpha/\beta$.
2. $Var(Y) = \alpha/\theta^2$.

Resultado D.2.5. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes con distribución Gamma con parámetro de forma α_i y parámetro de escala β para $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n X_i$ tiene distribución Gamma con parámetro de forma $\sum_{i=1}^n \alpha_i$ y parámetro de escala θ.*

Prueba. Análogo a la demostración del Resultado 1.1.5. ■

D.2.5 Distribución Gamma Inversa

DefiniciÃ³n D.2.5. *Una variable aleatoria Y tiene distribución Gamma Inversa con parámetro de forma $\alpha > 0$ y parámetro de escala $\beta > 0$ si su función de densidad está dada por:*

$$p(y \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} e^{-\beta/y} I_{(0,\infty)}(y). \tag{D.2.5}$$

donde $\Gamma(k) = \int_0^\infty u^{k-1}\exp(-u)\ du$.

La distribución Gamma inversa tiene dos parámetros: α y β, en este caso, el vector de hiper-parámetros es $\boldsymbol{\eta} = (\alpha, \beta)'$ donde el espacio paramétrico está dado por $\boldsymbol{\eta} = (0, \infty) \times (0, \infty)$.

Resultado D.2.6. *Si* Y *es una variable aleatoria con distribución Gamma inversa con parámetro de forma* α *y parámetro de escala* β*, entonces*

1. $E(Y) = \beta/(\alpha - 1)$.
2. $Var(Y) = \theta^2/(\alpha - 1)^2(\alpha - 2)$.

Resultado D.2.7. *Si* X *es una variable aleatoria con distribución* $Gamma(\alpha, \beta)$*, entonces* $1/X$ *tiene distribución* $Gamma - inversa(\alpha, 1/\beta)$.

D.2.6 Distribución exponencial

DefiniciÃ³nD.2.6. *Una variable aleatoria* Y *tiene distribución exponencial con parámetro de escala* $\theta > 0$ *si su función de densidad está dada por:*

$$f_Y(y) = \frac{1}{\theta} e^{-y/\theta} I_{(0,\infty)}(y) \tag{D.2.6}$$

Nota: La distribución exponencial es un caso particular de la distribución Gamma cuando el parámetro de forma k toma el valor 1, y usualmente se utiliza para describir la vida útil de un componente eléctrico o el tiempo necesario para la ocurrencia de algún evento.

Resultado D.2.8. *Si* Y *es una variable aleatoria con distribución exponencial con parámetro* θ*, entonces*

1. $E(Y) = \theta$.
2. $Var(Y) = \theta^2$.
3. $m_Y(t) = \frac{1}{1-\theta t}$ *para* $t < 1/\theta$*, y no existe para otros valores de* t.

Resultado D.2.9. *Sea* $Y_1, \cdots, Y_n$ *variables aleatorias independientes e idénticamente distribuidas con distribución exponencial con parámetro de escala* θ*, entonces la variable* $\sum_{i=1}^n X_i$ *tiene distribución Gamma con parámetro de forma* n *y parámetro de escala* θ.

D.2.7 Distribución Beta

DefiniciÃ³nD.2.7. *Una variable aleatoria* Y *tiene distribución Beta con parámetro de forma* $\alpha > 0$ *y parámetro de escala* $\beta > 0$ *si su función de densidad está dada por:*

$$f_Y(y) = \frac{1}{Beta(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} I_{[0,1]}(y). \tag{D.2.7}$$

donde $Beta(\alpha, \beta) = \dfrac{\gamma(\alpha)\gamma(\beta)}{\gamma(\alpha + \beta)}$.

La distribución Beta tiene dos parámetros: α y β, en este caso, el vector de parámetros es $\Theta = (\alpha, \beta)'$ donde el espacio paramétrico está dado por $\Theta = (0, \infty) \times (0, \infty)$. Pero cuando uno de los dos parámetros es fijo, por ejemplo, si θ es fijo, entonces la distribución tendría un sólo parámetro: k.

Resultado D.2.10. *Si Y es una variable aleatoria con distribución Gamma con parámetro de forma k y parámetro de escala θ, entonces*

1. $E(Y) = \frac{\alpha}{\alpha+\beta}$.
2. $Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

D.2.8 Distribución normal

La distribución normal también es llamada la distribución gaussiana, rindiendo homenaje al matemático alemán Carl Friedrich Gauss (1777-1855). La distribución normal es, sin duda, una de las distribuciones más importantes, puesto que una gran parte de la teoría estadística fue desarrollada inicialmente para variables con esta distribución; por el otro lado, gracias al teorema del límite central, muchas distribuciones ajenas a la normal puede ser aproximadas por esta.

DefiniciÃ³nD.2.8. *Una variable aleatoria Y tiene distribución normal con parámetros μ y σ^2 si su función de densidad está dada por:*

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} I_{\mathbb{R}}(y), \tag{D.2.8}$$

donde $\sigma > 0$ y se nota como $Y \sim N(\mu, \sigma^2)$.

La distribución normal tiene dos parámetros, representado como $\Theta = (\mu, \sigma^2)$ y $\Theta = \mathbb{R} \times (0, \infty)$.

Resultado D.2.11. *Si Y es una variable aleatoria con distribución normal con parámetros μ y σ^2, entonces*

1. $E(Y) = \mu$.
2. $Var(Y) = \sigma^2$.
3. $m_Y(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$.

Nota: Cuando $\mu = 0$ y $\sigma = 1$, se dice que Y tiene distribución normal estándar y usualmente se denota por Z.

Resultado D.2.12. *Si $Y \sim N(\mu, \sigma^2)$, y α, β son constantes, entonces la variable $\alpha Y + \beta$ tiene distribución $N(\alpha\mu + \beta, \alpha^2\sigma^2)$.*

Prueba. Se usará el hecho de que la función generadora de momentos caracteriza

la distribución probabilística. Se tiene que:

$$\begin{aligned} m_{\alpha Y+\beta}(t) &= E(e^{t(\alpha Y+\beta)}) \\ &= E(e^{\alpha t Y})e^{\beta t} \\ &= m_Y(\alpha t)e^{\beta t} \\ &= e^{\mu\alpha t+\sigma^2\alpha^2 t/2}e^{\beta t} \\ &= e^{(\alpha\mu+\beta)t+\sigma^2\alpha^2 t/2} \end{aligned}$$

el cual es la función generadora de momentos de una distribución $N(\alpha\mu+\beta, \alpha^2\sigma^2)$, y el resultado queda demostrado. ■

Como consecuencia inmediata del anterior resultado, se define la estandarización que es fundamental en la teoría relacionado con las distribuciones normales.

Estandarización: Si $Y \sim N(\mu, \sigma^2)$, entonces la variable $Z = \frac{Y-\mu}{\sigma}$ tiene distribución normal estándar, y la anterior transformación se conoce como la estandarización.

Resultado D.2.13. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes, donde $Y_i \sim N(\mu_i, \sigma_i^2)$ con $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n Y_i$ tiene distribución $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.*

D.2.9 Distribución log-normal

DefiniciÃ³n D.2.9. *Una variable aleatoria Y tiene distribución log-normal si su función de densidad está dada por:*

$$p(Y \mid \mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\{\frac{-1}{2\sigma^2(\ln(y)-\mu)^2}\} \tag{D.2.9}$$

Nótese que si μ y σ^2 son la media y la varianza de $\ln(Y)$, entonces $\ln(Y)$ tiene distribución normal de media μ y varianza σ^2.

Resultado D.2.14. *Si Y es una variable aleatoria con distribución log-normal, entonces*

1. $E(Y) = \exp(\mu + \sigma^2/2)$.
2. $Var(Y) = (\exp(\sigma^2 - 1)) \exp(2\mu + \sigma^2)$.

D.2.10 Distribución chi-cuadrado

DefiniciÃ³n D.2.10. *Una variable aleatoria Y tiene distribución chi-cuadrado con n grados de libertad, con n entero positivo, si su función de densidad está dada por:*

$$f_Y(y) = \frac{y^{(n/2)-1}e^{-y/2}}{2^{n/2}\Gamma(n/2)} I_{(0,\infty)}(y), \tag{D.2.10}$$

y se nota como $Y \sim \chi_n^2$.

Nota: La distribución chi-cuadrado con n grados de libertad es un caso particular de la distribución Gamma cuando el parámetro de forma k toma el valor $n/2$ y el parámetro de escala toma el valor 2.

También en la literatura estadística existe la siguiente definición para la distribución chi-cuadrado.

DefiniciÃ³n D.2.11. *Si $Z_1, \cdots, Z_n$ son variables aleatorias independientes e idénticamente distribuidas con distribución normal estándar, entonces la variable $\sum_{i=1}^{n} Z_i^2$ tiene distribución chi-cuadrado con n grados de libertad.*

Resultado D.2.15. *Si Y es una variable aleatoria con distribución chi-cuadro con n grados de libertad, entonces*

1. $E(Y) = n$.
2. $Var(Y) = 2n$.
3. $m_Y(t) = \left(\frac{1}{1-2t}\right)^{n/2}$ *para* $t < 1/2$, *y no existe para otros valores de* t.

Resultado D.2.16. *Sea $Z_1, \cdots, Z_m$ variables aleatorias independientes con distribución $\chi^2_{n_i}$ para $i = 1, \cdots, m$, entonces la variable $\sum_{i=1}^{m} Z_i$ tiene distribución chi-cuadrado con $\sum_{i=1}^{m} n_i$ grados de libertad.*

Prueba. Se deja como ejercicio. ■

D.2.11 Distribución t-student

El descubrimiento de la distribución t-student fue publicado por el estadístico inglés William Sealy Gosset (1876-1937) en el año 1908 cuando trabajaba en la famosa empresa cervecera Guinness. La publicación lo hizo de forma anónimo bajo el nombre de Student, pues Guinness le prohibía la publicación por ser el descubrimiento parte de resultados de investigación realizado por la empresa.

DefiniciÃ³n D.2.12. *Una variable aleatoria Y tiene distribución t-student con n grados de libertad si su función de densidad está dada por:*

$$f_Y(y) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\ \Gamma(\frac{n}{2})} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2} I_{\mathbb{R}}(y), \tag{D.2.11}$$

donde $n > 0$ y se nota como $Y \sim t_n$.

Otra definición que se encuentra frecuentemente en la literatura estadística es la siguiente:

DefiniciÃ³n D.2.13. *Sea Z una variable aleatoria con distribución normal estándar y Y una variable aleatoria con distribución chi-cuadrado con n grados de libertad, si Z y Y son independientes, entonces la variable $\frac{Z}{\sqrt{Y/n}}$ tiene distribución t-student con n grados de libertad.*

Nota: La función de densidad de la distribución t-student es muy parecida a la de distribución normal estándar, entre más grande sea el grado de libertad, más se parece a la distribución normal estándar.

Resultado D.2.17. *Si Y es una variable aleatoria con distribución t-student con n grados de libertad, entonces*

1. $E(Y) = 0$ *para* $n > 1$.
2. $Var(Y) = \frac{n}{n-2}$ *para* $n > 2$.

Nota: La distribución t-student no tiene función generadora de momentos.

D.2.12 Distribución t-student generalizada

DefiniciÃ³n D.2.14. *Una variable aleatoria Y tiene distribución t-student con n grados de libertad, parámetro de centralidad θ y parámetro de escala σ^2, si su función de densidad está dada por:*

$$f_Y(y) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}\sigma}\left[1 + \frac{1}{n}\left(\frac{y-\theta}{\sigma}\right)^2\right]^{-(n+1)/2} I_{\mathbb{R}}(y), \tag{D.2.12}$$

donde $n > 0$ y se nota como $Y \sim t_n(\theta, \sigma^2)$.

Resultado D.2.18. *Si Y es una variable aleatoria con distribución t-student generalizada, entonces*

1. $E(Y) = \theta$ *para* $n > 1$.
2. $Var(Y) = \frac{n}{n-2}\sigma^2$ *para* $n > 2$.

D.2.13 Distribución F

La distribución F también se conoce como la distribución F de Fisher o distribución de Fisher-Snedecor, refiriendo al gran estadístico Ronald Aylmer Fisher (1890-1962) y el fundador del primer departamento de estadística en los Estados Unidos, George Waddel Snedecor (1881-1974).

DefiniciÃ³n D.2.15. *Una variable aleatoria Y tiene distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador si su función de densidad está dada por:*

$$f_Y(y) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{m/2}\frac{z^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}z\right)^{\frac{m+n}{2}}}, \tag{D.2.13}$$

y se nota como $Y \sim F_n^m$.

Otra definición para la distribución F es como sigue:

DefiniciÃ³nD.2.16. *Sea Y y Y variables aleatorias independientes con distribuciones chi-cuadrado con m y n grados de libertad, respectivamente, entonces la variable $\frac{Y/m}{Y/n}$ tiene distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador.*

Resultado D.2.19. *Si Y es una variable aleatoria con distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador, entonces*

1. $E(Y) = \frac{n}{n-2}$ *para* $n > 2$.
2. $Var(Y) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ *para* $n > 4$.

Nota: La distribución F no tiene función generadora de momentos.

D.3 Distribuciones multivariadas

D.3.1 Distribución Multinomial

DefiniciÃ³nD.3.1. *Un vector aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_p')$ tiene distribución multinomial si su función de densidad está dada por:*

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \binom{n}{y_1, \ldots, y_p} \theta_1^{y_1} \cdots \theta_p^{y_p} \qquad \theta_i > 0\,, \ \sum_{i=1}^{p} \theta_i = 1 \ \ \boldsymbol{y} \ \ \sum_{i=1}^{p} y_i = p \quad \text{(D.3.1)}$$

donde

$$\binom{p}{y_1, \ldots, y_p} = \frac{p!}{y_1! \cdots y_p!}. \quad \text{(D.3.2)}$$

Como **?** afirma esta distribución es una generalización de la distribución binomial. La distribución marginal de una sola variable Y_i es $Binomial(p, \theta_i)$

Resultado D.3.1. *Si $\mathbf{Y}$ es una vector aleatorio con distribución multinomial, entonces*

1. $E(\mathbf{Y}) = p(\theta_1, \ldots, \theta_p)'$.
2. $Var(\mathbf{Y})_{ij} = \begin{cases} p\theta_i(1-\theta_i) & si\ i = j \\ -p\theta_i\theta_j & si\ i \neq j \end{cases}$.

D.3.2 Distribución Dirichelt

DefiniciÃ³nD.3.2. *Un vector aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_p')$ tiene distribución Dirichelt si su función de densidad está dada por:*

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \frac{\Gamma(\theta_1 + \cdots + \theta_p)}{\Gamma(\theta_1) \cdots \Gamma(\theta_p)} y^{\theta_1 - 1} \cdots y^{\theta_p - 1} \qquad \theta_i > 0 \ \ \boldsymbol{y} \ \ \sum_{i=1}^{p} \theta_i = 1. \quad \text{(D.3.3)}$$

Esta distribución es una generalización de la distribución beta. La distribución marginal de una sola variable Y_i es $Beta(\theta_i, (\sum_{i=1}^{p} \theta_i) - \theta_i)$

Resultado D.3.2. *Si* $\mathbf{Y}$ *es una vector aleatorio con distribución Dirichlet, entonces*

1. $E(\mathbf{Y}) = (\sum_{i=1}^{p} \theta_i)^{-1}(\theta_1, \ldots, \theta_p)'$.
2. $Var(\mathbf{Y})_{ij} = \begin{cases} \frac{\theta_i(\sum_{i=1}^{p} \theta_i - \theta_i)}{(\sum_{i=1}^{p} \theta_i)^2(\sum_{i=1}^{p} \theta_i + 1)} & si\ i = j \\ -\frac{\theta_i \theta_j)}{(\sum_{i=1}^{p} \theta_i)^2(\sum_{i=1}^{p} \theta_i + 1)} & si\ i \neq j \end{cases}$.

D.3.3 Distribución Normal Multivariante

DefiniciÃ³nD.3.3. *Un vector aleatorio* $\mathbf{Y} = (Y_1, \ldots, Y_p')$ *tiene distribución normal multivariante de orden p, denotada como* $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$*, si su función de densidad está dada por:*

$$p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\theta})'\boldsymbol{\Sigma}(\mathbf{y} - \boldsymbol{\theta})\right\} \tag{D.3.4}$$

donde $\mid \boldsymbol{\Sigma} \mid$ *se refiere al determinante de la matriz* $\boldsymbol{\Sigma}$*, la cual es simétrica y definida positiva de orden* $p \times p$.

La distribución Normal Multivariante es el baluarte de una gran cantidad de técnicas y métodos estadísticos como son los modelos lineales, los modelos lineales generalizados, el análisis factorial, etc. Algunas de sus propiedades se citan a continuación.

Resultado D.3.3. *Si* $\mathbf{Y} = (Y_1, \ldots, Y_p')$ *es una vector aleatorio con distribución normal multivariante, entonces*

1. *La distribución marginal de cualquier subconjunto de componentes de* $\mathbf{Y}$ *es también normal multivariante. Por ejemplo si* $\mathbf{Y}$ *es particionado en* $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')$*, entonces* $p(\mathbf{Y}_1)$ *seguiría una distribución normal multivariante, al igual que* $p(\mathbf{Y}_2)$.
2. *Cualquier transformación lineal de* $\mathbf{Y}$ *es normal multivariante y su dimensión equivale al rango de la transformación. en particular, la suma de las componentes del vector, dada por* $\sum_{i=1}^{p} Y_i$ *sigue una distribución normal univariada.*
3. *La distribución condicional de* $\mathbf{Y}$*, restringida a un subespacio lineal es normal.*
4. *La distribución condicional de cualquier sub-vector de elementos de* $\mathbf{Y}$ *dados los restantes elementos es normal multivariante. Más aún, si* $\mathbf{Y}$ *es particionado en* $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')$*, entonces* $p(\mathbf{Y}_1 \mid \mathbf{Y}_2)$ *es normal multivariada con*

$$E(\mathbf{Y}_1 \mid \mathbf{Y}_2) = E(\mathbf{Y}_1) + Cov(\mathbf{Y}_1, \mathbf{Y}_2)(Var(\mathbf{Y}_2))^{-1}(\mathbf{Y}_2 - E(\mathbf{Y}_2))$$
$$Var(\mathbf{Y}_1 \mid \mathbf{Y}_2) = Var(\mathbf{Y}_1) - Cov(\mathbf{Y}_1, \mathbf{Y}_2)(Var(\mathbf{Y}_2))^{-1}Cov(\mathbf{Y}_2, \mathbf{Y}_1)$$

5. *Si* $\mathbf{X}$ *es un vector con distribución normal multivariante, entonces* $\mathbf{X}+\mathbf{Y}$ *tiene una distribución normal multivariante. En particular si* $\mathbf{X}$ *es independiente de* $\mathbf{Y}$*, comparten el mismo orden* p *y* $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Gamma})$*, entonces* $\mathbf{X}+\mathbf{Y} \sim N_p(\boldsymbol{\mu}+\boldsymbol{\theta}, \boldsymbol{\Gamma}+\boldsymbol{\Sigma})$.

Resultado D.3.4. *Si* $\mathbf{Y}$ *es una vector aleatorio con distribución Normal Multivariante, entonces*

1. $E(\mathbf{Y}) = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$.
2. $Var(\mathbf{Y}) = \boldsymbol{\Sigma}$

Resultado D.3.5. *Dado* $\mathbf{Y}$ *un vector aleatorio particionado como* $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')$ *con esperanza* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ *y matrix de varianzas y covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Si $\mathbf{Y}_1 \mid \mathbf{Y}_2 \sim N(\boldsymbol{\theta}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\theta}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$ *y* $\mathbf{Y}_2 \sim N(\boldsymbol{\theta}_2, \boldsymbol{\Sigma}_{22})$, *entonces se tiene que*

$$\mathbf{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}).$$

Resultado D.3.6. *Si* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ *es una muestra aleatoria de vectores con distribución Normal Multivariante, entonces la verosimilitud de la muestra se puede escribir como*

$$\prod_{i=1}^{n} p(\mathbf{Y}_i \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}})\right\} \tag{D.3.5}$$

Donde $\mathbf{S}_{\boldsymbol{\theta}} = \sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})'$.

Prueba. La verosimilitud de la muestra aleatoria está dada por

$$\begin{aligned} \prod_{i=1}^{n} p(\mathbf{Y}_i \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) &\propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})\right\} \\ &= \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2} traza\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}}\right)\right\} \end{aligned}$$

Puesto que, por las propiedades del operador *traza*, se tiene que

- Si c es un escalar, entonces $c = traza(c)$.
- Si $\mathbf{A}$ y $\mathbf{B}$ son dos matrices, entonces $traza(\mathbf{AB}) = traza(\mathbf{BA})$
- Si $\mathbf{A}_i$ (i=1,. . . ,n) son matrices del mismo tamaño, entonces $\sum_{i=1}^{n} traza(\mathbf{A}_i) = traza\left(\sum_{i=1}^{n}\mathbf{A}_i\right)$

Por lo anterior,

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta}) &= traza\left[\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta})\right]\\ &= \sum_{i=1}^{n} traza[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta})(\mathbf{Y}_i-\boldsymbol{\theta})')]\\ &= traza\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})(\mathbf{Y}_i-\boldsymbol{\theta})')\right]\\ &= traza(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}})\end{aligned}$$

■

D.3.4 Distribución Wishart

DefiniciÃ³n D.3.4. *Sea* $\boldsymbol{\Sigma}$ *una matriz aleatoria simétrica y definida positiva de tamaño* $p\times p$. *Se dice que* $\boldsymbol{\Sigma}$ *tiene distribución Wishart con* v *grados de libertad, denotada como* $\mathbf{Y}\sim Wishart_v(\boldsymbol{\Lambda})$, *si su función de densidad está dada por:*

$$p(\boldsymbol{\Sigma}) = \left(2^{vp/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{v+1-i}{2}\right)\right)^{-1} \times \mid \boldsymbol{\Lambda}\mid^{-v/2}\mid \boldsymbol{\Sigma}\mid^{(v-p-1)/2}\exp\left\{-\frac{1}{2}traza(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma})\right\} \quad \text{(D.3.6)}$$

donde $\mid\boldsymbol{\Lambda}\mid$ *se refiere al determinante de la matriz* $\boldsymbol{\Lambda}$, *la cual es simétrica y definida positiva de orden* $p\times p$.

Resultado D.3.7. *Si* $\boldsymbol{\Sigma}$ *es una matriz aleatoria con distribución Wishart con* v *grados de libertad, entonces* $E(\boldsymbol{\Sigma})=v\boldsymbol{\Lambda}$

D.3.5 Distribución inversa-Wishart

DefiniciÃ³n D.3.5. *Sea* $\boldsymbol{\Sigma}$ *una matriz aleatoria simétrica y definida positiva de tamaño* $p\times p$. *Se dice que* $\boldsymbol{\Sigma}$ *tiene distribución Wishart con* v *grados de libertad, denotada como* $\mathbf{Y}\sim Wishart_v(\boldsymbol{\Lambda})$, *si su función de densidad está dada por:*

$$p(\boldsymbol{\Sigma}) = \left(2^{vp/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{v+1-i}{2}\right)\right)^{-1} \times \mid \boldsymbol{\Lambda}\mid^{v/2}\mid \boldsymbol{\Sigma}\mid^{-(v+p+1)/2}\exp\left\{-\frac{1}{2}traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1})\right\} \quad \text{(D.3.7)}$$

donde $\mid\boldsymbol{\Lambda}\mid$ *se refiere al determinante de la matriz* $\boldsymbol{\Lambda}$, *la cual es simétrica y definida positiva de orden* $p\times p$.

Resultado D.3.8. *Si* $\mathbf{\Sigma}$ *es una matriz aleatoria con distribución inversa-Wishart con* v *grados de libertad, entonces* $E(\mathbf{\Sigma}) = \dfrac{1}{v-p-1}\mathbf{\Lambda}$

Resultado D.3.9. *Si* $\mathbf{\Sigma}^{-1}$ *es una matriz aleatoria con distribución inversa-Wishart, entonces con* $\mathbf{\Sigma}$ *tiene distribución Wishart.*

E Algunos elementos de inferencia estadística

E.1 Matriz de información

DefiniciÃ³n E.1.1. *Dada X una variable aleatoria con función de densidad $f(x,\theta)$, donde θ es el parámetro de la distribución, y además existe $\dfrac{\partial}{\partial\theta}\ln f(x,\theta)$, entonces se define la información contenida en X acerca de θ como*

$$I_X(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2\right\}. \tag{E.1.1}$$

Resultado E.1.1. *En la anterior definición, si además existe $\dfrac{\partial^2}{\partial\theta^2}\ln f(x,\theta)$, entonces se tiene que*

$$I_X(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right\}. \tag{E.1.2}$$

Las anteriores definiciones introducen la información contenida en una variable, sin embargo, cuando tenemos disponible una muestra aleatoria, es necesario definir la información contenida en una muestra aleatoria acerca de algún parámetro.

DefiniciÃ³n E.1.2. *Dada $X_1, \cdots, X_n$ variables aleatorias con función de densidad $f(x_i,\theta)$, donde θ es el parámetro de la distribución, y además existe $\dfrac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(x_i,\theta)$, entonces se define la información contenida en la muestra aleatoria acerca de θ como*

$$I_{X_1,\cdots,X_n}(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta)\right]^2\right\}. \tag{E.1.3}$$

Resultado E.1.2. *Dada $X_1, \cdots, X_n$ una muestra aleatoria, entonces*

$$I_{X_1,\cdots,X_n}(\theta) = nI_X(\theta),$$

donde $I_X(\theta) = I_{X_i}(\theta)$, con $i = 1, \cdots, n$. Es decir, en una muestra aleatoria, cada variable aporta la misma cantidad de información, y la cantidad total de información en la muestra es la suma de la información en cada variable.

Prueba.

$$
\begin{aligned}
I_{X_1,\cdots,X_n}(\theta) &= E\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i,\theta)\right]^2\right\} \\
&= E\left\{\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} \\
&= E\left\{\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} + \\
&\qquad \underbrace{E\left\{\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\frac{\partial}{\partial\theta}\ln f(X_j,\theta)\right]\right\}}_{=0,\ \text{por la independencia entre } X_i \text{ y } X_j} \\
&= \sum_{i=1}^{n} E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} \\
&= \sum_{i=1}^{n} I_X(\theta) = nI_X(\theta).
\end{aligned}
$$

■

Ejemplo E.1.1. *Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de la distribución $N(\mu,\sigma^2)$, la información contenida en la muestra acerca de μ es n/σ^2. Para verificar esta afirmación, calculamos la información acerca de μ en una variable X con distribución $N(\mu,\sigma^2)$. Tenemos:*

$$
\begin{aligned}
I_X(\mu) &= -E\left\{\frac{\partial^2}{\partial\mu^2}\ln f(X,\theta)\right\} \\
&= -E\left\{\frac{\partial^2}{\partial\mu^2}\left[-\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2\right]\right\} \\
&= -E\left\{\frac{\partial}{\partial\mu}\left[\frac{X-\mu}{\sigma^2}\right]\right\} \\
&= -E\left\{-\frac{1}{\sigma^2}\right\} \\
&= \frac{1}{\sigma^2}.
\end{aligned}
$$

Ahora, usando el Resultado 2.3.4, se tiene que $I_{X_1,\cdots,X_n}(\mu) = n/\sigma^2$.

Nótese que esta información, en primer lugar, depende del tamaño n de manera que entre más grande sea la muestra, hay mayor información acerca de μ; en segundo lugar, entre más pequeña sea la varianza σ^2, la cantidad de información acerca de μ también incrementa, esto es natural, puesto que si σ^2 es pequeña, los

datos de la muestra están muy concentrados alrededor de μ, entonces estos datos aportan más información que otros datos con más dispersión.

DefiniciÃ³nE.1.3. *Dada una variable aleatoria X con función de densidad $f(x,\boldsymbol{\theta})$, la matriz de información contenida en X acerca de $\boldsymbol{\theta}$ se define como*

$$I_X(\boldsymbol{\theta}) = E\left\{\frac{\partial \ln f(X,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln f(X,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right\} \tag{E.1.4}$$

DefiniciÃ³nE.1.4. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad $f(x_i,\boldsymbol{\theta})$, la matriz de información contenida en la muestra acerca de $\boldsymbol{\theta}$ se define como*

$$I_{X_1,\cdots,X_n}(\boldsymbol{\theta}) = E\left\{\frac{\partial \ln \prod_{i=1}^n f(X_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln \prod_{i=1}^n f(X_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right\}$$

Ejemplo E.1.2. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución común $N(\mu,\sigma^2)$, vamos a hallar la matriz de información contenida en la muestra acerca del vector de parámetros (μ,σ^2). Tenemos que*

$$\begin{aligned}
&I_{X_1,\cdots,X_n}(\mu,\sigma^2)\\
&= E\left\{\begin{pmatrix}\dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \mu}\\ \dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \sigma^2}\end{pmatrix}\begin{pmatrix}\dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \mu} & \dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \sigma^2}\end{pmatrix}\right\}\\
&= E\left\{\begin{pmatrix}\dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma^2}\\ \dfrac{\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2}{2\sigma^4}\end{pmatrix}\begin{pmatrix}\dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma^2} & \dfrac{\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2}{2\sigma^4}\end{pmatrix}\right\}\\
&= E\left\{\begin{pmatrix}\dfrac{(\sum_{i=1}^n X_i - n\mu)^2}{\sigma^4} & \dfrac{(\sum_{i=1}^n X_i - n\mu)(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)}{2\sigma^6}\\ \dfrac{(\sum_{i=1}^n X_i - n\mu)(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)}{2\sigma^6} & \dfrac{(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)^2}{4\sigma^8}\end{pmatrix}\right\}
\end{aligned}$$

Donde el primer elemento diagonal de la anterior matriz está dada por

$$\begin{aligned}
E\left\{\frac{(\sum_{i=1}^n X_i - n\mu)^2}{\sigma^4}\right\} &= \left[Var\left(\sum_{i=1}^n X_i - n\mu\right) + (E\left(\sum_{i=1}^n X_i - n\mu\right))^2\right]/\sigma^4\\
&= n\sigma^2/\sigma^4 = n/\sigma^2.
\end{aligned}$$

El segundo elemento diagonal está dada por

$$E\left\{\frac{(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2)^2}{4\sigma^8}\right\} \tag{E.1.5}$$

$$=\frac{1}{4\sigma^8}E\left\{\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]^2+n^2\sigma^4-2n\sigma^2\sum_{i=1}^{n}(X_i-\mu)^2\right\} \tag{E.1.6}$$

$$=\frac{1}{4\sigma^8}\left\{Var(\sum_{i=1}^{n}(X_i-\mu)^2)+\left[E(\sum_{i=1}^{n}(X_i-\mu)^2)\right]^2+n^2\sigma^4-2n\sigma^2E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]\right\} \tag{E.1.7}$$

Usando el hecho de que

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}\sim\chi_n^2$$

*y la esperanza y varianza de la distribución χ_n^2, tenemos que la expresión (**??**) está dada por*

$$\frac{1}{4\sigma^8}\left\{2n\sigma^4+\left[n\sigma^2\right]^2+n^2\sigma^4-2n\sigma^2n\sigma^2\right\}=\frac{n}{2\sigma^4}.$$

Finalmente, el elemento fuera de la diagonal de la matriz $I_{X_1,\cdots,X_n}(\mu,\sigma^2)$ está dado por

$$\begin{aligned}
&E\left\{\left(\sum_{i=1}^{n}X_i-n\mu\right)\left(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2\right)\right\}\\
&=E\left\{\sum_{i=1}^{n}X_i\left(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2\right)-n\mu\left(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2\right)\right\}\\
&=E\left\{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}(X_i-\mu)^2\right\}-n\sigma^2E\left(\sum_{i=1}^{n}X_i\right)-n\mu E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)+n^2\mu\sigma^2\\
&=E\left(\sum_{i=1}^{n}X_i\sum_{i=1}^{n}X_i^2\right)-2\mu E\left[(\sum_{i=1}^{n}X_i)^2\right]+n^2\mu^3-n^2\mu\sigma^2-n^2\mu\sigma^2+n^2\mu\sigma^2\\
&=E\left(\sum_{i=1}^{n}X_i^3+\sum_{i\neq j}X_iX_j^2\right)-2\mu(n\sigma^2+n^2\mu^2)+n^2\mu^3-n^2\mu\sigma^2\\
&=\sum_{i=1}^{n}\left[3\mu E(X_i^2)-2\mu^3\right]+\sum_{i\neq j}E(X_i)E(X_j^2)-2n\mu\sigma^2-2n^2\mu^3+n^2\mu^3-n^2\mu\sigma^2\\
&=3n\mu(\sigma^2+\mu^2)-2n\mu^3+\mu(\sigma^2+\mu^2)(n^2-n)-2n\mu\sigma^2-2n^2\mu^3+n^2\mu^3-n^2\mu\sigma^2\\
&=0
\end{aligned}$$

De donde obtenemos finalmente la matriz de información $I_{X_1,\cdots,X_n}(\mu,\sigma^2)$ *dada por*

$$I_{X_1,\cdots,X_n}(\mu,\sigma^2)=\begin{pmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{n}{2\sigma^4}\end{pmatrix}$$

F Simulación de distribuciones de probabilidad

Como lo afirma **?** la simulación numérica es parte central del análisis bayesiano puesto que la generación de datos provenientes de una distribución de probabilidad se puede realizar fácilmente, incluso cuando la forma estructural de ésta no es conocida o es muy complicada computacionalmente. A lo largo de la historia del desarrollo de la teoría estadística, la simulación de distribuciones de probabilidad ha jugado un papel importante. Aunque son innumerables los métodos de generación de datos, en este apartado, se da cuenta de unos pocos, quizás lo más usados en este auge computacional.

F.1 Método de la transformación uniforme

Al momento de la simulación estocástica de observaciones provenientes de alguna distribución de interés, la distribución uniforme es quizás la más usada y la más importante. El siguiente resultado adaptado de **?** así lo confirma.

Resultado F.1.1 (Transformación integral de probabilidad)**.** *Si U es una variable aleatoria con distribución uniforme en el intervalo $(0, 1)$, entonces la la variable aleatoria $F^{-1}(U)$ tiene distribución F.*

Aunque la función F no necesariamente es una función uno a uno, por lo menos no lo es en el caso discreto, sí se puede verificar que $F^{-1}(U)$ es única con probabilidad uno. Una definición general, que encaja en el caso continuo o discreto, de la función F inversa es la siguiente

DefiniciÃ³nF.1.1. *Para cualquier función F definida sobre $\mathbb{R}$, se define la función inversa generalizada de F como*

$$F^{-1}(u) = \text{ínf}\{x \mid F(x) \geq u\} \tag{F.1.1}$$

Ejemplo F.1.1. *Suponga que X es una variable aleatoria con distribución exponencial dada por (B.2.4). De esta forma, su función de densidad acumulativa viene dada por*

$$F(x) = 1 - \exp\{-\theta x\}$$

Del anterior resultado se tiene que si u es una realización de una variable $U \sim Uniforme(0, 1)$, entonces $F^{-1}(u)$ es una realización de una variable con

distribución exponencial. Como $x = F^{-1}(u)$, *entonces* $F(x) = u$ *y despejando* x, *se llega a que la siguiente expresión*

$$F^{-1}(u) = -\frac{\ln(1-u)}{\theta}$$

entrega una forma diáfana para la simulación de una observación con distribución exponencial. Para simular una muestra de n *observaciones, simplemente se repite el anterior procedimiento* n *veces. En* R, *el código necesario para la simulación de una muestra de tamaño 1000 proveniente de una distribución exponencial con parámetro* $\theta = 5$ *es*

```
> theta <- 5
> u <- runif(1000)
> rexpo <- log(1-u)/(-theta)
> 1/mean(rexpo)
[1] 5.076471

> hist(rexpo,breaks=100)
> lines(density(rexpo),col=2)
```

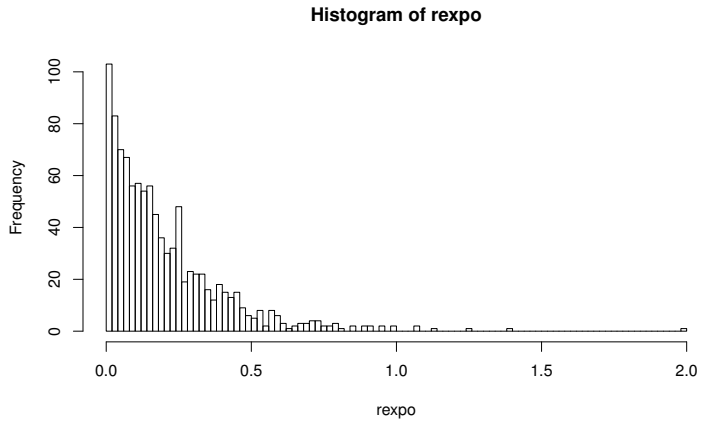

Figura F.1: *Histograma de* $n = 1000$ *observaciones con distribución exponencial*

F.1.1 Método de la grilla

Existen distribuciones de probabilidad cuya forma estructural es muy compleja. Mas aún, existen distribuciones de probabilidad conocidas para las cuales la inversa de la función de de densidad acumulativa es difícil de solucionar analíticamente. En los anteriores casos, el método analítico dado por el teorema de la transformación integral de probabilidad no siempre resulta efectivo. Sin embargo, es posible realizar una variante, manteniendo el espíritu de la anterior técnica.

El presente método utiliza una distribución discreta para aproximar cualquier tipo de distribución (discreta o continua) sin importar su nivel de complejidad. El algoritmo que enmarca este método se da a continuación:

1. Escribir la densidad de interés como $f(\cdot)$ y establecer el rango de la variable aleatoria de interés.
2. Fijar un conjunto de n valores $x_1 < \cdots < x_n$ equiespaciados que cubran una gran parte del rango de la variable aleatoria.
3. Para x_k $(k = 1, \ldots, n)$ calcular $f(x_k)$ que equivale al valor de la densidad en el punto x_k. Nótese que si $f(\cdot)$ es una función de densidad continua, entonces $f(x_k)$ no corresponde a una probabilidad;
4. Calcular la probabilidad asociada al punto x_k definida por la aproximación discreta a $f(\cdot)$ y dada por

$$p(x_k) = \frac{f(x_k)}{\sum_{k=1}^{n} f(x_k)}$$

5. Calcular la función de densidad acumulativa aproximada definida como

$$F(x) = \begin{cases} 0, & \text{si } x < x_1 \\ \sum_{l=1}^{k} p(x_l), & \text{si } x_k \leq x < x_{k+1} \\ 1, & \text{si } x > x_n \end{cases}$$

6. Simular una observación u proveniente de una distribución uniforme continua en el intervalo $(0, 1)$.
7. Si $F(x_k) < u \leq F(x_{k+1})$, entonces $F^{-1}(u) = x_{k+1}$ y por consiguiente el valor x_{k+1} es una pseudo-observación proveniente de la densidad de interés.

Nótese que en el anterior proceso, la unidad x_{k+1} es seleccionada con probabilidad p_{k+1}; puesto que

$$\begin{aligned} P(F(x_k) < U \leq F(x_{k+1})) &= F(x_{k+1}) - F(x_k) \\ &= \sum_{l=1}^{k+1} p(x_l) - \sum_{l=1}^{k} p(x_l) = p_{k+1} \end{aligned}$$

Si se quiere extraer una muestra aleatoria de N observaciones provenientes de la distribución de interés, entonces basta con repetir el anterior proceso N veces. Por supuesto, como se trata de una muestra aleatoria cada selección se debe realizar con repetición; de esta manera no importa si $N > n$. Suponiendo que el conjunto $x_1, \ldots, x_n$ conforma una grilla de puntos lo suficientemente cercanos y que no sucede nada importante entre cada uno de ellos, entonces esta técnica debe tener un buen funcionamiento.

Distribución exponencial

un sólo parámetro univariada y continua

```
> ## Simulación para la distribución exponencial (theta=5)

> theta<-5
> x.grid<-seq(0,100,by=0.01)
> p.exp<-theta*exp(-theta*x.grid)

> r.exp<-sample(x.grid,10000,prob=p.exp,replace=T)
> 1/mean(r.exp)
[1] 5.139459
> hist(r.exp,breaks=100,freq=F)
```

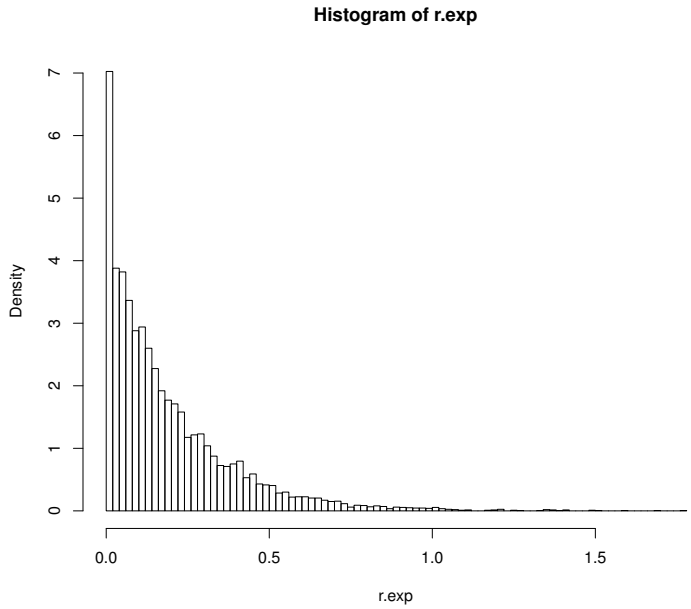

Figura F.2: *Histograma de observaciones con distribución exponencial*

Distribución Poisson

un sólo parámetro univariada y discreta

```
> ## Simulación para la distribución poisson (theta=2)

> p.poisson <-function(theta, x.grid){
+ N<-length(x.grid)
+ res <- rep(NA, N)
+ for(k in 1:N){
```

```
+ P1 <- exp(-theta)*theta^(x.grid[k])
+ P2 <- factorial(x.grid[k])
+ res[k] <-P1/P2
+ }
+ res
+ }

> t <- 2
> x.grilla <- seq(0,100,by=1)
> p.x <- p.poisson(t, x.grilla)/sum(p.poisson(t, x.grilla))
> sum(p.x)
[1] 1

> r.pois <- sample(x.grilla,10000,prob=p.x,replace=T)
> mean(r.pois)
[1] 1.9929
> var(r.pois)
[1] 1.981248
> hist(r.pois,breaks=100,freq=F)
```

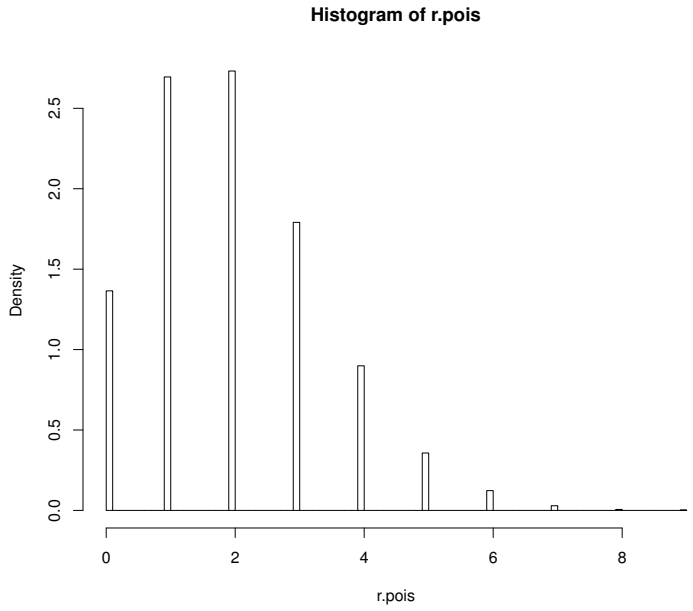

Figura F.3: *Histograma de observaciones con distribución Poisson*

Distribución Gamma

biparamétrica, univariada y continua

```
> ## Simulación para la distribución gamma (alpha=4, beta=2)

> p.gamma <- function(a, b, x.grid){
+ N<-length(x.grid)
+ res <- rep(NA, N)
+ for(k in 1:N){
+ P1<- (b^a)/gamma(a)
+ P2<- x.grid[k]^(a-1)
+ P3<- exp(-b*x.grid[k])
+ res[k] <- P1*P2*P3
+ }
+ res
+ }

> alpha <- 4
> beta <- 2
> x.grilla <- seq(0,100,by=0.1)
> p.x <- p.gamma(alpha, beta, x.grilla)/sum(p.gamma(alpha, beta, x.grilla))
> sum(p.x)
[1] 1

> r.gamma <- sample(x.grilla,10000,prob=p.x,replace=T)
> mean(r.gamma)
[1] 2.00477
> var(r.gamma)
[1] 0.9956038
> hist(r.gamma,breaks=100,freq=F)
```

Distribución Normal bivariada

multiparamétrica, bivariada y continua

```
> ## Simulación para la distribución normal bivariada
> ## mu=(2,4), Sigma = (25, 30, 30, 16)

> p.normal2 <- function(mu, Sigma, x, y){
+ P1<- 1/(2*pi)
+ P2<- 1/sqrt(det(Sigma))
+ P3a <- t((c(x,y)-mu))%*%solve(Sigma)%*%(c(x,y)-mu)
+ P3 <- exp((-1/2)*P3a)
+ res <- P1*P2*P3
```

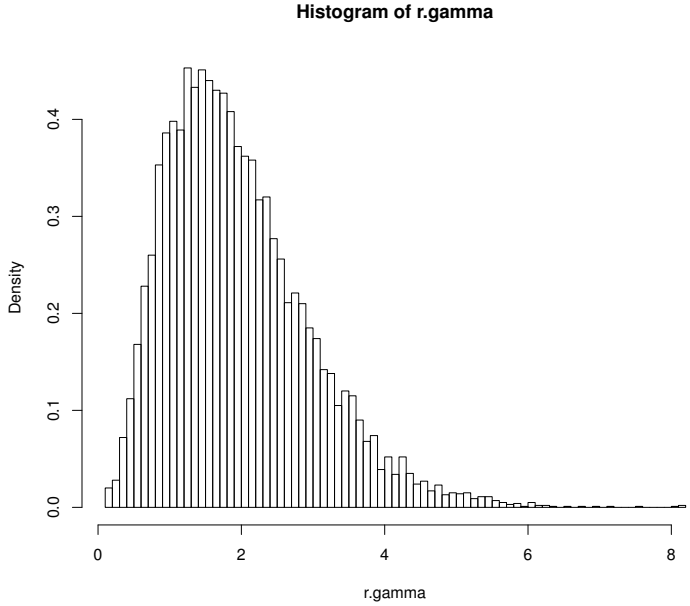

Figura F.4: *Histograma de observaciones con distribución Gamma*

```
+ res
+ }

> grilla<-function(a,b){
+ A<-seq(1:length(a))
+ unoA <-rep(1,length(A))
+ B<-seq(1:length(b))
+ unoB <-rep(1,length(B))
+ P1<-kronecker(A,unoB)
+ P2<-kronecker(unoA,B)
+ grid<-cbind(a[P1],b[P2])
+ return(grid)
+ }

> mu1 <- c(2,4)
> Sigma1 <- matrix(c(25, 30, 30, 100), nrow=2)
> x.grid<-seq(mu1[1]-3*sqrt(Sigma1[1,1]),mu1[1]+3*sqrt(Sigma1[1,1]),by=0.5)
> y.grid<-seq(mu1[2]-3*sqrt(Sigma1[2,2]),mu1[2]+3*sqrt(Sigma1[2,2]),by=0.5)
> xy.grid<-grilla(x.grid,y.grid)
> N.grid<-dim(xy.grid)[1]

> p.xy <- rep(NA, N.grid)
```

```
> for(j in 1:N.grid){
+ p.xy[j] <- p.normal2(mu1, Sigma1, xy.grid[j,1], xy.grid[j,2])
+ }

> p<-as.vector(p.xy/sum(p.xy))
> sum(p)
[1] 1
> r.normal2<-sample(N.grid,5000,prob=p,replace=T)
> rxy.normal2<-xy.grid[r.normal2,]
> rx.normal<-rxy.normal2[,1]
> ry.normal<-rxy.normal2[,2]

> mean(rx.normal)
[1] 2.116
> mean(ry.normal)
[1] 4.0845
> var(rxy.normal2)
         [,1]     [,2]
[1,] 24.58536 29.33552
[2,] 29.33552 97.52201

> hist(rx.normal,breaks=100,freq=F)
> hist(ry.normal,breaks=100,freq=F)

> a<-x.grid
> b<-y.grid

> mat<-matrix(NA, nrow=length(a), ncol=length(b))
> for(i in 1:length(a)){
+ for(j in 1:length(b)){
+ mat[i,j]<-p.normal2(mu1,Sigma1,a[i],b[j])
+ }
+ }
> mat<-mat/(sum(mat))

> contour(a,b,mat)
> persp(a,b,mat)
```

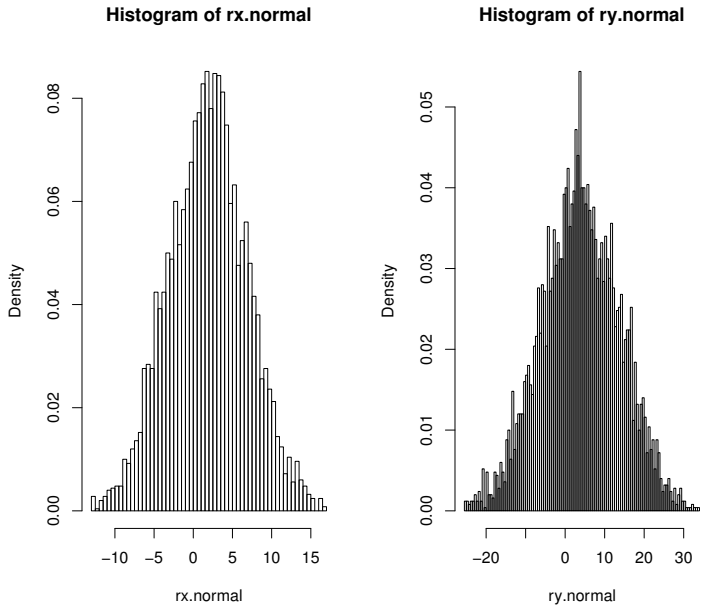

Figura F.5: *Diagrama de contorno de observaciones con distribución exponencial*

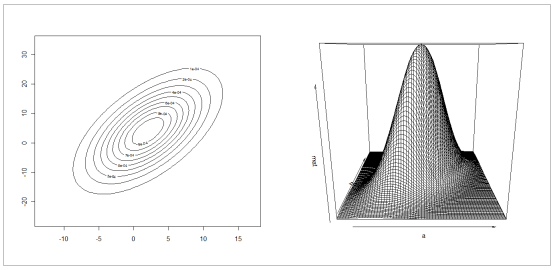

Figura F.6: *Diagrama de contorno de observaciones con distribución exponencial*

Índice de figuras

Ãndice de Tablas

Bibliografía

Agresti, A., Booth, J. G., Hobart, J. P. and Caffo, B. (2000), 'Random-effects modeling of categorical response data', *Sociological Methodology* **30**, 27 – 80.

Agresti, A. and Coull, B. A. (1998), 'Approximate is better than exact for interval estimation of binomial proportions', *The American Statistician* **52**(2), 119–126.

Agresti, A. and Min, Y. (2005), 'Frequentist performance of bayesian confidence intervals for comparing proportions in 2×2 contingency tables', *Biometrics* **61**, 515–523.

Aitkin, M. (1987), 'Modelling variance heterogeneity in normal regression using glim', *Applied Statistics* **36**, 332 – 339.

Aitkin, M. (1997), 'Modelling variance heterogeneity in normal regression using glim', *Applied Statistics* **36**, 332 – 339.

Albert, J. (2007), *Bayesian Computation with R*, Springer.

Anderson, D. A. and Aitkin, M. (1985), 'Variance component models with binary response: Inteviewer variability', *Journal of the Royal Statistical Association* **B. 47**, 203 – 210.

Apostol, T. M. (1957), *Mathematical Analysis*, McGraw - Hill.

Bailey, W. N. (1934), 'On the Reducibility of Appell's Function F_4', *Quart. J. Math* **5**, 291–292.

Barndorf-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, Wiley.

Barry, S. C., Brooks, S. P., Catchpole, E. A. and Morgan, B. J. T. (2003), 'The analysis of ring-recovery data using random effects', *Biometrics* **59**, 54 – 65.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, 2 edn, Springer.

Bernardo, J. M. and Smith, A. F. M. (2000), *Bayesian Theory*, Wiley.

Bickel, P. J. and Doksum, K. A. (1977), *Mathemtical Statistics*, Holden-Day.

Bock, R. D. (1989), *Multilevel Analysisi of Educational Data*, academic Press.

Bolfarine, H. and Zacks, S., eds (1991), *Prediction Theory for Finite Populations*, Springer-Verlag.

Box, G. E. P. and Tiao, G. C. (1992), *Bayesian Inference in Statistical Analysis*, 1 edn, Wiley.

Brown, H. and Prescott, R. (1999), *Applied Mixed Models in Medicine*, Wiley.

Canavos, G. C. (1988), *Probabilidad y estadística aplicaciones y métodos*, McGraw-Hill.

Carlin, B. P. and Louis, T. A. (1996), *Bayes and Empirical Bayes for Data Analysis*, 1 edn, Chapman and Hall/CRC.

Carlin, B. P. and Louis, T. A. (2009), *Bayesian Methods for Data Analysis*, 3 edn, CRC.

Cepeda, E. (2001), Variability Modeling in Generalized Linear Models, PhD thesis, Mathematics Institute, Universidade Federal do Rio de Janeiro.

Cepeda, E. and Achcar, J. A. (2010), 'Heteroscedastic nonlinear regression models', *Comunnications in Statistics Simulation and Computation* **39**, 405 – 419.

Cepeda, E. and Gamerman, D. (2001), 'Bayesian modeling of variance heterogeneity in normal regression models', *Brazilian Journal of Probability and Statistics* **14**, 207 – 221.

Cepeda, E. and Gammerman, D. (2005), 'Bayesian methodology for modeling parameters in the two parameter exponential family', *Estadística* **57**, 93 – 105.

Cepeda, E. and Nuñez, V. A. (2009), 'Bayesian modeling of the mean and covarince matrix in normal nonlinear models', *Journal of Statistical Computation and Simulation* **79**, 837 – 853.

Chambers, R. L. and Skinner, C. J. (2003*a*), *Analysis of Survey Data*, Wiley.

Chambers, R. L. and Skinner, C. J., eds (2003*b*), *Analysis of Survey Data*, Wiley.

Chaudhuri, A. (1994), 'Small domain statistics: A review', *Statistica Neerlandica* **48**, 215 – 236.

Clayton, D. G. and Kaldor, J. M. (1987), 'Empirical bayes estimates of age-standardized relative risk for use in disease mapping', *Biometrics* **43**, 671 – 682.

Cochran, W. G. (1976), *Sampling Techniques*, Wiley.

Cox, D. R. and Reid, N. (1987), 'Parameter orthogonality and approximate conditional inference (with discussion)', *Journal of the Royal Statistical Association* **B. 49**, 1 – 39.

Crowley, J. and Hu, M. (1977), 'Covariance analysis of heart transplant survival data', *Journal of the American Statistical Association* **72**, 27 – 36.

Davidian, M. (2009), *Longitudinal Data Analysis,*, Chapman and Hall, chapter Non-linear mixed-effects models, pp. 107 – 141.

Davidian, M. and Giltinan, D. M. (1995), *Non-linear Models for Repeated Measurement Data*, Chapman and hall.

Demidenko, E. (2004), *Mixed Models: Theory and Application*, Wiley.

Dempster, A. P. (1974), The direct use of likelihood for significance testing, *in* 'Proceedings of Conference on Foundational Questions in Statistical Inference', Department of Theoretical Statistics: University of Aarhus., pp. 335 – 352.

Denis, J. B. (1983), 'Extension du modele additif danalyse de variance par modelisation multiplicative des variances', *Biometrics* **39**, 849 – 856.

Dey, D. K., Gelfand, A. E. and Peng, F. (1997), 'Overdispersed generalizaed linear models', *Journal of Statistical Planning and Inference* **64**, 93 – 107.

Diggle, P., Heagerty, P., Liang, K. Y. and Zeger, S. (2002), *Analysis of Longitudinal Data*, Oxford University Press.

Efron, B. (2010), *Large-Scale Inference. Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press.

Efron, B. and Morris, C. (1975), 'Data analysis using stein's estimator and its generalizations', *Journal of the American Statistical Association* **70**, 311 – 319.

Fahrmeir, L. and Tutz, G. T. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer-Verlag.

Faraway, J. J. (2006), *Extending the Linear Model with R*, CRC.

Fisher, R. A. (1970), *Statistical Methods for Research Workers*, 15 edn, Macmillan Pub. Co.

Foulley, J. L. and Quaas, R. L. (1995), 'Heterogeneous variances in gaussian linear mixed models', *Genetics Selection Evolution* **27**, 211 – 228.

Gamerman, D. (1997), 'Sampling from the posterior distributions in generalized linear mixed models', *Statistics and Computing* **7**, 57 – 68.

Gamerman, D. and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman and Hall/CRC.

Gelfand, A. E. and Dalal, S. R. (1990), 'A note on overdispersed exponential families', *Biometrika* **77**, 55 – 64.

Gelfand, A. E. and Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Society* **85**, 398 – 409.

Gelman, A. (2008), 'Objections to bayesian statistics', *Bayesian Analysis* **3**(3), 445–450.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995), *Bayesian Data Analysis*, 1 edn, Chapman and Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003), *Bayesian Data Analysis*, 2 edn, Chapman and Hall/CRC.

Gelman, A. and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

Gelman, A. and Shirley, K. (2010), *Handbook of Markov Chain Monte Carlo*, CRC, chapter Inference from Simulations and Monitoring Convergence.

Geman, S. and Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721 – 741.

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998), 'Generalized linear models for small area estimation', *Journal of American Statististical Association* **93**, 273 – 282.

Ghosh, M. and Rao, J. N. K. (1994), 'Small area estimation: An appraisal', *Statistical Sciences* **9**(1), 55–93.

Gilmour, A. R. and Anderson, R. D. an Rae, A. L. (1985), 'The analysis of binomial

data by a generalized linear mixed model', *Biometrika* **72**, 593 – 599.

Giltinan, D. and Davidian, M. (1995), *Nonlinear Models for Repeted Measurement*, Chapman and Hall.

Goldstein, H. (1990), *Multilevel Statistical Models*, Wiley.

Gonzalez, M. E. (1973), Use and evaluation of synthetic estimators, *in* 'Proceedings of the Social Statistics Section', American Statistical Association, pp. 33 – 36.

Halimi, R. E. (2005), Nonlinear Mixed-effects Models and Nonparametric Inference, PhD thesis, Departament dÉstadística, Universitat de Barcelona.

Harvey, A. (1976), 'Estimating regression models with multiplicative heteroscedasticity', *Econometrica* **44**, 461 – 465.

Hastings, W. K. (1970), 'Monte carlo sampling methods using markov chains and their applications', *Biometrika* **57**, 97 – 109.

Hedeker, D. (2005), 'Generalized linear mixed models', *Encyclopedia of Statistics in Behavioral Science* .

Henderson, C. R. (1950), 'Estimation of genetic parameters', *Annals of Mathematical Statistics* **21**, 309 – 310.

Henderson, C. R., Kempthorne, O., Searle, S. R. and VonKrosigk, C. M. (1959), 'The estimation of enviromental and genetic trends from records subject to culling', *Biometrics* **15**, 192 – 218.

Ibrahim, J. G., Chen, M. and Sinha, D. (2001), *Bayesian Survival Analysis*, Springer.

Jack, A., Woodard, D., Hoffman, J. and O'Connell, M. (2007), *Bayesian Modeling with S-PLUS and the flexBayes Library*, Insightful Corporation.

Jackman, S. (2011), *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*, Department of Political Science, Stanford University, Stanford.

Jeffreys, H. (1961), *The Theory of Probability*, Oxford.

Johnson, V. E. (1996), 'On bayesian analysis of multirater ordinal data: An application to automated essay grading', *Journal of the american Statistical Association* **91**, 42 – 51.

Jordan, M. I. (2004), The exponential family and generalized linear models.

Jorgersen, B. (1987), 'Exponential dispersion models (with discussion)', *Journal of the Royal Statistical Association* **B. 49**, 150.

Karim, M. R. and Zeger, S. L. (1992), 'Generalized linear models with random effects; salamander mating revisited', *Biometrics* **48**, 631 – 644.

Kizilkaya, K. and Tempelman, R. J. (2006), 'Bayesian heteroskedastic generalized linear models for animal breeding applications', *Statistical Science* **21**, 35 – 51.

Kreft, I. and DeLeeuw, J. (1998), *Introducing Multilevel Modeling*, Sage.

Laird, N. M. and Ware, J. H. (1982), 'Random effects models for longitudinal data', *Biometrics* **38**, 963 – 974.

Leonard, T. (1975), 'A bayesian approach to the linear model with unequal variances', *Technometrics* **17**, 95 – 102.

Lindstrom, M. J. and Bates, D. M. (1990), 'Nonlinear mixed effects models for repeated measures data', *Biometrics* **46**, 673 – 687.

MacCullagh, P. and Nelder, J. A. (1996), *Generalized Linear Models*, Chapman and Hall.

Magidson, J. (1982), 'Some common pitfalls in causal analysis of categorical data', *Journal of Marketing Research* **19**, 461–471.

Magidson, J. (2004), 'Epidat 3.0 programa para análisis epidemiológico de datos tabulados', *Revista Española de Salud Pública* **78**(2), 277–280.

Martin, A. D., Quinn, K. M. and Park, J. H. (2011), 'MCMCpack: Markov chain monte carlo in R', *Journal of Statistical Software* **42**(9), 22.
URL: *http://www.jstatsoft.org/v42/i09/*

McCulloch, C. E. and Searle, S. R. (2001), *Generalized, Linear and Mixed Models*, Wiley.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equations of state calculations by fast computing machines', *Journal of Chemical Physics* **21**, 1087 – 1092.

Migon, H. S. and Gamerman, D. (1999), *Statistical Inference: An Integrated Approach*, Arnold.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3 edn, McGraw - Hill.

Mukhopadhyay, P. (1998), *Small Area Estimation in Survey Sampling*, Narosa Publishing House.

Nair, V. N. and Pregibon, D. (1988), 'Analysing dispersion effects from replicated factorial experiments', *Technometrics* **30**, 247 – 257.

Novick, M. R., Lewis, C. and Jackson, P. H. (1973), 'The estimation of proportions in m groups', *Psychometrika* **38**, 19 –46.

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*, Wiley.

Pfefferman, D. (2002), 'Small area estimation: New developments and directions', *International Statistic Review* **70**, 125 – 143.

Pham-Gia, T. and Turkkan, N. (1993), 'Bayesian analysis of the difference of two proportions', *Communications in Statistics: Theory and Methods* **22**(6), 1755–1771.

Pinhero, J. C. and Bates, D. M. (1995), 'Approximations to the log-likelihood function in the nonlinear mixed-effects model', *Journal of the Computational and Graphical Statistics* **4**, 12 – 35.

Platek, R., Rao, J. N. K., Sarndal, C. E. and Singh, M. P. (1987), *Small Area Statistics*, Wiley.

Platek, R. and Singh, M. P. (1986), *Small Area Statistics: Contributed Papers*, Laboratory for Research in Statistics and Probability.

Pope, J. L. (1984), *Investigaciòn de mercados. Guía maestra para el profesional*, Grupo editorial norma.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J. N. K. (1986), Synthetic estimators, spree and best model based predictors, *in* 'Proceedings of the Conference on Survey Research Methods in Agriculture', U. S. Department of Agriculture, pp. 1 – 16.

Rao, J. N. K. (2001), *Empirical Bayes and Likelihood Inference*, Springer, chapter EB and EBLUP in Small Area Estimation, pp. 33 – 43.

Rao, J. N. K. (2003), *Small Area Estimation*, Wiley.

Raudenbush, S. W. and Bryk, A. S. (2002), *Hierarchical Linear Models*, Thousand Oaks.

Robert, C. P. and Casella, G. (1999), *Monte Carlo Statistical Methods*, Springer.

Robert, C. P. and Casella, G. (2009), *Introducing Monte Carlo Methods with R*, Springer.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997), 'Weak convergence and optimal scaling of random walk metropolis algorithms', *Annals of Applied Probability* **7**, 110 – 120.

Roberts, G. O. and Rosenthal, J. S. (2001), 'Optimal scaling for various metropolis-hastings algorithms', *Statistical Science* **16**, 351 – 367.

Robinson, G. K. (1991), 'That blup is a good thing: The estimation of randon effects', *Statistical Science* **6**, 15 – 51.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge Series in Statistical and Probabilistic Mathematics.

S., F. and F., C.-N. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**(7), 799 – 815.

Sarndal, C. E., Swenson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.

SAS (2006), *Preliminary Capabilities for Bayesian Analysis in SAS/STAT Software*.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461 – 464.

Searle, S., Casella, G. and McCulloch, C. E. (1992), *Variance Components*, Wiley.

Searle, S. R. (1971), *Linear Models*, Wiley.

Sheiner, L. B. and Beal, S. L. (1985), 'Pharmacokinetic parameter estimates from several least squares procedures: Superiority of extended least squares', *Journal of Pharmacokinetics and Biopharmaceutics* **13**, 185 – 201.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, Wiley.

Smith, J. (2004), *Analysis of Failure and Survival Data*, CRC.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and VanderLinde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society* **B 64**, 583 – 639.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2004), *WinBUGS User Manual*.

Stiratelli, R., Laird, N. M. and Ware, J. H. (1984), 'Random effects models for serial observations with binary response', *Biometrics* **40**, 961 – 971.

Sudgen, R. A. and Smith, T. M. F. (1984), 'Ignorable and informative designs in survey sampling inference.', *Biometrika* **71**, 495–506.

Therneau, T. and Lumley, T. (2011), *survival: Survival analysis, including penalised likelihood.* R package version 2.36-5.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000), *Finite Population Sampling and Inference*, Wiley.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2001), *Finite Population Sampling and Inference: A Prediction Approach*, Wiley.

Verbeke, G. and Molenberghs, G. (2000), *Linear Models for Longitudinal Data*, Springer-Verlag.

Vonesh, E. F. and Carter, R. L. (1992), *Mixed effects nonlinear regression for unbalanced repeated mesures*, Vol. 48.

Vonesh, E. F. and Chinchilli, V. M. (1997), *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, Marcel Dekker.

Walker, S. (1996), 'An em algorithm for nonlinear random effects models', *Biometrics* **52**, 934 – 944.

West, M. (1985), *Bayesian Statistics 2*, Oxford University Press, chapter Generalized Linear Models: Outlier Acommodation, Scale Parameters and Prior distributions (with discussion), pp. 461 – 484.

Wikipedia (2011*a*), 'Hit — Wikipedia, the free encyclopedia'.

Wikipedia (2011*b*), 'Porcentaje de bateo. Wikipedia'.

Wolfinger, R. (1993), 'Laplaces approximation for non-linear mixed effects models', *Biometrika* **80**, 791 – 795.

Wolfinger, R. and Oconnell, M. (1993), 'Generalized linear mixed models: A pseudo-likelihood approach', *Journal of Statistical Computation and Simulation* **48**, 233 – 243.

Yee, T. W. (2012), *VGAM: Vector Generalized Linear and Additive Models.*, URL http://CRAN.R-project.org/package=VGAM. R package version 0.9-0.

Zacks, S. (1971), *The Theory of Statistical Inference*, Wiley.

Zeger, S. L. and Karim, M. R. (1991), 'Generalized linear models with random effects. a gibbs sampling approach', *Journal of the american Statistical Association* **86**, 79 – 86.

Zhang, H. and Gutiérrez, H. A. (2010), *Teoría estadística. Aplicación y métodos.*, Universidad Santo Tomás.

Zhao, Y., Staudenmayer, J., Coull, B. A. and Wand, M. P. (2006), 'General design bayesian generalized linear mixed models', *Statistical Science* **21**, 35 – 51.

Zhiyu, G., Bickel, P. J. and Rice, J. A. (2004), 'An approximate likelihood approach to nonlinear mixed effects models via spline', *Computational Statistics and Data Analysis* **46**, 747 – 776.

Zimmerman, D. L. and Nuñez, V. A. (2009), *Antedependence Models For Longitudinal Data*, CRC Press.