Modelos jerárquicos bayesianos

Martín Alcalde Navarro

Trabajo de fin de grado de Matemáticas
Universidad de Zaragoza

Directores del trabajo: Ana Carmen Cebrián,
Jorge Castillo-Mateo

27 de junio de 2022

Prólogo

En 1763, a título póstumo, se publicó el conocido Teorema de Bayes [1]. Este teorema supuso no solo el desarrollo de una potente herramienta de cálculo de probabilidades, sino la inauguración de una perspectiva nueva de la probabilidad, entendiéndola como medida de incertidumbre.

Esta perspectiva se prolonga de forma consecuente en la disciplina de la estadística matemática, dando forma a la estadística bayesiana. Entender lo incierto como aleatorio lleva a la interpretación de los parámetros desconocidos como variables aleatorias con una cierta distribución de probabilidad. El objetivo de la estadística bayesiana consiste en tratar de estimar la distribución *a posteriori* de estos parámetros tras obtener una muestra de la población estudiada y conjugarla con las premisas de las que se partía —distribución *a priori.* Esta visión, ya desde los inicios de la estadística clásica o frecuentista, despertó el rechazo de varios de sus principales exponentes como Ronald Fisher —véase Fisher, R., (1949) [4]. No obstante, el tiempo y la práctica parecen haber demostrado la poca justedad de sus afirmaciones. En los últimos años, la estadística bayesiana ha vivido un repunte significativo y en especial con el gran desarrollo que ha experimentado el estudio de métodos computacionales. En su favor, encontramos la gran flexibilidad que permite en la modelización y que exploraremos brevemente en los Capítulos 3 y 4, la sencilla interpretación de algunas de sus herramientas frente a sus homólogas frecuentistas —como se observa en los intervalos de credibilidad bayesianos y los intervalos de confianza frecuentista—, y, además, el hecho de que ha permitido dar salida a limitaciones, cada vez más evidentes para mayor parte de la comunidad científica, del paradigma clásico; por ejemplo, en los contrastes de hipótesis y p-valores —McShane, B. B. et al (2019) [8].

El paradigma bayesiano se ha convertido en el referente en varios escenarios diferentes. Entre ellos encontramos: el análisis del índice de paternidad; en problemas legales y juicios, por ejemplo, para desenmascarar la falacia del fiscal; e incluso los filtros de spam del correo electrónico, basados en lo que se conoce como filtros bayesianos, esto es, filtros que irán aprendiendo a decidir a partir de los datos que el usuario le va enseñando paulatinamente, de forma que la probabilidad de filtrar solamente el material indeseado aumenta.

Volviendo al terreno de la modelización estadística, en este trabajo se presentan, primero, los fundamentos de la estadística bayesiana como enfoque diferenciado, y, posteriormente, se centra en uno de sus pilares fundamentales: los modelos jerárquicos. Más precisamente, son cuatros los puntos principales que se incluyen en este trabajo, divididos en cuatro capítulos principales.

En el Capítulo 1, junto con la introducción al enfoque bayesiano, se recogen algunos apuntes sobre sus herramientas fundamentales, tales como la elección de la distribución a priori, la necesidad de los métodos MCMC o los estimadores Bayes e intervalos de credibilidad.

En cuanto al Capítulo 2, se desarrolla el cálculo de las distribuciones a posteriori de parámetros asociados a variables normales según diferentes casos —media desconocida y varianza desconocida, el caso opuesto y ambos parámetros desconocidos.

Ya en el Capítulo 3, se retoman las cuestiones más de concepto, pues se exponen las características principales de los modelos jerárquicos bayesianos. Su desarrollo se justifica por su capacidad para relaciones de dependencia, a la par de producir modelos más realistas al reconocer los parámetros que determinan la distribución a priori de los parámetros —llamados hiperparámetros— como desconocidos. Nuevamente, se desarrolla un ejemplo relativo a variables normales para ejemplificar los pasos habituales para la caracterización de las diferentes distribuciones a posteriori del modelo.

Finalmente, en el Capítulo 4, en una primera parte, se estudian algunos modelos de regresión ex-

plorando y haciendo hincapié, nuevamente, en la flexibilidad del análisis bayesiano. Esto se expresa, principalmente, en su capacidad para generalizar el modelo de regresión clásico a situaciones con datos correlados o heterocedásticos partiendo de cálculos sencillos. En la segunda parte del Capítulo 4, se propone un ejemplo de cómo la obtención de las distribuciones a posteriori conjugadas se pueden combinar con técnicas MCMC para ajustar un modelo bayesiano de forma eficiente. El ejemplo propuesto sobre las temperaturas medias del verano es meramente ilustrativos, ya que un ajuste óptimo requeriría un modelo más complicado con más términos para representar la variabilidad espacial, y queda fuera del objetivo de esta memoria.

Abstract

Two main points that distinguish Bayesian analysis from classical statitistics are the subjective way of comprehending probability as a measure of uncertainty and regarding the unknown parameter θ in a statistical model as a random variable. Therefore, θ has its own distribution.

Bayes's Theorem establishes that the parameters distributions depend on the quantity of *prior* information that we have available, which is reflected on the prior distribution or density of θ, denoted by $[\theta]$, and one sample $y = (y_1, \ldots, y_n)$ of the studied population. The joint distribution of y —usually called likelihood— depends on θ and is denoted by $[y|\theta]$. According to the introduced notation, Bayes's Theorem for absolutely continuous random variables asserts that

$$[\theta|y] = \frac{[\theta][y|\theta]}{[y]} = \frac{[\theta][y|\theta]}{\int_\Theta [\theta][y|\theta]\,\mathrm{d}\theta},$$

where $[\theta|y]$ is the posterior distribution or density and Θ is the support of the density of θ. For Bayesian inference, the posterior distribution is the main goal. It expresses all our knowledge about the parameter θ and, therefore, it is the distribution that we use to estimate.

In Chapter 1, besides of going over all these concepts, we also include a brief discussion on the different kinds of prior distributions, especially, conjugate prior distributions due to their advantages related to posterior distribution computations. Furthermore, we introduce the necessity of MCMC methods in Bayesian analysis and frequent tools that Bayesian analysis uses to summarize posterior distributions, such that Bayes estimators and credible intervals.

In Chapter 2, we focus on showing posterior distributions computations in detail for normal variables, that is, $y \sim \mathrm{N}(\mu, \sigma^2)$. The aim is to look into different case studies depending on which parameter is supposed to be unknown: μ, σ^2 or both, and making different choices about the prior distributions: non-informative prior distributions and conjugate prior distributions. These results will be frequently used throughout the following chapters. Besides, two important remarks are raised. One of them shows that the influence of the prior distribution in the posterior distribution is very slight if we have a substantial quantity of data, what justifies making use of non-informative prior information in such cases. The other remark is an actual defence of Bayesian analysis, since we will see that classical properties of frequentist statistics can be followed easily thanks to Bayes's Theorem, and without necessity of counting with a great amount of prior information.

As for Chapter 3, we first go back to a more conceptual framework and principal issues regarding hierarchical modeling are brought up. The usual scenes where hiercachical model shows up are models with clustered data as it allows us to establish a depency relation easily. We consider the hierarchical model given by normal data $y_{i,j} \sim \mathrm{N}(\theta_j, \sigma^2)$ clustered in J different groups. The parameters θ_j are assumed to be unknown and we consider them as a sample such that $\theta_j|\mu, \tau \sim \mathrm{N}(\mu, \tau^2)$ for all j with parameters —usually called hyperparameters— $\phi = (\mu, \tau)$ unkown too. As a result, hierarchical modeling is more realistic than usual modeling since it also takes the uncertainty associated to the hyperparameters into consideration. The prior information will not be $[\theta]$ any longer, but $[\theta|\phi] = [\theta|\mu, \tau]$ together with $[\phi] = [\mu, \tau]$. Thus, we include a brief discussion on what non-informative prior distribution for ϕ one can choose in a model like ours. Our main objetive will be to characterize the posterior conditional distribution of the parameters given the hyperparameters $[\theta|\phi, y]$ and the marginal posterior distributions of the hyperparameters $[\phi|y]$. We use this example to show the usual steps followed in these characterizations in Bayesian hierarchical models.

Finally, in Chapter 4, we collect some results related to different linear regression models. Given an explained variable y and explanatory variables $x_1, \ldots, x_k$, a linear regression model is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k + \varepsilon,$$

where $\beta = (\beta_0, \ldots, \beta_k)$ are the regression coefficients and ε is a normal variable that expresses the error. Different assumptions about this error term determine the complexity of the model. The classical regression model is the first model we will cope with. Our goal will be the posterior distributions of the regression coefficients and the variance. Afterwards, in order to insist on the flexibity Bayesian statistics provides, we will show that some models with correlated and heterocedastic data can be faced with similar computations to those that are developed for the classical regression model. In addition, we study the applications of hierarchical modeling in the field of regression models by including random effects. As an example of the use of the previous results, two simple models based on summer temperatures in la Comunidad Autónoma de Aragón are studied, so that we can show how combining the analytical derivations of the posterior distributions and MCMC methods leads us to a more efficient way to fit a Bayesian model. The given example related to summer temperatures is merely illustrative, given that an optimal adjustment would require a more complex model with specific terms to express the spatial variability, what is further from our purposes.

To obtain the desired marginal posterior distributions, we develop a Gibbs sampling algorithm. The `R` codes used to implement the approach, together with the results of the simulation —the marginal posterior densities, posterior expectations and credible intervals—, are provided in the Appendix.

Índice general

Prólogo III

Abstract V

1. Generalidades del análisis bayesiano 1
1.1. Introducción al análisis bayesiano . . . 1
1.2. Distribución a posteriori. Teorema de Bayes . . . 1
1.3. Distribuciones a priori. Consideraciones . . . 2
1.3.1. Distribuciones conjugadas . . . 2
1.3.2. Distribuciones a priori informativas, no informativas y débilmente informativas . 3
1.4. Cálculo de la distribución a posteriori. Métodos MCMC . . . 3
1.4.1. Método de *Gibbs sampling* . . . 3
1.5. Resumen de la distribución a posteriori . . . 4
1.5.1. Funciones de pérdida y estimadores Bayes . . . 4
1.5.2. Intervalos de credibilidad . . . 4

2. Análisis bayesiano en variables normalmente distribuidas 5
2.1. Modelos uniparamétricos . . . 5
2.1.1. Media desconocida y varianza conocida . . . 5
2.1.2. Media conocida y varianza desconocida . . . 6
2.2. Modelos multiparamétricos. Media y varianza desconocidas . . . 7
2.2.1. Distribución a posteriori conjunta con distribución a priori impropia . . . 7
2.2.2. Distribución a posteriori conjunta con distribución a priori conjugada . . . 9

3. Modelos jerárquicos 11
3.1. Derivación analítica de las distribuciones a posteriori . . . 12
3.2. Modelo jerárquico conjugado para verosimilitudes normales . . . 12
3.2.1. Distribuciones a posteriori del modelo jerárquico normal . . . 13
3.2.2. Distribución condicional y marginal a posteriori de los hiperparámetros . . . 14
3.2.3. Distribución a priori de los hiperparámetros . . . 15

4. Modelos de regresión lineal bayesianos 17
4.1. Modelos regresión . . . 17
4.2. Análisis bayesiano del modelo de regresión clásico . . . 17
4.3. Modelos con varianzas desiguales y correlaciones . . . 18
4.4. Modelos de regresión jerárquicos . . . 20
4.4.1. Modelo de regresión con efectos aleatorios . . . 20
4.5. Aplicación al análisis espacio-temporal de temperaturas medias en verano en una región 21
4.5.1. Modelo de regresión no jerárquico . . . 21
4.5.2. Modelo jerárquico . . . 23
4.6. Conclusiones finales . . . 25

A. Simulación en R **27**
A.1. Mapa de la Comunidad Autónoma de Aragón y alrededores 27
A.2. Modelo no jerárquico 27
A.2.1. Convergencia de las cadenas de Markov 28
A.2.2. Densidades a posteriori marginales 29
A.2.3. Resumen númerico. Esperanzas a posteriori e intervalos de credibilidad de los parámetros 30
A.3. Modelo jerárquico 31
A.3.1. Convergencia de las cadenas de Markov 31
A.3.2. Densidades a posteriori marginales 33
A.3.3. Resumen númerico. Esperanzas a posteriori e intervalos de credibilidad de los parámetros 34

B. Implementación del código de R **35**
B.1. Modelo no jerárquico 35
B.2. Modelo jerárquico 38

Capítulo 1

Generalidades del análisis bayesiano

1.1. Introducción al análisis bayesiano

La probabilidad como medida de incertidumbre es utilizada con frecuencia en el lenguaje cotidiano. Afirmaciones como «es cien por cien seguro que ocurrirá» expresan la creencia de un sujeto ante un fenómeno, reflejando una perspectiva condicionada por la información de la que un sujeto parte. En este sentido, podemos decir que la probabilidad puede entenderse de forma intuitiva como medida de incertidumbre. La estadística bayesiana es una rama de la estadística que se basa en esta interpretación de la probabilidad como medida de incertidumbre, en lugar de la interpretación frecuentista de la probabilidad, fundamento de la estadística clásica. Además, otra diferencia esencial entre ambos paradigma es la forma de considerar los parámetros desconocidos. Para la estadística clásica o frecuentista, tales parámetros se suponen valores fijos y uno de los objetivos princpales es obtener estimadores de esos parámetros. Por el contrario, la estadística bayesiana reconoce a los parámetros como variables aleatorias y el objetivo es caracterizar su distribución.

1.2. Distribución a posteriori. Teorema de Bayes

El objetivo principal de la estadística bayesiana es caracterizar la distribución de θ, el parámetro o vector de parámetros bajo estudio, incuyendo la información que proporciona una muestra observada y la información a priori que podemos tener sobre él. Esta distribución se denomina distribución a posteriori y se trata de la distribución del parámetro condicionada a la muestra observada. Esta distribución condicionada se calcula utilizando el Teorema de Bayes —Bayes, T. (1763) [1]— como se muestra en el siguiente teorema. El resultado se enuncia para variables absolutamente continuas, pero también es válido para variables discretas sustituyendo las densidades por probabilidades. A lo largo de todo el estudio denotaremos por $\Theta \subseteq \mathbb{R}^J$ al dominio paramétrico.

Teorema 1.1. *Sea $y = (y_1, \ldots, y_n)$ una muestra con densidad $[y|\theta]$ dependiente de un parámetro θ. Entonces, si suponemos que θ tiene densidad $[\theta]$, se tiene que su densidad a posteriori es*

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]} = \frac{[y|\theta][\theta]}{\int_\Theta [y|\theta][\theta]\,\mathrm{d}\theta}. \tag{1.1}$$

La densidad $[\theta]$ se denomina *densidad a priori*, ya que refleja la información sobre θ previa a la muestra y, y la densidad $[y|\theta]$ es la *verosimilitud* de la muestra, que satisface

$$[y|\theta] = \prod_{i=1}^{n} [y_i|\theta],$$

puesto que las variables $y_i|\theta$, $i = 1, \ldots, n$, se suponen independientes entre sí.

Dado que el objetivo es caracterizar la distribución a posteriori $[\theta|y]$, se puede considerar como constante todos los términos que no dependan de θ y se puede prescindir de ellos en el cálculo porque

su único cometido es corregir la densidad, de manera que su integral sobre el dominio paramétrico sea 1. Por este motivo, en el análisis bayesiano, es habitual utilizar la proporcionalidad en lugar de la igualdad y trabajar con el *kernel* de una distribución, que es, precisamente, la parte de la densidad dependiente de θ, en lugar de la distribución completa. De este modo, por ejemplo, el Teorema de Bayes resulta

$$[\theta|y] \propto [y|\theta][\theta]. \tag{1.2}$$

De esta última expresión, se deduce que la distribución a posteriori solo depende realmente de la verosimilitud de la muestra y y de la distribución a priori de θ. Notemos que la verosimilitud es clara de partida, en tanto que corresponde a la modelización de los datos que estudiamos, pero, en cuanto a la distribución a priori, es oportuno detenerse sobre algunas consideraciones alrededor de su elección.

1.3. Distribuciones a priori. Consideraciones

La distribución a priori representa la información que se tiene sobre el parámetro antes de observar la muestra. Se distinguen diferentes distribuciones, que son las distribuciones informativas, no informativas y débilmente informativas, según el grado de conocimiento que tengamos. Por otro lado, para el análisis bayesiano, son especialmente importantes por las simplificaciones de cálculo que ofrecen, las distribuciones a priori conjugadas puesto que, como veremos a continuación, proporcionan una distribución a posteriori que pertenece a una familia conocida.

1.3.1. Distribuciones conjugadas

Dada una verosimilitud $[y|\theta]$, se dice que la distribución a priori $[\theta]$ es conjugada si la distribución a posteriori $[\theta|y]$ continúa siendo de la misma familia que $[\theta]$, es decir, sus densidades tienen la misma forma funcional. La principal ventaja de las distribuciones conjugadas es que, al conservar la familia de distribuciones, reducen el problema de caracterización de la distribución a posteriori al de la estimación de los parámetros, además de que nos permite llegar a una distribución conocida explícita, lo que facilita su simulación.

Observación 1.2. Pese a no existir siempre, para muchas verosimilitudes puede encontrarse una distribución a priori conjugada. De hecho, siempre existe para datos procedentes de la familia exponencial —que engloba a la mayoría de distribuciones habituales— como demostraremos a continuación. Recordemos que una *familia exponencial* es aquella cuya distribución $[y_i|\theta]$ puede escribirse en términos de funciones f, g, φ y h medibles satisfaciendo que

$$[y_i|\theta] = f(y_i)g(\theta)\exp\left(\varphi(\theta)^t h(y_i)\right), \quad i = 1, \ldots, n.$$

Así, bajo la hipótesis de independencia de los datos $y_i|\theta$, tenemos que la verosimilitud es

$$[y|\theta] = \left(\prod_{i=1}^{n} f(y_i)\right) g(\theta)^n \exp\left(\varphi(\theta)^t \sum_{i=1}^{n} h(y_i)\right).$$

Y si escogemos

$$[\theta] = g(\theta)^k \exp\left(\varphi(\theta)^t u\right) \tag{1.3}$$

para $k \geq 0$ y u una tupla de la dimensión adecuada, entonces,

$$[\theta|y] \propto g(\theta)^{n+k} \exp\left(\varphi(\theta)^t \left(u + \sum_{i=1}^{n} h(y_i)\right)\right).$$

Lo que prueba que, en efecto, la familia (1.3) es conjugada para la verosimilitud dada.

1.3.2. Distribuciones a priori informativas, no informativas y débilmente informativas

Una de las mayores virtudes del análisis bayesiano es que nos permite introducir, como punto de partida, los resultados de otras investigaciones, incluyendo, de este modo, una considerable cantidad de información añadida junto con nuestros datos. Esta es la atribución de las distribuciones informativas, que pueden tener bastante influencia en la distribución a posteriori. Por este motivo, en caso de carecer de información previa, la imposición de una distribución que condicione considerablemente la distribución a posteriori no está justificada y podría inducir a un análisis poco fiable. Por ello, en este caso, la mejor elección es tomar distribuciones *no informativas* y *débilmente informativas*, esto es, distribuciones que, en la mayor medida de lo posible, no incluyan regiones especialmente probables, de manera que la distribución a posteriori no se incline hacia ellas. Un ejemplo sencillo de distribución no informativa es la distribución uniforme. En la práctica, las distribuciones no informativas pueden ser *impropias*, i.e., con densidad no integrable, lo que no es ningún problema siempre que la distribución a posteriori sea propia. Un ejemplo de este tipo de distribuciones a priori es la ley uniforme en $\mathbb{R}$. El hecho de que las distribuciones a priori impropias puedan generar distribución a posteriori impropias da paso a la necesidad de las distribuciones débilmente informativas, que tratan de garantizar la integrabilidad de la densidad a posteriori, pero respetando el protagonismo de la verosimilitud, como pueden ser aquellas distribuciones con varianza muy grande aproximando, así, a una distribución uniforme.

1.4. Cálculo de la distribución a posteriori. Métodos MCMC

Ya se ha expuesto que, si se cuenta con una distribución conjugada, el cálculo de la distribución a posteriori se simplifica drásticamente, pudiendo obtener una familia conocida. Sin embargo, en la práctica, lo general, y más aún cuando el número de parámetros J es elevado, es que la expresión de la distribución (1.2) sea demasiado complicada para realizar cálculos analíticos a partir de ella, lo que impide caracterizarla o resumirla. En esta situación, se hace obligatorio el uso de métodos computaciones, principalmente métodos de Monte Carlo basados en cadenas de Markov, más conocidos como métodos MCMC. Existen múltiples métodos de este tipo y aunque no es objetivo de este trabajo la descripción de los mismos, se expone brevemente el método de *Gibbs sampling*, ya que se utilizará en la aplicación desarrollada en el Capítulo 4. Una revisión detallada de estos métodos se puede encontrar en Robert, C. y Casella, G. (1992) [9] y (1998) [10].

1.4.1. Método de *Gibbs sampling*

Sea $\theta = (\theta_1, \ldots, \theta_J)$ un vector de parámetros. Supongamos que podemos simular a partir de las *distribuciones a posteriori completamente condicionadas*, i.e.,

$$[\theta_j | \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_J, y], \quad j = 1, \ldots, J.$$

Entonces, el algoritmo *Gibbs sampling* viene dado por la siguiente transición de $\theta^{(t)}$ a $\theta^{(t+1)}$: dada $(\theta_1^{(t)}, \ldots, \theta_J^{(t)})$, se generan

$$\begin{aligned}
\theta_1^{(t+1)} &\sim \left[\theta_1 \,\middle|\, \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_J^{(t)}, y\right], \\
\theta_2^{(t+1)} &\sim \left[\theta_2 \,\middle|\, \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_J^{(t)}, y\right], \\
&\vdots \\
\theta_J^{(t+1)} &\sim \left[\theta_J \,\middle|\, \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{J-1}^{(t+1)}, y\right].
\end{aligned}$$

El algoritmo, así, fabrica una cadena de Markov $(\theta^{(t)})_{t \geq 1}$ y asegura que la distribución estacionaria de las observaciones así generadas es la distribución a posteriori conjunta $[\theta_1, \ldots, \theta_J | y]$. Además, si consideramos las componentes del vector de forma separada, la distribución estacionaria de las observaciones

$(\theta_j^{(t)})_{t\geq 1}$ es la distribución marginal a posteriori $[\theta_j|y]$ para $j = 1,\ldots,J$. En la aplicación del Capítulo 4, obtendremos las densidades a posteriori completamente condicionadas analíticas, de modo que la aplicación del algoritmo de Gibbs nos ofrecerá una estimación mucho más rápida, ya que no será necesario combinarlo con otros algoritmos MCMC aproximar las densidades.

1.5. Resumen de la distribución a posteriori

Una vez caracterizada la distribución a posteriori, pueden calcularse diferentes medidas que resumen en un valor algunas de sus características principales. Algunas de ellas especialmente habituales son los *estimadores Bayes* y *los intervalos de credibilidad*.

1.5.1. Funciones de pérdida y estimadores Bayes

Definición 1.3. Una *función de pérdida* es una función $\ell : \Theta \times \Theta \to [0,+\infty)$ medible tal que $\ell(\theta,\theta) = 0$ para todo θ. En esta situación, se define el *estimador Bayes* como el valor $\hat{\theta} \in \Theta$, si existe, tal que la esperanza a posteriori $\mathrm{E}(\ell(\hat{\theta},\theta)|y)$ toma valor mínimo.

Es una comprobación sencilla, demostrar que, para el caso $\ell(x,y) = (x-y)^2$ —denominada función de pérdida cuadrática—, el estimador Bayes es la esperanza a posteriori, esto es, $\hat{\theta} = \mathrm{E}(\theta|y)$.

1.5.2. Intervalos de credibilidad

Definamos, a continuación, los intervalos de credibilidad bayesianos.

Definición 1.4. Dada una distribución a posteriori $\theta|y$ y $\alpha \in (0,1)$, se denomina *intervalo de credibilidad* o *intervalo a posteriori a nivel* $1-\alpha$ a un intervalo de la forma $[(\theta|y)_{\alpha_1},(\theta|y)_{1-\alpha_2}]$, donde $(\theta|y)_\alpha$ denota al cuantil α de la distribución a posteriori $\theta|y$ y $\alpha_1,\alpha_2 > 0$ tales que $\alpha_1+\alpha_2 = \alpha$. Además, si $\alpha_1 = \alpha_2 = \alpha/2$, el intervalo se dirá también *centrado.*

Los intervalos de credibilidad son la versión bayesiana de los intervalos de confianza frecuentista. Algo especialmente notorio de estos intervalos es que son muy sencillos de estimar, siendo suficiente con tomar cuantiles muestrales. Además, su interpretación es más intuitiva. Al considerar los parámetros como variables aleatorias, el nivel de confianza $1-\alpha$ del intervalo de credibilidad es la probabilidad de que el parámetro pertenezca a dicho intervalo.

Capítulo 2

Análisis bayesiano en variables normalmente distribuidas

Procedemos a ilustrar los cálculos y distribuciones características del análisis bayesiano para el caso particular de verosimilitudes normales. Sea $y = (y_1, \ldots, y_n)$ una muestra aleatoria simple con $[y_i|\mu, \sigma^2] = \mathrm{N}(y_i|\mu, \sigma^2)$ para $i = 1, \ldots, n$, entonces

$$[y|\mu, \sigma^2] = \prod_{i=1}^{n} \mathrm{N}(y_i|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right), \tag{2.1}$$

donde μ es la *media* de la distribución y σ^2 la *varianza*. La distribución posee dos parámetros, por lo que se puede o bien suponer uno de ellos conocido y estimar el otro, o bien considerar el vector de parámetros (μ, σ^2). Dado que este último caso exige el desarrollo del primero, estudiamos primero dos modelos uniparamétricos para μ y σ^2 respectivamente partiendo de distribuciones a priori conjugadas. Los fundamentos teóricos para este capítulo y el siguiente pueden consultarse en Gelman, A. et al (2021) [5].

2.1. Modelos uniparamétricos

2.1.1. Media desconocida y varianza conocida

Sea y una muestra con distribución normal de parámetros μ y σ^2 —como en (2.1)— tal que $\sigma^2 > 0$ es conocido. Por la Observación 1.2, una distribución conjugada para μ es

$$[\mu] = \mathrm{N}(\mu|\mu_0, \tau_0^2) \propto \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right).$$

En efecto, por el Teorema de Bayes —véase la ecuación (1.2)—,

$$\begin{aligned}
[\mu|y] &\propto [\mu][y|\mu] \\
&\propto \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)\mu^2 - 2\mu\left(\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}\right)\right)\right).
\end{aligned} \tag{2.2}$$

Y denotando a los términos

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{y} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \tag{2.3}$$

de (2.2), se sigue que

$$\mu|y \sim \mathrm{N}(\mu_1, \tau_1^2). \tag{2.4}$$

Observación 2.1. a) La expresión de $1/\tau_1^2$ en (2.3) demuestra que la relación de varianzas entre distribución a priori, a posteriori y la verosimilitud viene dada por la suma de sus inversas. Como consecuencia, para el análisis bayesiano, se prefiere la parametrización de la distribución normal en términos de $1/\sigma^2$, que recibe el nombre de *precisión*. En lo que a las medias se refiere, en (2.3), la media a posteriori μ_1 aparece como una media ponderada de la media a priori y la media muestral, cuyos pesos son las precisiones de sendas distribuciones.

b) De estas relaciones, puede deducirse el comportamiento asintótico de la distribución a posteriori, de manera que, para n suficientemente grande,

$$[\mu|y] \approx \mathrm{N}(\mu|\bar{y}, \sigma^2/n);$$

puesto que, como puede comprobarse fácilmente,

$$\left.\frac{\mu-\bar{y}}{\sigma/\sqrt{n}}\right| y \xrightarrow{(\mathscr{L})} \mathrm{N}(0,1).$$

En esta distribución asintótica, los parámetros de las distribución a priori —denominados *hiperparámetros*— ya no intervienen; concluyendo que, cuando el número de datos es suficientemente grande, la distribución a posteriori no depende asintóticamente de la elección de la distribución a priori. Esta observación justifica el uso de distribuciones a priori no informativas en la práctica. En efecto, si consideramos la distribución a priori no informativa impropia $[\mu] \propto 1$, que se trata de la distribución uniforme sobre $\mathbb{R}$, es inmediato comprobar por (2.1), que

$$[\mu|y] \propto \exp\left(-\frac{1}{2\sigma^2}(n\mu^2 - 2n\mu\bar{y})\right) \propto \mathrm{N}(\mu|\bar{y}, \sigma^2/n),$$

por lo que no hay diferencia con respecto al caso conjugado si el volumen de datos es suficientemente grande.

2.1.2. Media conocida y varianza desconocida

Por la Observación 2.1, en lugar de trabajar con la distribución de la varianza, es habitual proponer una distribución a priori para la precisión, o equivalentemente, una distribución *inversa* para σ^2.

Definición 2.2. Sean $\nu_0 \in \mathbb{N}$ y $s_0 > 0$, se dice que el parámetro θ tiene *distribución inversa-χ^2* con ν_0 grados de libertad y parámetro de escala s_0 si $\nu_0 s_0^2/\theta \sim \chi^2_{\nu_0}$, es decir,

$$[\theta] = \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} s_0^{\nu_0} \theta^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0 s_0^2}{2\theta}\right) \propto \theta^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0 s_0^2}{2\theta}\right), \quad \theta > 0.$$

La distribución inversa-χ^2 para σ^2 es conjugada para la verosimilitud normal si μ es conocido. Efectivamente, si y es la verosimilitud normal de (2.1) con media $\mu \in \mathbb{R}$ conocida y $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2)$,

$$\begin{aligned}[\sigma^2|y] &\propto [\sigma^2][y|\sigma^2] \\ &\propto (\sigma^2)^{-((\nu_0+n)/2+1)} \exp\left(-\frac{1}{2\sigma^2}\left(\nu_0 s_0^2 + \sum_{i=1}^n (y_i-\mu)^2\right)\right),\end{aligned}$$

que es, por la Definición 2.2, el *kernel* de la distribución

$$[\sigma^2|y] = \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0 s_0^2 + n s_\mu^2}{\nu_0 + n}\right), \tag{2.5}$$

donde $s_\mu^2 = \sum_{i=1}^n (y_i - \mu)^2/n$.

Observación 2.3. a) Notemos que la distribución de (2.5) cuenta, por un lado, con el cuadrado del parámetro de escala igual a una media ponderada por los grados de libertad del cuadrado del parámetro de escala a priori, s_0^2, y la varianza muestral de los datos respecto de μ, s_μ^2; y, por otro, con grados de libertad igual a la suma de los grados de libertad a priori y el número de datos.

b) La distribución a priori conjugada escogida es una distribución impropia si $\nu_0 = 0$, ya que

$$[\sigma^2] \propto \frac{1}{\sigma^2} \tag{2.6}$$

no es integrable sobre $(0, +\infty)$. Sin embargo, por la distribución (2.5), es claro que la distribución a posteriori continúa siendo propia, por lo que (2.6) es una distribución a priori no informativa *válida*.

2.2. Modelos multiparamétricos. Media y varianza desconocidas

Tras haber desarrollado los modelos uniparamétricos, es más sencillo abordar el problema de caracterizar la distribución a posteriori conjunta de la media y varianza, así como las distribuciones condicionales y marginales. En concreto, utilizaremos las distribuciones de μ dado σ^2 desarrolladas en la Sección 2.1.1, tanto para obtener la distribución conjunta de (μ, σ^2) como la a priori y a posteriori. Por ejemplo, en cuanto a la distribución a priori, notemos que la distribución conjunta del parámetro (μ, σ^2) puede expresarse como

$$[\mu, \sigma^2] = [\mu|\sigma^2][\sigma^2],$$

donde la distribución $[\mu|\sigma^2]$ se corresponde con la distribución a priori $[\mu]$ especificada en la sección ya mencionada —ya que se trata de la distribución de μ supuesto σ^2 dado o *conocido*. Esto significa que podemos escoger las mismas distribuciones para $\mu|\sigma^2$ y σ^2 que en la Sección 2.1 —tanto las distribuciones conjugadas como las impropias no informativas— resolviendo de forma inmediata el problema de la elección de la distribución conjunta a priori.

Observación 2.4. Antes de proceder con las demostraciones, introducimos otra forma de expresar la verosimilitud normal que será útil en varias ocasiones. Dada y una muestra con distribución normal de parámetros μ y σ^2, entonces, la verosimilitud puede expresarse a través de los estadísticos suficientes s^2 —que es cuasivarianza muestral de y— y la media $\bar{y}$, puesto que

$$\begin{aligned}
[y|\mu, \sigma^2] &\propto (\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\
&= (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right)\right) \\
&= (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y} - \mu)^2)\right).
\end{aligned} \tag{2.7}$$

Con esta observación, damos paso a la caracterización de las distribuciones a posteriori en ambos casos.

2.2.1. Distribución a posteriori conjunta con distribución a priori impropia

Comenzando por las distribuciones impropias de las Observaciones 2.1 y 2.3, consideramos

$$[\mu|\sigma^2] \propto 1, \quad [\sigma^2] \propto (\sigma^2)^{-1},$$

esto es,

$$[\mu, \sigma^2] \propto (\sigma^2)^{-1}.$$

Así, utilizando la verosimilitud (2.7) y el Teorema de Bayes, la distribución a posteriori resulta

$$\begin{aligned}[\mu,\sigma^2|y] &\propto [\mu,\sigma^2][y|\mu,\sigma^2] \\ &= (\sigma^2)^{-(n/2+1)} \exp\left(-\frac{1}{2\sigma^2}((n-1)s^2+n(\bar{y}-\mu)^2)\right). \end{aligned} \tag{2.8}$$

Al contrario de lo que sucedía en los modelos uniparamétricos, la distribución aquí obtenida no corresponde con ninguna distribución conocida. Sin embargo, la caracterización de la distribución conjunta no es imprescindible en tanto que puede describirse por medio de distribuciones condicionales y marginales a posteriori.

- *Distribución condicional a posteriori* $[\mu|\sigma^2,y]$. Por las observaciones hechas al comienzo de la sección, $[\mu|\sigma^2,y]$ se corresponde con lo que, en la Observación 2.3, era la distribución a posteriori de μ, luego

$$\mu|\sigma^2,y \sim \mathrm{N}(\bar{y},\sigma^2/n). \tag{2.9}$$

- *Distribución marginal a posteriori* $[\sigma^2|y]$. Al contar la distribución conjunta a posteriori, sabemos que

$$[\sigma^2|y] = \int_{-\infty}^{+\infty} [\mu,\sigma^2|y]\,\mathrm{d}\mu.$$

  Luego, utilizando (2.8),

$$[\sigma^2|y] \propto (\sigma^2)^{-(n/2+1)} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2\right)\mathrm{d}\mu,$$

  donde

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2/n}(\bar{y}-\mu)^2\right)\mathrm{d}\mu = \sqrt{2\pi\sigma^2/n},$$

  puesto que el integrando se trata del *kernel* de la distribución $\mathrm{N}(\mu|\bar{y},\sigma^2/n)$. En consecuencia,

$$[\sigma^2|y] \propto (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \propto \text{Inv-}\chi^2(\sigma^2|n-1,s^2). \tag{2.10}$$

  Y obtenemos una distribución inversa-χ^2 de parámetros $n-1$ y s^2.

A partir de estos resultados, es inmediato caracterizar la distribución a posteriori conjunta como producto de (2.9) y (2.10). Sin embargo, por el especial interés que posee el parámetro μ —que es, generalmente, sobre el que se desea hacer inferencia—, también es interesante calcular su distribución marginal a posteriori.

- *Distribución marginal a posteriori* $[\mu|y]$. Utilizando el mismo procedimiento, sabemos que

$$[\mu|y] = \int_{0}^{+\infty} [\mu,\sigma^2|y]\,\mathrm{d}\sigma^2.$$

  Para ello, proponemos el cambio de variable

$$z = \frac{(n-1)s^2+n(\mu-\bar{y})^2}{2\sigma^2},$$

  con el que la integral se transforma en

$$((n-1)s^2+n(\mu-\bar{y})^2)^{-n/2} \int_0^{+\infty} z^{n/2-1}\exp(-z)\,\mathrm{d}z = ((n-1)s^2+n(\mu-\bar{y})^2)^{-n/2}\Gamma(n/2).$$

Por tanto,

$$\begin{aligned}[\mu|y] &\propto ((n-1)s^2 + n(\mu-\bar{y})^2)^{-n/2} \\ &\propto \left(1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right)^{-n/2} \\ &\propto t_{n-1}(\mu|\bar{y}, s^2/n), \end{aligned} \tag{2.11}$$

que se trata de una distribución *t de Student* no centrada de media $\bar{y}$ y parámetro de escala $s/\sqrt{n}$ de $n-1$ grados de libertad.

Observación 2.5. Las distribuciones (2.10) y (2.11) pueden escribirse también como

$$\left.\frac{(n-1)s^2}{\sigma^2}\right| y \sim \chi^2_{n-1} \quad \text{y} \quad \left.\frac{\mu-\bar{y}}{\sqrt{s^2/n}}\right| y \sim t_{n-1},$$

respectivamente. Esto establece una conexión muy interesante entre el paradigma bayesiano y el frecuentista. Se observa que, en este caso, se deducen resultados análogos al Teorema de Fisher clásico desde el enfoque bayesiano suponiendo que no contábamos con información a priori. Esta observación pone de relieve las ventajas del análisis bayesiano: notemos, primero, que, a partir de unos cálculos sencillos, hemos llegado a un resultado fundamental para el enfoque frecuentista; pero, más aún, obsérvese que si nos hubiésemos apoyado en una distribución a priori informativa, con un procedimiento similar, podríamos haber obtenido propiedades que desde la estadística clásica no serían evidentes.

2.2.2. Distribución a posteriori conjunta con distribución a priori conjugada

Veamos las densidades de los parámetros en el caso con distribución a priori conjugada. De acuerdo a las distribuciones a priori conjugadas de las Secciones 2.1.1 y 2.1.2, escogemos la distribución conjunta a priori $[\mu, \sigma^2]$ tal que

$$\begin{aligned}\mu|\sigma^2 &\sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0), \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2).\end{aligned}$$

Es decir,

$$[\mu, \sigma^2] \propto (\sigma^2)^{-(\nu_0+1)/2-1} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2)\right). \tag{2.12}$$

Definición 2.6. Sean $\mu_0 \in \mathbb{R}$, $\sigma^2, \kappa_0 > 0$ y $\nu_0 \in \mathbb{N}$. A la distribución (2.12) se la denomina *distribución normal-inversa-χ^2* de parámetros $\mu_0, \sigma_0^2, \kappa_0$ y ν_0. La denotaremos como

$$(\mu, \sigma^2) \sim \text{N-Inv-}\chi^2(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \sigma_0^2).$$

Esta distribución es conjugada para la verosimilitud normal. Más concretamente, se cumple

$$\mu, \sigma^2|y \sim \text{N-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2), \tag{2.13}$$

donde

$$\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}, \tag{2.14}$$

$$\kappa_n = \kappa_0 + n, \tag{2.15}$$

$$\nu_n = \nu_0 + n, \tag{2.16}$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2. \tag{2.17}$$

En efecto, como comprobaremos a continuación, si consideramos la verosimilitud expresada en términos de s^2 y $\bar{y}$ —véase (2.7)—,

$$[\mu, \sigma^2|y] \propto (\sigma^2)^{-(\nu_0+n+1)/2-1} \exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2 + (n-1)s^2 + n(\bar{y}-\mu)^2\right)\right). \quad (2.18)$$

Y, gracias a la ecuación (2.17),

$$\nu_0\sigma_0^2 + (n-1)s^2 = \nu_n\sigma_n^2 - \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2,$$

por lo que, sustituyendo en la expresión (2.18), se obtiene que

$$[\mu, \sigma^2|y] \propto (\sigma^2)^{-(\nu_n+1)/2-1} \exp\left(-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2 - \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2 + \kappa_0(\mu_0-\mu)^2 + n(\bar{y}-\mu)^2\right)\right). \quad (2.19)$$

Por otro lado, es una simple comprobación ver que

$$\kappa_0(\mu-\mu_0)^2 + n(\bar{y}-\mu)^2 = (\kappa_0+n)\left(\mu - \frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}\right)^2 + \kappa_0\mu_0 + n\bar{y} - \frac{(\kappa_0\mu_0+n\bar{y})^2}{\kappa_0+n}.$$

Y esta última expresión es

$$\kappa_n(\mu-\mu_n)^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2,$$

por las definiciones (2.14) y (2.15) y agrupando el resto de sumandos. Luego, sustituyendo en (2.19), se tiene (2.13).

A diferencia del caso anterior, esta vez sí que es conocida la distribución conjunta a posteriori, siendo esta, precisamente, la ventaja de trabajar con distribuciones a priori conjugadas. Además, este resultado es también interesante porque, en él, se observa ese compromiso particular de la distribución normal que ya se comentó en el estudio de modelos uniparamétricos. Nos referimos aquí a μ_n, que vuelve a ser un promedio ponderado entre la media a priori y la media muestral de los datos, y a σ_n^2, que es la suma de las *incertidumbres* a priori y muestral, añadiéndose un sumando más por la diferencia de las medias μ_0 y $\bar{y}$.

En cuanto a las distribuciones condicionales y marginales, utilizando los mismos argumentos que en el caso de distribución a priori impropia, puede probarse que tales distribuciones son las siguientes.

- *Distribución condicional a posteriori* $[\mu|\sigma^2, y]$. Simplemente, aplicando (2.4),

$$\mu|\sigma^2, y \sim \mathrm{N}\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}\right) = \mathrm{N}(\mu_n, \sigma^2/\kappa_n).$$

- *Distribución marginal a posteriori* $[\sigma^2|y]$. Es suficiente con calcular la integral de (2.13) respecto de μ, resultando

$$\sigma^2|y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2).$$

- *Distribución marginal a posteriori* $[\mu|y]$. Integrando (2.13) respecto de σ^2 y haciendo uso del mismo cambio de variable que en (2.11), se tiene

$$[\mu|y] \propto \left(1 + \frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2} \propto t_{\nu_n}(\mu|\mu_n, \sigma^2/\kappa_n).$$

Capítulo 3

Modelos jerárquicos

En estadística, es frecuente el tratamiento de modelos que, por la estructura del problema, contengan varios parámetros $\theta_1, \ldots, \theta_J$ que son dependientes entre sí. En este capítulo, consideraremos una estructura *jerárquica*, lo que, en esencia, significa abordar un modelo con las siguientes propiedades. En primer lugar, consideraremos a los parámetros θ_j como observaciones de una distribución a priori $[\theta|\phi]$ dependiente de algún hiperparámetro ϕ, que suponemos desconocido. En segundo lugar, una hipótesis básica de un modelo jerárquico es que la distribución de y solo depende de ϕ a través de θ, es decir, $[y|\phi,\theta] = [y|\theta]$.

En este capítulo, consideraremos datos $y_{i,j}$ con distribución normal $\mathrm{N}(\theta_j, \sigma^2)$, para $j = 1, \ldots, J$ e $i = 1, \ldots, n_j$, y supondremos que $\theta_j \sim \mathrm{N}(\mu, \tau^2)$. Gráficamente, la estructura de dependencia jerárquica se refleja en la Figura 3.1.

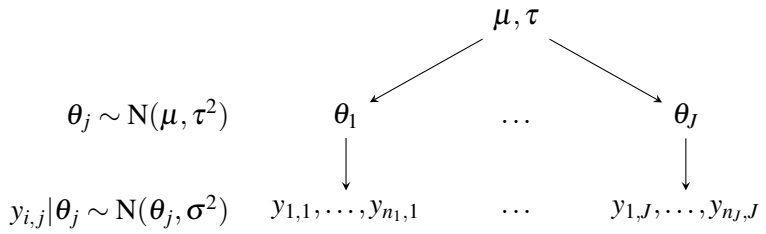


Figura 3.1: Modelo jerárquico para variables normales.

El potencial de los modelos jerárquicos se basa en que permite considerar parámetros relacionados entre sí, pero no iguales. Por ejemplo, en el modelo jerárquico anterior, los parámetros θ_j son las medias de J grupos diferentes. En un modelo no jerárquico podríamos considerar que esas medias son todas iguales, o bien que son diferentes e independientes. La primera opción es muy restrictiva, ya que impone que todas las observaciones compartan la misma media, lo que es una hipótesis poco oportuna en varias situaciones. En cuanto a la segunda opción, pese a ser más flexible, no permite capturar las posibles, y probables, relaciones de dependencia entre las medias de diferentes grupos. Un modelo jerárquico nos permite representar un rango de situaciones mucho más general, que incluye los dos casos anteriores como particulares. Además, la estructura jerárquica consigue evitar problemas de sobreajuste, al permitir una gran flexibilidad en el modelo con un número reducido de parámetros.

Otra ventaja del modelo jerárquico es que ofrece la posibilidad de cuantificar la incertidumbre asociada a la estimación del modelo de una forma más realista, puesto que recoge la incertidumbre asociada a los parámetros θ_j y a los correspondientes hiperparámetros. Esta aproximación implica que debemos asignar una distribución a priori al vector (ϕ, θ), que puede expresarse como $[\phi,\theta] = [\theta|\phi][\phi]$. Y la distribución conjunta a posteriori será $[\phi,\theta|y] \propto [y|\phi,\theta][\phi,\theta] = [y|\theta][\phi,\theta]$, donde la última igualdad se sigue de que la verosimilitud y solo depende de ϕ a través de θ. A continuación, presentamos un modelo jerárquico centrándonos, nuevamente, en variables normales. Previamente introducimos el procedimien-

to habitual para el cálculo de las distribuciones a posteriori en modelos jerárquicos.

3.1. Derivación analítica de las distribuciones a posteriori

La derivación analítica se resume en tres pasos:

I) Escribir la distribución conjunta a posteriori, $[\phi, \theta|y]$, como proporcional al producto $[\phi][\theta|\phi][y|\theta]$.

II) Determinar analíticamente la distribución condicional a posteriori $[\theta|\phi, y]$ como función de ϕ.

III) Calcular la distribución marginal a posteriori $[\phi|y]$.

Los pasos II) y III) proporcionan un procedimiento sencillo para obtener muestras de la distribución a posteriori, que, con frecuencia, no puede utilizarse en la práctica por su complejidad. El paso I) es inmediato, ya que se corresponde con el paso habitual de cálculo de la distribución a posteriori ya tratado. En cuanto al paso II), en el caso particular de familias conjugadas, es especialmente sencillo, ya que

$$[y|\theta] = \prod_{j=1}^{J} \prod_{i=1}^{n_j} [y_{i,j}|\theta_j] = \prod_{j=1}^{J} [y_j|\theta_j],$$

donde $[y_j|\theta_j]$ representa la verosimilitud asociada al grupo j con datos $y_{1,j}, \ldots, y_{n_j,j}$ para cada $j = 1, \ldots, J$. Así, es claro que

$$[\theta|\phi, y] \propto [y|\theta][\theta|\phi] = \prod_{j=1}^{J} [y_j|\theta_j] \prod_{j=1}^{J} [\theta_j|\phi] \propto \prod_{j=1}^{J} [\theta_j|\phi, y_{.j}],$$

por lo que la distribución a posteriori condicional se calcula como producto de las distribuciones a posteriori condicionales de las componentes.

Por último, acerca del paso III), cabe señalar que existen dos posibles vías —ambas basadas en la distribución conjunta a posteriori del paso I).

a) Por un lado, el proceso habitual de integración de la distribución conjunta, esto es,

$$[\phi|y] = \int_{\Theta} [\theta, \phi|y] \, \mathrm{d}\theta.$$

b) Y, por otro, en algunos casos como con variables normales, puede ser útil la propia definición de la densidad condicionada $[\theta|\phi, y]$, que permite expresar $[\phi|y]$ como

$$[\phi|y] = \frac{[\theta, \phi|y]}{[\theta|\phi, y]}.$$

Un ejemplo de aplicación de este procedimiento se muestra en la Sección 3.2.1.

3.2. Modelo jerárquico conjugado para verosimilitudes normales

Consideramos un modelo en el que partimos de $J \in \mathbb{N}$ grupos de manera que los datos observados en cada uno de ellos, que denotaremos como $y_{i,j}$, son normalmente distribuidos. Cada una de estas distribuciones normales tendrá una media θ_j distinta y desconocida y varianza σ^2 común a todos ellos y conocida. En suma, si $\theta = (\theta_1, \ldots, \theta_J)$ es el vector de las medias, partimos de la verosimilitud $[y|\theta] = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}(y_{i,j}|\theta_j, \sigma^2)$, donde n_j es el número de datos del grupo j.

Con el objetivo de simplificar la notación, definimos

$$\overline{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j}, \quad j = 1, \ldots, J.$$

Notemos que, como, para j fijo, los datos $y_{1,j}, y_{2,j}, \ldots, y_{n_j,j}$ son normales independientes e idénticamente distribuidos, se tiene que

$$[\overline{y}_{.j}|\theta_j] = \mathrm{N}(\overline{y}_{.j}|\theta_j, \sigma_j^2), \tag{3.1}$$

con $\sigma_j^2 = \sigma^2/n_j$. Con estas elecciones, la expresión de la verosimilitud se simplifica, ya que $[y|\theta]$ es proporcional, respecto de θ, a $\prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j}|\theta_j, \sigma_j^2)$. Utilizaremos esta verosimilitud en el resto del capítulo.

3.2.1. Distribuciones a posteriori del modelo jerárquico normal

Supongamos que nuestras medias θ_j tienen distribución normal de media μ y varianza τ^2 para todo $j = 1,\dots,J$, y que son independientes condicionadas a estos hiperparámetros. Es claro, por tanto, que estas medias son intercambiables y su distribución conjunta es

$$[\theta] = [\theta_1,\dots,\theta_J] = \iint_{\mathbb{R}\times(0,+\infty)} \left(\prod_{j=1}^{J}[\theta_j|\mu,\tau]\right)[\mu,\tau]\,\mathrm{d}\mu\,\mathrm{d}\tau.$$

Esta distribución podría utilizarse para determinar la distribución a posteriori $[\theta|y]$, pero, en la práctica, en los modelos jerárquicos, esta no es la distribución de interés. En este caso, caracterizaremos la distribución a posteriori de las medias dados los hiperparámetros, $[\theta_j|\mu,\tau,y]$ para $j = 1,\dots,J$, y la distribución a posteriori de los hiperparámetros, $[\mu,\tau|y]$.

Notemos que la distribución a priori de los hiperparámetros $[\mu,\tau]$ se puede expresar como $[\tau][\mu|\tau]$. Por tanto, podemos razonar como en el estudio de modelos multiparamétricos del Capítulo 2. En esta situación, fijaremos $[\mu|\tau] \propto 1$ y, al final, discutiremos qué opciones pueden tomarse para la distribución marginal a priori de τ. De momento, consideraremos $[\mu,\tau] \propto [\tau]$.

Determinemos las distribuciones a posteriori aplicando el procedimiento descrito en la Sección 3.1.

I) *Distribución conjunta a posteriori*. Utilizando la simplificación anterior de la verosimilitud, se tiene

$$[\theta,\mu,\tau|y] \propto [\mu,\tau][\theta|\mu,\tau][y|\theta] \propto [\mu,\tau]\prod_{j=1}^{J}\mathrm{N}(\theta_j|\mu,\tau^2)\prod_{j=1}^{J}\mathrm{N}(\overline{y}_{.j}|\theta_j,\sigma_j^2).$$

II) *Distribución condicional a posteriori de las medias dados los hiperparámetros* $[\theta_j|\mu,\tau,y]$. Al contar con J medias que, una vez condicionadas a μ y τ, son independientes y poseen distribución a priori normal, sabemos que, por (2.4),

$$\theta_j|\mu,\tau,y \sim \mathrm{N}(\hat{\theta}_j,V_j), \tag{3.2}$$

donde

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\overline{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}} \quad \text{y} \quad \frac{1}{V_j} = \frac{1}{\sigma_j^2}+\frac{1}{\tau^2}.$$

De nuevo, la media a posteriori, $\hat{\theta}_j$, se trata de una ponderación de la media a priori de la población y la media muestral del j-ésimo grupo con pesos dados por las precisiones de sendas distribuciones. La precisión a posteriori es la suma de las dos precisiones. Para acabar con este punto, obsérvese que $[\theta|\mu,\tau,y]$ es

$$\prod_{j=1}^{J}[\theta_j|\mu,\tau,y] = \prod_{j=1}^{J}\mathrm{N}(\theta_j|\hat{\theta}_j,V_j).$$

III) En cuanto a la caracterización de la *distribución a posteriori de los hiperparámetros* $[\mu,\tau|y]$, por tratarse de la verosimilitud normal, aplicamos el Teorema de Bayes, ya que la distribución $[y|\mu,\tau]$ se determina de forma sencilla —no es así en verosimilitudes cualesquiera. En efecto, considerando la distribución conjunta de θ y $\overline{y}_{.j}$ dados μ y τ e integrando respecto de θ se tiene que

$$[\overline{y}_{.j}|\mu,\tau] = \int_{\mathbb{R}^J}[\overline{y}_{.j},\theta|\mu,\tau]\,\mathrm{d}\theta = \int_{\mathbb{R}}[\overline{y}_{.j}|\theta_j][\theta_j|\mu,\tau]\,\mathrm{d}\theta_j, \quad j = 1,\dots,J. \tag{3.3}$$

Gracias a las distribuciones (3.2) y (3.1), el integrando $[\overline{y}_{.j}|\theta_j][\theta_j|\mu,\tau]$ resulta

$$(\sigma_j^2\tau^2)^{-1/2}\exp\left(-\frac{(\theta_j-\overline{y}_{.j})^2}{2\sigma_j^2}\right)\exp\left(-\frac{(\theta_j-\mu)^2}{2\tau^2}\right). \tag{3.4}$$

Denotando a los términos

$$\tilde{\mu}_j = \frac{\frac{\bar{y}_{.j}}{\sigma_j^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{y} \quad \frac{1}{\tilde{\sigma}_j^2} = \frac{1}{\sigma_j^2} + \frac{1}{\tau^2},$$

la expresión (3.4) se puede reescribir como

$$\begin{aligned}
&(\sigma_j^2\tau^2)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}_{.j}^2}{\sigma_j^2} + \frac{\mu^2}{\tau^2}\right)\right) \exp\left(-\frac{1}{2}\left(\frac{1}{\tilde{\sigma}_j^2}\theta_j^2 - 2\frac{\tilde{\mu}_j}{\tilde{\sigma}_j^2}\theta_j\right)\right) \\
&= (\sigma_j^2\tau^2)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}_{.j}^2}{\sigma_j^2} + \frac{\mu^2}{\tau^2}\right)\right) \exp\left(\frac{\tilde{\mu}_j^2}{2\tilde{\sigma}_j^2}\right) \exp\left(-\frac{(\theta_j - \tilde{\mu}_j)^2}{2\tilde{\sigma}_j^2}\right).
\end{aligned}$$

Y notando que el último factor es el *kernel* de una distribución normal $\mathrm{N}(\tilde{\mu}_j, \tilde{\sigma}_j^2)$, resulta que la distribución (3.3) se trata de

$$\begin{aligned}
[\bar{y}_{.j}|\mu,\tau] &\propto (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}_{.j}^2}{\sigma_j^2} + \frac{\mu^2}{\tau^2}\right)\right) \exp\left(\frac{\tilde{\mu}_j^2}{2\tilde{\sigma}_j^2}\right) \\
&= (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(\bar{y}_{.j} - \mu)^2}{2(\sigma_j^2 + \tau^2)}\right), \quad j = 1, \ldots, J.
\end{aligned}$$

Luego $\bar{y}_{.j}|\mu,\tau$ tiene distribución normal $\mathrm{N}(\mu, \sigma_j^2 + \tau^2)$ para todo j. Y, por tanto, se deduce que

$$[y|\mu,\tau] \propto \prod_{j=1}^{J} \mathrm{N}(\bar{y}_{.j}|\mu, \sigma_j^2 + \tau^2), \tag{3.5}$$

respecto de μ y τ. Así, por el Teorema de Bayes, la distribución a posteriori de los hiperparámetros resulta

$$[\mu,\tau|y] \propto [\mu,\tau][y|\mu,\tau] \propto [\mu,\tau] \prod_{j=1}^{J} \mathrm{N}(\bar{y}_{.j}|\mu, \sigma_j^2 + \tau^2). \tag{3.6}$$

3.2.2. Distribución condicional y marginal a posteriori de los hiperparámetros

Si bien el procedimiento descrito en la Sección 3.6 permite caracterizar la distribución conjunta a posteriori $[\theta,\mu,\tau|y]$; en la práctica, el objetivo es caracterizar las distribuciones a posteriori condicional $[\mu|\tau,y]$ y marginal $[\tau|y]$ porque ofrecen una factorización de la distribución conjunta a posteriori que puede usarse para la simulación.

a) *Distribución a posteriori condicionada* $[\mu|\tau,y]$. De nuevo, por el Teorema de Bayes y (3.5),

$$[\mu|\tau,y] \propto [\mu|\tau][y|\mu,\tau] \propto \prod_{j=1}^{J} \mathrm{N}(\mu|\bar{y}_{.j}, \sigma_j^2 + \tau^2),$$

que será una distribución normal de media $\hat{\mu}$ y varianza V_μ con

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \quad \text{y} \quad \frac{1}{V_\mu} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}. \tag{3.7}$$

b) *Distribución a posteriori marginal* $[\tau|y]$. Esta distribución, dada su distribución a priori, puede obtenerse por medio de la expresión de la densidad condicionada y las distribuciones ya calculadas. Teniendo en mente que $[\mu,\tau] \propto [\tau]$, se tiene

$$[\tau|y] = \frac{[\mu,\tau|y]}{[\mu|\tau,y]} \propto \frac{[\tau]\prod_{j=1}^{J} \mathrm{N}(\bar{y}_{.j}|\hat{\mu}, \sigma_j^2 + \tau^2)}{\mathrm{N}(\mu|\hat{\mu}, V_\mu)}.$$

A continuación, nótese que el lado izquierdo de la igualdad no depende de μ. Esto nos proporciona una vía para simplificar su expresión al poder fijar $\mu = \hat{\mu}$, de manera que

$$[\tau|y] \propto \frac{[\tau]\prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j}|\hat{\mu}, \sigma_j^2+\tau^2)}{\mathrm{N}(\hat{\mu}|\hat{\mu}, V_\mu)} \propto [\tau] V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2+\tau^2)^{-1/2} \exp\left(-\frac{(\overline{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right). \tag{3.8}$$

Por último, puesto que, en todos estos desarrollos, hemos indicado la distribución a priori de τ, pero sin escoger ninguna en particular, propondremos algunas elecciones.

3.2.3. Distribución a priori de los hiperparámetros

Dado que en la práctica es frecuente trabajar con distribuciones a priori no informativas para las varianzas, las densidades que presentaremos a continuación son opciones razonables. Notemos que, al ser $\tau > 0$, estudiaremos la integrabilidad de la densidad $[\tau|y]$ sobre $(0, +\infty)$.

a) *Distribución a priori* $[\tau] \propto 1/\tau$. Tras las demostraciones ya vistas, una primera apuesta razonable es $[\tau] \propto 1/\tau$. Sin embargo, notemos que, al sustituir en (3.8), este factor hace aparecer un problema en el comportamiento en $\tau = 0$. En efecto, es claro que

$$V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2+\tau^2)^{-1/2} \exp\left(-\frac{(\overline{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right)$$

tiende a una constante positiva si $\tau \to 0^+$. Por este motivo, gracias al criterio de comparación por paso al límite, si estudiamos la integrabilidad de la densidad en $(0,1)$, esta tiene el mismo carácter que $1/\tau$, que es divergente. Por tanto, esta densidad a priori no es válida en esta situación.
Por contra, sí que lo será la distribución no informativa por antonomasia: la distribución uniforme.

b) *Distribución a priori uniforme* $[\tau] \propto 1$. Comprobemos que la distribución a posteriori de τ es propia. Por (3.8),

$$\begin{aligned}
[\tau|y] &\propto V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2+\tau^2)^{-1/2} \exp\left(-\frac{(\overline{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right) \\
&= \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\right)^{-1/2} \prod_{j=1}^{J} (\sigma_j^2+\tau^2)^{-1/2} \exp\left(-\frac{(\overline{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right) \\
&= \left(\sum_{j=1}^{J} \prod_{\substack{i=1 \\ i\neq j}}^{J} (\sigma_i^2+\tau^2)\right)^{-1/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{J} \frac{(\overline{y}_{.j}-\hat{\mu})^2}{\sigma_j^2+\tau^2}\right) \\
&\leq \frac{1}{J^{1/2}\tau^{J-1}} \exp\left(-\frac{1}{2}\sum_{j=1}^{J} \frac{(\overline{y}_{.j}-\hat{\mu})^2}{\sigma_j^2+\tau^2}\right).
\end{aligned} \tag{3.9}$$

Gracias a que la densidad es continua sobre cualquier intervalo acotado en $[0, +\infty)$, bastará estudiar su comportamiento cuando $\tau \to +\infty$. Con tal fin, consideramos la cota (3.9). Recuperando la expresión de (3.7), es fácil ver que

$$\lim_{\tau \to +\infty} \hat{\mu} = \frac{1}{J}\sum_{j=1}^{J} \overline{y}_{.j} = \overline{\overline{y}}_{.},$$

que es una cantidad finita, por lo que el límite de la exponencial resulta

$$\lim_{\tau \to +\infty} \exp\left(-\frac{1}{2}\sum_{j=1}^{J} \frac{(\overline{y}_{.j}-\hat{\mu})^2}{\sigma_j^2+\tau^2}\right) = 1,$$

puesto que cada sumando $(\overline{y}_{.j}-\hat{\mu})^2/(\sigma_j^2+\tau^2)$ tiende a 0. En consecuencia, se tiene que (3.9) es equivalente a $(J^{1/2}\tau^{J-1})^{-1}$ cuando τ tiende a $+\infty$, y, así, puede garantizarse que la distribución a priori uniforme de τ produce una distribución a posteriori propia si el número de grupos es mayor o igual que 3.

Capítulo 4

Modelos de regresión lineal bayesianos

4.1. Modelos regresión

Los modelos de regresión son una de las técnicas estadísticas más utilizadas con aplicaciones en múltiples campos. En estos modelos, se plantea que una variable y, denominada *respuesta*, es una función lineal de una o varias variables explicativas o predictoras x_i. En general, supondremos k el número de variables regresoras, de este modo, el modelo de regresión se expresa como

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

donde $\beta = (\beta_0, \ldots, \beta_k)$ son los *coeficientes de regresión* y ε es un término de error aleatorio. El objetivo de un modelo de regresión habitual consiste en tratar de obtener una estimación de los coeficientes para obtener un buen ajuste de la variable y. Estas estimaciones pueden hacerse a partir de una muestra de observaciones $(y_1, \ldots, y_n)$ y una matriz del diseño $n \times (k+1)$, $\mathbf{X}$.

Nótese que las hipótesis que se tomen sobre la distribución de y_i determinan la complejidad del modelo. En el caso de considerar el modelo de regresión *clásico*, se supone que $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ para todo i, por lo que, en consecuencia:

- Las variables y_i son *homocedásticas*, esto es, $\mathrm{Var}(y_i|\beta, \sigma^2) = \sigma^2$ para todo $i = 1, \ldots, n$.
- Las variables y_i tienen distribución normal dados β, σ^2 y $\mathbf{X}$.

Además, las variables y_i deben ser independientes entre sí dados β, σ^2 y $\mathbf{X}$; y se añade una condición de no colinealidad sobre la matriz $\mathbf{X}$, lo que significa que sus columnas deben ser linealmente independientes, luego $\mathrm{ran}(\mathbf{X}) = k+1$.

4.2. Análisis bayesiano del modelo de regresión clásico

Los parámetros que intervienen en el modelo anterior son $\theta = (\beta_0, \ldots, \beta_k, \sigma^2)$ y nuestro objetivo será determinar las distribuciones a posteriori condicional $[\beta|\sigma^2, \mathbf{X}, y]$ y marginal $[\sigma^2|\mathbf{X}, y]$ puesto que ofrecen una factorización de la distribución conjunta a posteriori. Tales distribuciones las denotaremos como $[\beta|\sigma^2, y]$ y $[\sigma^2|y]$ respectivamente, eliminando la indicación explícita de $\mathbf{X}$ ya que siempre lo supondremos conocido. Para caracterizar estas distribuciones es necesario fijar la distribución a priori de θ. Si el número de parámetros a estimar es mayor que el número de datos será preciso utilizar una distribución a priori informativa. En otro caso, es razonable utilizar una distribución a priori no informativa, siendo una elección habitual

$$[\beta, \sigma^2] \propto (\sigma^2)^{-1}. \tag{4.1}$$

Las distribuciones que se presentan a continuación se desarrollan a partir de esta distribución a priori. En el caso de modelos de regresión, la verosimilitud corresponde a la distribución del vector y. En el modelo clásico, la verosimilitud es

$$[y|\beta, \sigma^2] = \mathrm{N}(y|\mathbf{X}\beta, \sigma^2 \mathbf{I}_n) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} (y - \mathbf{X}\beta)^t (y - \mathbf{X}\beta) \right), \tag{4.2}$$

con $\mathbf{I}_n$ la matriz indentidad de dimensión n. Y la distribución conjunta a posteriori es

$$[\beta, \sigma^2|y] \propto (\sigma^2)^{-(n/2+1)} \exp\left(-\frac{1}{2\sigma^2}(y-\mathbf{X}\beta)^t(y-\mathbf{X}\beta)\right). \tag{4.3}$$

Procedemos con la caracterización de las distribuciones a posteriori condicional $[\beta|\sigma^2, y]$ y marginal $[\sigma^2|y]$.

- *Distribución condicional a posteriori* $[\beta|\sigma^2, y]$. Por (4.1) y (4.2),

$$\begin{aligned}[\beta|\sigma^2, y] &\propto \exp\left(-\frac{1}{2\sigma^2}(y-\mathbf{X}\beta)^t(y-\mathbf{X}\beta)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\beta^t\mathbf{X}^t\mathbf{X}\beta - 2\beta^t\mathbf{X}^t y)\right) \\ &\propto \exp\left(-\frac{1}{2}(\beta^t(\sigma^2)^{-1}\mathbf{X}^t\mathbf{X}\beta - 2\beta^t(\sigma^2)^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y)\right).\end{aligned}$$

  Por ser $\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ una matriz simétrica y definida positiva puesto que $\text{ran}(\mathbf{X}) = k+1$, se tiene que

$$[\beta|\sigma^2, y] = \text{N}(\beta|\hat{\beta}, V_\beta), \tag{4.4}$$

  con

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y \quad \text{y} \quad V_\beta = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}.$$

- *Distribución marginal a posteriori* $[\sigma^2|y]$. En este caso, por la definición de distribución condicional $[\beta|\sigma^2, y]$,

$$[\sigma^2|y] = \frac{[\beta, \sigma^2|y]}{[\beta|\sigma^2, y]}.$$

  Y dado que la expresión anterior no depende de β, basta con sustituir $\beta = \hat{\beta}$ en (4.3) y (4.4). Así,

$$\begin{aligned}[\sigma^2|y] &\propto \frac{(\sigma^2)^{-(n/2+1)}}{(\sigma^2)^{-(k+1)/2}} \exp\left(-\frac{n-k-1}{2\sigma^2(n-k-1)}(y-\mathbf{X}\hat{\beta})^t(y-\mathbf{X}\hat{\beta})\right) \\ &\propto \text{Inv-}\chi^2(\sigma^2|n-k-1, s^2),\end{aligned}$$

  con

$$s^2 = \frac{1}{n-k-1}(y-\mathbf{X}\hat{\beta})^t(y-\mathbf{X}\hat{\beta}).$$

De estas distribuciones, se deduce que la distribución conjunta (4.3) es propia ya que $n > k+1$ y el rango de $\mathbf{X}$ es $k+1$. Así, se comprueba que el número de datos ha de ser siempre mayor que el número de parámetros, $k+1$, en el caso de ausencia de información previa y, por otro lado, que las columnas de $\mathbf{X}$ han de ser linealmente independientes. Además, se observa otra vez la relación entre el paradigma bayesiano y frecuentista comentado en la Observación 2.5 en las distribuciones $[\beta|\sigma^2, y]$ y $[\sigma^2|y]$ —compárense con las distribuciones de los estimadores por mínimos cuadrados frecuentistas $\hat{\beta}$ y $\widehat{\sigma}^2$.

4.3. Modelos con varianzas desiguales y correlaciones

Una generalización básica del modelo de regresión básico es permitir que las varianzas de la variable respuesta sean diferentes y/o que las variables no sean independientes entre sí. Esto supone considerar un modelo en el que la matriz de varianzas-covarianzas de y sea una matriz Σ_y $n \times n$, simétrica y definida positiva no restringida a la forma $\sigma^2\mathbf{I}_n$. En consecuencia, la distribución de la respuesta es

$$[y|\beta,\Sigma_y] = \mathrm{N}(y|\mathbf{X}\beta,\Sigma_y) \propto \det(\Sigma_y)^{-1/2} \exp\left(-\frac{1}{2}(y-\mathbf{X}\beta)^t\Sigma_y^{-1}(y-\mathbf{X}\beta)\right). \tag{4.5}$$

En primer lugar, determinaremos $[\beta|\Sigma_y,y]$ y, después, la distribución marginal $\Sigma_y|y$, que dependerá la elección de Σ_y.

- *Distribución condicional* $[\beta|\Sigma_y,y]$. Análogamente a (4.4), puede verse que

$$[\beta|\Sigma_y,y] = \mathrm{N}(\beta|\hat{\beta},V_\beta),$$

donde

$$\hat{\beta} = (\mathbf{X}^t\Sigma_y^{-1}\mathbf{X})^{-1}\mathbf{X}^t\Sigma_y^{-1}y \quad \text{y} \quad V_\beta = (\mathbf{X}^t\Sigma_y^{-1}\mathbf{X})^{-1}.$$

- *Distribución marginal* $[\Sigma_y|y]$.

$$\begin{aligned}[\Sigma_y|y] &= \frac{[\beta,\Sigma_y|y]}{[\beta|\Sigma_y,y]} = \frac{[\Sigma_y]\mathrm{N}(y|\beta,\Sigma_y)}{\mathrm{N}(\beta|\hat{\beta},V_\beta)} \\ &= [\Sigma_y]\det(\Sigma_y)^{-1/2}\det(V_\beta)^{1/2}\exp\left(-\frac{1}{2}(y-\mathbf{X}\hat{\beta})^t\Sigma_y^{-1}(y-\mathbf{X}\hat{\beta})\right),\end{aligned} \tag{4.6}$$

donde la ultima igualdad se deduce al tomar $\beta = \hat{\beta}$.

A continuación, presentamos la distribución (4.6) para algunos ejemplos particulares de Σ_y.

Ejemplo 4.1. a) *Matriz de varianzas-covarianzas conocida salvo por un factor escalar.* En este caso suponemos que la matriz de varianzas-covarianzas es $\Sigma_y = \sigma^2 Q_y$, con Q_y conocida y σ^2 desconocido. Notemos que se trata de una generalización del modelo clásico, que considera $Q_y = \mathbf{I}_n$. La distribución a posteriori, dada la distribución a priori $[\beta,\sigma^2] \propto (\sigma^2)^{-1}$, se calcula de forma análoga. Así, es inmediato comprobar que

$$[\sigma^2|y] = \text{Inv-}\chi^2(\sigma^2|n-k-1,s^2),$$

y

$$\hat{\beta} = (\mathbf{X}^tQ_y^{-1}\mathbf{X})^{-1}\mathbf{X}^tQ_y^{-1}y, \quad V_\beta = \sigma^2(\mathbf{X}^tQ_y^{-1}\mathbf{X})^{-1}, \quad s^2 = \frac{1}{n-k-1}(y-\mathbf{X}\hat{\beta})^tQ_y^{-1}(y-\mathbf{X}\hat{\beta}).$$

Nótese que este ejemplo incluye como caso particular el modelo *regresión con pesos*, pues este se corresponde con el caso $Q_y = \mathrm{diag}(1/w_1,\ldots,1/w_n)$, donde $w_i \geq 0$ para todo i y $\sum_{i=1}^n w_i = 1$.

b) *Grupos con misma varianza*. En este ejemplo, tratamos un modelo de regresión en el que las n observaciones pueden dividirse en I grupos con misma varianza en cada uno. Supongamos que tenemos n_i datos del grupo i para $i = 1,\ldots,I$ —de manera que $n_1+\cdots+n_I = n$— y que, para tales grupos, los datos tienen distribución normal con varianza σ_i^2. En tal caso,

$$\Sigma_y = \begin{pmatrix} \sigma_1^2\mathbf{I}_{n_1\times n_1} & O & \cdots & O \\ O & \sigma_2^2\mathbf{I}_{n_2\times n_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \cdots & O & \sigma_I^2\mathbf{I}_{n_I\times n_I} \end{pmatrix}.$$

Una distribución a priori no informativa de los parámetros $(\beta,\sigma_1^2,\ldots,\sigma_I^2)$ es

$$[\beta,\sigma_1^2,\ldots,\sigma_I^2] \propto \prod_{i=1}^I(\sigma_i^2)^{-1},$$

que proporciona distribuciones a posteriori marginales para σ_i^2 si $n_i \geq 2$ —Gelman, A. et al (2021) [5]. Sustityendo en (4.6), se tiene

$$[\sigma_1^2,\ldots,\sigma_I^2|y] \propto \prod_{i=1}^I(\sigma_i^2)^{-1}\det(V_\beta)^{1/2}\det(\Sigma_y)^{-1/2}\exp\left(-\frac{1}{2}(y-\mathbf{X}\hat{\beta})^t\Sigma_y^{-1}(y-\mathbf{X}\hat{\beta})\right).$$

4.4. Modelos de regresión jerárquicos

Los modelos de regresión jerárquicos permiten representar situaciones donde los predictores tienen distintos *niveles de variación*. Para fijar ideas se expone brevemente el siguiente ejemplo —Gelman, A. y Hill, J. (2007) [6]. Supongamos que se desea analizar los efectos de los programas de preparación de colegios diferentes para un test de aptitud. En este ejemplo, se puede disponer de información a distintos niveles: a nivel individual de cada alumno —como su entorno familiar—, información relativa a la clase —características del profesor— o a nivel de toda la escuela. También, por otro lado, los modelos jerárquicos sirven para modelizar datos que provienen de un muestreo por grupos.

4.4.1. Modelo de regresión con efectos aleatorios

El modelo de regresión jerárquico más sencillo es un modelo de efectos aleatorios simple. En esta situación, se consideran grupos de coeficientes β que son intercambiables y normalmente distribuidos. Más concretamente, el modelo queda especificado por la verosimilitud (4.5) y la distribución a priori

$$\beta|b,\sigma_\beta^2 \sim \mathrm{N}(b\mathbf{1},\sigma_\beta^2\mathbf{I}_{k+1}), \tag{4.7}$$

donde $\mathbf{1}=(1,\dots,1)^t \in \mathbb{R}^{k+1}$ y los hiperpámetros son desconocidos. Notemos que esta distribución incluye como casos particulares tanto el hecho de que los parámetros β_j no estén relacionados, lo que se expresa escogiendo $\sigma_\beta^2=\infty$, o que sean iguales, esto es, $\sigma_\beta^2=0$; por lo que, en efecto, permite expresar de forma sencilla una estructura de dependencia jerárquica como la ya vista en el Capítulo 3.

Ejemplifiquemos esto último retomando el ejemplo de los programas de estudio en los colegios. Sea J el número de colegios. Si denotamos por $y_{i,j}$ al resultado de un alumno i en el colegio j, es razonable suponer que

$$y_{i,j}=\mu+\beta_j+\varepsilon_{i,j}, \quad \varepsilon_{i,j}\sim \mathrm{N}(0,\sigma^2),$$

donde $\beta=(\beta_1,\dots,\beta_J)$ tiene densidad conjunta (4.7). De este modo, los resultados del test de alumnos de un mismo colegio se ven relacionados al considerarse observaciones de una misma distribución $\mathrm{N}(\mu+\beta_j,\sigma^2)$, y, además, se establece una estructura jerárquica que relaciona a todos los colegios. Tenemos entonces un efecto aleatorio que actúa en el nivel de los colegios.

Para completar la especificación de un modelo jerárquico como el del comienzo, es preciso escoger una distribución a priori de los hiperparámetros b y σ_β^2. En esta situación, pueden suponerse ambos hiperparámetros independientes. Para b, escogemos $[b]\propto 1$. En cuanto a $[\sigma_\beta^2]$, consideraremos la distribución inversa-χ^2. Más concretamente,

$$[b]\propto 1, \quad [\sigma_\beta^2]=\text{Inv-}\chi^2(\sigma_\beta^2|\nu,s^2), \tag{4.8}$$

donde $\mu\in\mathbb{R}$, $\sigma_\beta^2,s^2>0$ y $\nu\in\mathbb{N}$.

Observación 4.2. Es importante destacar que el modelo de efectos aleatorios planteado, con independencia de las elecciones de las distribuciones a priori de los hiperparámetros, permite representar situaciones en las que existe una correlación dentro de las observaciones de un mismo grupo. En efecto, puede probarse —veáse Gelman, A. et al (2021) [5, p. 382]— que los modelos

I) $y=(y_1,\dots,y_n)$ distribuidos en J grupos diferentes y con distribución $\mathrm{N}(b\mathbf{1},\Sigma_y)$, donde $\mathrm{Var}(y_i)=\eta^2$ para todo i y $\mathrm{Cov}(y_{i_1},y_{i_2})=\rho\eta^2\geq 0$ si los datos pertenecen al mismo grupo y 0 en otro caso;

II) $y=(y_1,\dots,y_n)$ con distribución $\mathrm{N}(\mathbf{X}\beta,\sigma^2 I_n)$ y β como (4.7), donde $\mathbf{X}$ es una matriz indicadora con $\mathbf{X}_{i,j}=1$ si i está en el grupo j y 0 en caso contrario

son equivalentes siempre que $\eta^2=\sigma^2+\sigma_\beta^2$ y $\rho=\sigma_\beta^2/(\sigma^2+\sigma_\beta^2)$.

En resumen, la conclusión que se extrae es que se puede construir un modelo que capture la correlación existente entre observaciones de un mismo grupo incluyendo los efectos aleatorios apropiados. Esta observación es especialmente interesante puesto que abre la posibilidad de modelizar dependencias dentro de un grupo.

4.5. Aplicación al análisis espacio-temporal de temperaturas medias en verano en una región

En esta sección, se desarrolla un modelo bayesiano de regresión para representar las temperaturas medias durante el verano en un área alrededor de la Comunidad Autónoma de Aragón. Para ello, contamos con un conjunto de datos que contiene observaciones de temperaturas máximas *diarias* medidas en grados centígrados en 18 localidades proporcionadas por la Agencia Estatal de Meteorología (AEMET) —véase la Figura A.1. Estos datos recogen temperaturas desde 1956 hasta 2015 durante el período estival, desde junio hasta agosto inclusive. Computando las medias de tales datos, se han construido las temperaturas medias en verano recogidas en la matriz de datos y.

Para el estudio de las temperaturas medias cabe tener en cuenta las siguientes características. En primer lugar, es esperable una evolución creciente a lo largo del tiempo como consecuencia del calentamiento global y que suele modelizarse como una tendencia lineal en el tiempo o mediante efectos aleatorios—vid. Castillo-Mateo, J. (2022) et al [3]. Además, es claro que las temperaturas medias presentarán una dependencia de las diferentes características geográficas de las localidades puesto que la zona de estudio está conformada por regiones con climas diferentes como son el Valle del Ebro, los Pirineos y el Sistema Ibérico. Estas diferencias climáticas se ven influenciadas, en gran medida, por la diferente altitud respecto del nivel del mar, motivo por el cual es sensato considerar la altitud —medida en metros— como variable explicativa. Junto con todo lo anterior, es también razonable suponer que existe dependencia de la temperatura de un año con la del año anterior en una misma localidad —esto es, que existe una correlación serial. Esta correlación suele representarse mediante una *estructura autorregresiva*, es decir, establecer que dicha dependencia sea lineal —vid. Brockwell, P. J. y Davis, R. A. (2006) [2]— y que se satisfaga una relación markoviana de dependencia de las temperaturas respecto de t, i.e., $[y_{t,s}|y_{t-1,s},\dots,y_{1,s}] = [y_{t,s}|y_{t-1,s}]$. Como resultado de imponer esta estructura, el modelo representará la distribución condicionada de la temperatura de un año por la del año anterior. Como última observación, notemos que, en un modelo para temperaturas en distintas localidades espaciales, cabe esperar que las observaciones de todos los observatorios correspondientes a un mismo año no sean independientes, sino que exista relación entre ellas. Esta dependencia se puede representar en un modelo jerárquico con un efecto aleatorio asociado a cada año, como los descritos anteriormente.

Teniendo en cuenta todo lo anterior, se plantean dos modelos, uno no jerárquico —por lo que no incluirá el efecto aleatorio en la tendencia— y otro jerárquico. La diferencia entre ambos estará, así, en que el modelo no jerárquico modeliza la tendencia con una covariable asociada a una tendencia lineal, mientras que el jerárquico permite mayor flexibilidad y lo hace mediante efectos aleatorios para cada año. En ambos casos, los ejemplos propuestos reflejarán cómo las distribuciones a posteriori conjugadas junto con técnicas MCMC permiten ajustar un modelo bayesiano de manera eficiente. Es importante destacar que estos modelos sobre las temperaturas medias del verano son meramente ilustrativo, ya que un ajuste óptimo requeriría un modelo más complicado con más términos para representar la variabilidad espacial. Entre ellos, por ejemplo, tenemos términos independientes, tendencias específicas de cada localidad espacial s y la parametrización en términos de anomalías —véase Castillo-Mateo, J. et al (2022) [3]. Todos estos elementos quedan fuera de los objetivos de este trabajo.

4.5.1. Modelo de regresión no jerárquico

Se considera el modelo

$$y_{t,s} = \beta_0 + \alpha t + \rho y_{t-1,s} + \gamma\,\mathrm{alt}(s) + \varepsilon_{t,s}, \quad \varepsilon_{t,s} \sim \mathrm{N}(0,\sigma^2), \quad t = 1,\dots,60, \quad s = 1,\dots,18,$$

donde $y_{t,s}$ representa la temperatura en el año t y localidad s; α expresa la tendencia lineal respecto del tiempo; $\mathrm{alt}(s)$, la altitud de la localidad s y ρ es el factor de correlación de la serie temporal, que puede considerarse en el intervalo $(-1,1)$ para que la serie sea estacionaria —Brockwell, P. J. y Davis, R. A. (2006) [2]. De este modo, los parámetros de este modelo son $\theta = (\beta_0,\alpha,\gamma,\rho,\sigma^2)$.

Distribución conjunta a posteriori

Dada la estructura autorregresiva del modelo, la verosimilitud, $[y|\theta]$, se expresa como

$$\prod_{s=1}^{18}\prod_{t=1}^{60}[y_{t,s}|y_{t-1,s},\theta] = \prod_{s=1}^{18}\prod_{t=1}^{60}\mathrm{N}(y_{t,s}|\beta_0+\alpha t+\rho y_{t-1,s}+\gamma\,\mathrm{alt}(s),\sigma^2)$$
$$\propto (\sigma^2)^{-540}\exp\left(-\frac{1}{2\sigma^2}\sum_{s=1}^{18}\sum_{t=1}^{60}(y_{t,s}-\beta_0-\alpha t-\rho y_{t-1,s}-\gamma\,\mathrm{alt}(s))^2\right). \tag{4.9}$$

Puesto que, como se indicaba, condicionamos a $y_{t-1,s}$, la verosimilitud es análoga a (4.2).

Suponiendo que las distribuciones a priori siguientes son independientes,

$$\begin{array}{lll}\beta_0\sim\mathrm{N}(\mu_{\beta_0},\tau^2_{\beta_0}), & \alpha\sim\mathrm{N}(\mu_\alpha,\tau^2_\alpha), & \gamma\sim\mathrm{N}(\mu_\gamma,\tau^2_\gamma),\\ \rho\sim\mathrm{U}(-1,1), & \sigma^2\sim\text{Inv-}\chi^2(n_0,s_0^2), & \end{array}$$

se tiene que la densidad conjunta a posteriori es proporcional a

$$[\beta_0][\alpha][\gamma][\rho][\sigma^2](\sigma^2)^{-540}\exp\left(-\frac{1}{2\sigma^2}\sum_{s=1}^{18}\sum_{t=1}^{60}(y_{t,s}-\beta_0-\alpha t-\rho y_{t-1,s}-\gamma\,\mathrm{alt}(s))^2\right). \tag{4.10}$$

Dado que la distribución a posteriori obtenida es multivariante y no corresponde a una distribución conocida, es necesario, a la hora de hacer inferencia, recurrir a métodos MCMC para obtener simulaciones de los valores de los parámetros y, a partir de ello, simulaciones de valores de la respuesta. En este caso, utilizaremos un método de *Gibbs sampling*, cuyo funcionamiento ya fue expuesto brevemente en el Capítulo 1. Para implementar el algoritmo, se han de determinar las *distribuciones completamente condicionales a posteriori*, esto es, las distribuciones condicionales a los datos y a los demás parámetros.

Distribuciones completamente condicionales

A estas distribuciones completamente condicionales las denotaremos como $[\beta_0|y,\ldots]$, $[\alpha|y,\ldots]$, etc. para simplificar la notación.

Observemos que el cálculo de las distribuciones a posteriori condicionadas a todos los demás parámetros es sencillo por la observación siguiente. Centrándonos en β_0 —es análogo para las demás—, nótese que, por definición de densidad condicional,

$$[\beta_0|\alpha,\gamma,\rho,\sigma^2,y] = \frac{[\beta_0,\alpha,\gamma,\rho,\sigma^2|y]}{[\alpha,\gamma,\rho,\sigma^2|y]} \propto [\beta_0,\alpha,\gamma,\rho,\sigma^2|y].$$

Esto significa que las densidades completamente condicionales a posteriori son proporcionales a la densidad conjunta a posteriori. Calcularemos de forma detallada la distribución completamente condicionada de β_0 y las demás se expondrán brevemente.

- *Distribución* $[\beta_0|y,\ldots]$. Por la observación anterior,

$$[\beta_0|y,\ldots]\propto[\beta_0]\exp\left(-\frac{1}{2\sigma^2}\sum_{s=1}^{18}\sum_{t=1}^{60}(y_{t,s}-\beta_0-\alpha t-\rho y_{t-1,s}-\gamma\,\mathrm{alt}(s))^2\right).$$

  Y, dado que $\beta_0\sim\mathrm{N}(\mu_{\beta_0},\tau^2_{\beta_0})$,

$$[\beta_0|y,\ldots]\propto\exp\left(-\frac{(\beta_0-\mu_{\beta_0})^2}{2\tau^2_{\beta_0}}\right)\exp\left(-\frac{1}{2\sigma^2}\left(1080\beta_0^2-2\beta_0\sum_{s=1}^{18}\sum_{t=1}^{60}(y_{t,s}-\alpha t-\rho y_{t-1,s}-\gamma\,\mathrm{alt}(s))\right)\right).$$

  Es claro, así, que la distribución será normal al ser el exponente una función cuadrática de β_0. De este modo, es suficiente con calcular los coeficientes de β_0^2 y β_0, de manera que resulta

$$[\beta_0|y,\ldots]=\mathrm{N}\left(\beta_0\left|\frac{\mu_{\beta_0}/\tau^2_{\beta_0}+\sum_{s=1}^{18}\sum_{t=1}^{60}(y_{t,s}-\alpha t-\rho y_{t-1,s}-\gamma\,\mathrm{alt}(s))/\sigma^2}{1/\tau^2_{\beta_0}+1080/\sigma^2},\frac{1}{1/\tau^2_{\beta_0}+1080/\sigma^2}\right.\right).$$

  De forma completamente análoga, se deducen las distribuciones de α y γ.

- *Distribución* $[\alpha|y,\ldots]$.

$$[\alpha|y,\ldots] = \mathrm{N}\left(\alpha \,\middle|\, \frac{\mu_\alpha/\tau_\alpha^2 + \sum_{s=1}^{18}\sum_{t=1}^{60} t(y_{t,s} - \beta_0 - \rho y_{t-1,s} - \gamma\,\mathrm{alt}(s))/\sigma^2}{1/\tau_\alpha^2 + 1328580/\sigma^2}, \frac{1}{1/\tau_\alpha^2 + 1328580/\sigma^2}\right).$$

- *Distribución* $[\gamma|y,\ldots]$.

$$[\gamma|y,\ldots] = \mathrm{N}\left(\gamma \,\middle|\, \frac{\mu_\gamma/\tau_\gamma^2 + \sum_{s=1}^{18}\mathrm{alt}(s)\sum_{t=1}^{60}(y_{t,s} - \beta_0 - \alpha t - \rho y_{t-1,s})/\sigma^2}{1/\tau_\gamma^2 + 60\sum_{s=1}^{18}\mathrm{alt}(s)^2/\sigma^2}, \frac{1}{1/\tau_\gamma^2 + 60\sum_{s=1}^{18}\mathrm{alt}(s)^2/\sigma^2}\right).$$

- *Distribución* $[\rho|y,\ldots]$. Al ser la distribución a priori de ρ uniforme en $(-1,1)$, su distribución a posteriori estará concentrada $(-1,1)$. De hecho, para $\rho \in (-1,1)$, es claro que $[\rho|y,\ldots]$ será propocional a una densidad normal, por lo que se trata de una densidad normal truncada. Más concretamente,

$$\begin{aligned}[\rho|y,\ldots] &\propto \exp\left(-\frac{1}{2\sigma^2}\left(\rho^2\sum_{s=1}^{18}\sum_{t=1}^{60} y_{t-1,s}^2 - 2\rho\sum_{s=1}^{18}\sum_{t=1}^{60} y_{t-1,s}(y_{t,s} - \beta_0 - \alpha t - \gamma\,\mathrm{alt}(s))\right)\right)\mathbf{1}_{(-1,1)}(\rho)\\ &\propto \mathrm{N}\left(\rho \,\middle|\, \frac{\sum_{s=1}^{18}\sum_{t=1}^{60} y_{t-1,s}(y_{t,s} - \beta_0 - \alpha t - \gamma\,\mathrm{alt}(s))}{\sum_{s=1}^{18}\sum_{t=1}^{60} y_{t-1,s}^2}, \frac{\sigma^2}{\sum_{s=1}^{18}\sum_{t=1}^{60} y_{t-1,s}^2}\right)\mathbf{1}_{(-1,1)}(\rho).\end{aligned}$$

- *Distribución* $[\sigma^2|y,\ldots]$. Notemos que de (4.9) y la elección de $[\sigma^2]$, es inmediato que

$$[\sigma^2|y,\ldots] = \text{Inv-}\chi^2\left(\sigma^2 \,\middle|\, 1080 + n_0, \frac{1}{1080+n_0}\left(n_0 s_0^2 + \sum_{s=1}^{18}\sum_{t=1}^{60}(y_{t,s} - \beta_0 - \alpha t - \rho y_{t-1,s} - \gamma\,\mathrm{alt}(s))^2\right)\right).$$

Los resultados de la simulación basada en *Gibbs sampling*, así como los códigos utilizados, pueden consultarse en el Apéndice A en la Sección A.2. También se incluyen criterios para comprobar la convergencia del método, las densidades marginales a posteriori de los parámetros del modelo lineal y sus esperanzas a posteriori junto con intervalos de credibilidad al 95 %. En el Apéndice B Sección B.1 puede consultarse la implementación de código de `R`.

4.5.2. Modelo jerárquico

Según lo expuesto en la Sección 4.4.1, el sentido de un efecto aleatorio es incorporar, a un conjunto de datos en un mismo grupo, un parámetro común a todos ellos para establecer una relación entre sus valores en la variable explicada.

Por la estructura de nuestros datos $y_{t,s}$, en los que aparece el año y la localidad, para crear una relación de tipo espacial, hemos de añadir un parámetro δ_t para cada año $t = 1,\ldots,60$. De este modo, en cada año t, incluimos ese término aleatorio mencionado que afectará a todas las localidades por igual. Como ya indicábamos en la sección anterior, estos efectos aleatorios sustituirán a la tendencia lineal α del modelo no jerárquico. Además, el parámetro β_0 desaparece de la verosimilitud y se incluye como hiperparámetro de los δ_t para todo t. Así, el efecto aleatorio δ_t representará la aleatorización de esa temperatura de base β_0 en el año t. Consideramos, así,

$$\delta_t|\beta_0,\tau^2 \sim \mathrm{N}(\beta_0,\tau^2), \quad t = 1,\ldots,60,$$

junto con las distribuciones independientes a priori de los hiperámetros (4.8):

$$[\beta_0] \propto 1, \quad [\tau^2] = \text{Inv-}\chi^2(\tau^2|n_1,s_1^2).$$

El modelo jerárquico a estudiar ahora es

$$y_{t,s} = \delta_t + \rho y_{t-1,s} + \gamma\,\mathrm{alt}(s) + \varepsilon_{t,s}, \quad \varepsilon_{t,s} \sim \mathrm{N}(0,\sigma^2), \quad t = 1,\ldots,60, \quad s = 1,\ldots,18,$$

con vector de parámetros $\theta = (\delta_1, \ldots, \delta_{60}, \rho, \gamma, \sigma^2)$ e hiperparámetros $\phi = (\beta_0, \tau^2)$. La distribución conjunta a posteriori será a análoga a (4.10) pero añadiendo los hiperparámetros.

$$\begin{aligned}[\theta, \phi | y] &= [\delta_1, \ldots, \delta_{60}, \beta_0, \tau^2, \alpha, \rho, \gamma, \sigma^2 | y] \\ &\propto [\delta_1 | \beta_0, \tau^2] \cdots [\delta_{60} | \beta_0, \tau^2][\beta_0][\tau^2][\alpha][\rho][\gamma][\sigma^2] \\ &\times (\sigma^2)^{-540} \exp\left(-\frac{1}{2\sigma^2} \sum_{s=1}^{18} \sum_{t=1}^{60} (y_{t,s} - \delta_t - \alpha t - \rho y_{t-1,s} - \gamma \operatorname{alt}(s))^2 \right)\end{aligned}$$

Distribuciones completamente condicionales

Para el cálculo de las densidades completamente condicionales, observemos que, para los parámetros ρ, γ y σ^2, el resultado debe ser análogo, ya que son independientes de los hiperparámetros y, como sabemos, estos no aparecen en la verosimilitud. De hecho, es suficiente con sustituir $\beta_0 + \alpha t$ por los nuevos parámetros δ_t. En efecto, veámoslo para γ.

- *Distribución* $[\gamma | y, \ldots]$. Como antes, la distribución completamente condicional a posteriori $[\gamma | y, \ldots]$ será propocional a la densidad conjunta, que ahora es $[\theta, \phi | y]$. Luego,

$$\begin{aligned}[\gamma | y, \ldots] &\propto [\gamma] \exp\left(-\frac{1}{2\sigma^2} \sum_{s=1}^{18} \sum_{t=1}^{60} (y_{t,s} - \delta_t - \rho y_{t-1,s} - \gamma \operatorname{alt}(s))^2 \right) \\ &\propto [\gamma] \exp\left(-\frac{1}{2\sigma^2} \left(60\gamma^2 \sum_{s=1}^{18} \operatorname{alt}(s)^2 - 2\gamma \sum_{s=1}^{18} \sum_{t=1}^{60} \operatorname{alt}(s)(y_{t,s} - \delta_t - \rho y_{t-1,s}) \right) \right).\end{aligned}$$

  Como en el caso no jerárquico, sustituyendo $[\gamma]$ por su expresión, se deduce inmediatamente que

$$[\gamma | y, \ldots] = \mathrm{N}\left(\alpha \,\middle|\, \frac{\mu_\gamma / \tau_\gamma^2 + \sum_{s=1}^{18} \sum_{t=1}^{60} \operatorname{alt}(s)(y_{t,s} - \delta_t - \rho y_{t-1,s})/\sigma^2}{1/\tau_\gamma^2 + 60 \sum_{s=1}^{18} \operatorname{alt}(s)^2/\sigma^2}, \frac{1}{1/\tau_\gamma^2 + 60 \sum_{s=1}^{18} \operatorname{alt}(s)^2/\sigma^2} \right).$$

- Por el mismo razonamiento, las distribuciones de ρ y σ^2 pueden consultarse en el modelo no jerárquico con la sustitución indicada.
- *Distribuciones* $[\delta_1 | y, \ldots], \ldots, [\delta_{60} | y, \ldots]$. Sea $t' = 1, \ldots, 60$. Observemos que, si nos centramos en el parámetro t'-ésimo, la expresión de la verosimilitud puede simplificarse, resultando

$$\begin{aligned}[y | \theta] &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{s=1}^{18} (y_{t',s} - \delta_{t'} - \rho y_{t'-1,s} - \gamma \operatorname{alt}(s))^2 \right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \left(18\delta_{t'}^2 - 2\delta_{t'} \sum_{s=1}^{18} (y_{t',s} - \rho y_{t'-1,s} - \gamma \operatorname{alt}(s)) \right) \right).\end{aligned}$$

  Así, multiplicando la verosimilitud anterior por la expresión de la distribución a priori $[\delta_{t'} | \beta_0, \tau^2]$, se obtiene

$$\begin{aligned}[\delta_{t'} | y, \ldots] &\propto \exp\left(-\frac{1}{2} \left(\frac{1}{\tau^2} + \frac{18}{\sigma^2} \right) \delta_{t'}^2 - 2\delta_{t'} \left(\frac{\beta_0}{\tau^2} + \frac{1}{\sigma^2} \sum_{s=1}^{18} (y_{t',s} - \rho y_{t'-1,s} - \gamma \operatorname{alt}(s)) \right) \right) \\ &\propto \mathrm{N}\left(\delta_{t'} \,\middle|\, \frac{\beta_0/\tau^2 + \sum_{s=1}^{18} (y_{t',s} - \rho y_{t'-1,s} - \gamma \operatorname{alt}(s))/\sigma^2}{1/\tau^2 + 18/\sigma^2}, \frac{1}{1/\tau^2 + 18/\sigma^2} \right).\end{aligned}$$

- En cuanto a los hiperparámetros, su distribución a posteriori completamente condicional es especialmente simple de calcular, ya que, al no aparecer en la verosimilitud, se satisface que tales densidades serán proporcionales a

$$[\theta, \phi | y] \propto [\delta_1 | \beta_0, \tau^2] \cdots [\delta_{60} | \beta_0, \tau^2][\beta_0][\tau^2]. \tag{4.11}$$

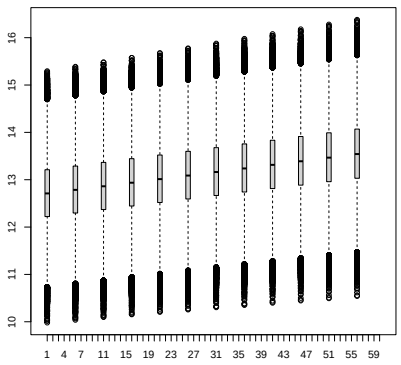
- *Distribución* $[\beta_0|y,\ldots]$. Dado que $[\beta_0|y,\ldots]$ es proporcional a (4.11), calculando los coeficientes de β_0^2 y β_0, se tiene que

$$[\beta_0|y,\ldots] \propto \exp\left(-\frac{1}{2\tau^2}\left(60\beta_0^2 - 2\beta_0\sum_{t=1}^{60}\delta_t\right)\right) \propto \mathrm{N}\left(\beta_0\left|\frac{1}{60}\sum_{t=1}^{60}\delta_t, \frac{\tau^2}{60}\right.\right).$$

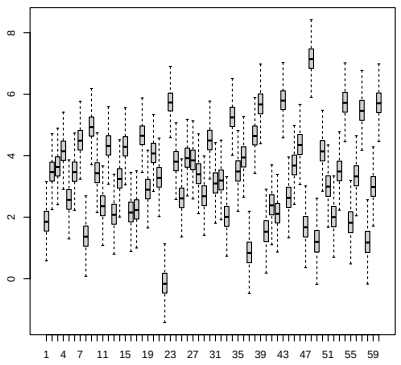
- *Distribución* $[\tau^2|y,\ldots]$. De nuevo, por (4.11) y sustituyendo $[\tau^2]$ por su expresión, puede verse inmediatamente agrupando las exponenciales que

$$\begin{aligned}[\tau^2|y,\ldots] &\propto (\tau^2)^{-(n_1+60)/2-1}\exp\left(-\frac{1}{2\tau^2}\left(\sum_{t=1}^{60}(\delta_t-\beta_0)^2 + s_1^2\right)\right)\\ &\propto \text{Inv-}\chi^2\left(\tau^2\left|n_1+60, \frac{1}{n_1+60}\left(\sum_{t=1}^{60}(\delta_t-\beta_0)^2 + s_1^2\right)\right.\right).\end{aligned}$$

Como ejemplo de los resultados, se muestra una comparativa de los *boxplots* de la tendencia en el modelo no jerárquico, reflejada por $\beta_0 + \alpha * t$, que representaremos solo cada 5 años para mayor claridad del gráfico, y su versión jerárquica, que son los efectos aleatorios $\delta_1,\ldots,\delta_{60}$. Se observa esa mayor flexibilidad del modelo jerárquico, puesto que no se impone ninguna dependencia lineal respecto de t. En el Apéndice A Sección A.3, se incluyen, además, los mismos análisis que en el modelo no jerárquico. En el Apéndice B Sección B.2 se incluye el código de `R` utilizado.



(a) *Boxplot* de $\beta_0 + \alpha t$ cada 5 años.



(b) *Boxplot* de $\delta_1,\ldots,\delta_{60}$.

4.6. Conclusiones finales

En este trabajo, se ha presentado de forma general los conceptos fundamentales y elementos básicos del análisis bayesiano, como la distribución a priori y la distribución a posteriori de un parámetro y otras medidas propias de la inferencia bayesiana. En más detalle, se han presentado el desarrollo de las distribuciones necesarias para realizar inferencia en algunos de los casos más importantes relacionados con la distribución normal. También se han introducido los modelos jerárquicos en el marco bayesiano. El interés de estos modelos es que permiten representar situaciones donde existen distintos niveles de variación y expresar la dependencia de parámetros procedentes de poblaciones agrupadas. Finalmente, se han presentado el uso de los resultados anteriores al caso particular de los modelos de regresión y los modelos de regresión jerárquicos. Estos modelos son de gran importancia en el campo de la modelización estadística. Como un ejemplo ilustrativo de la metodología presentada, se han obtenido las distribuciones completamente condicionadas necesarias para implementar un algoritmo de *Gibbs sampling* que permite estimar un modelo no jerárquico y otro jerárquico para las series de temperaturas medias del verano de un conjunto de localidades en una región.

Apéndice A

Simulación en R

A.1. Mapa de la Comunidad Autónoma de Aragón y alrededores

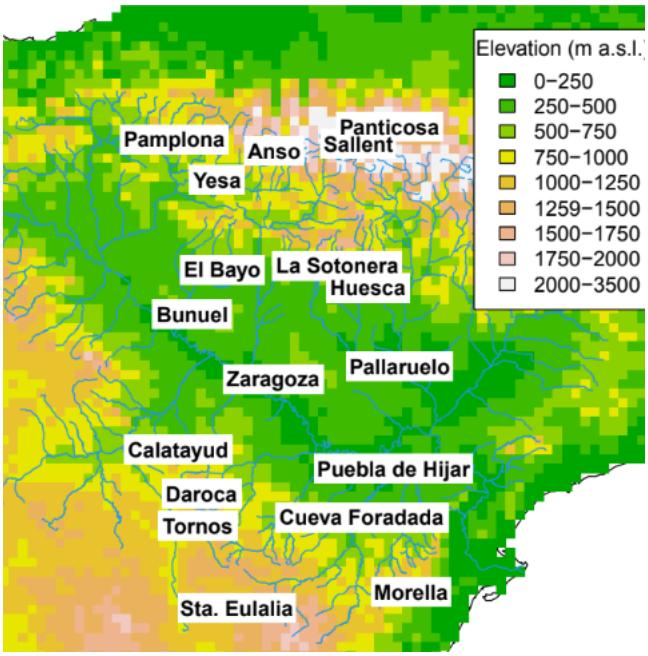


Figura A.1: Mapa geográfico de Aragón y localidades estudiadas.

A.2. Modelo no jerárquico

A la hora de implementar el algoritmo de *Gibbs sampling*, hemos escogido los hiperparámetros de las distribuciones a priori utilizadas en la Sección 4.5.1, de manera que resulten débilmente informativas considerando distribuciones con varianzas de gran magnitud —salvo ρ—, ya que no contamos con ningún estudio previo. Más precisamente,

$$\begin{aligned} \beta_0, \alpha, \gamma &\sim \mathrm{N}(0, 5000), \\ \rho &\sim \mathrm{U}(-1, 1), \\ \sigma^2 &\sim \text{Inv-}\chi^2(1, 1). \end{aligned}$$

Además, dado que el algoritmo se basa en cadenas de Markov, hemos considerado dos cadenas de 200 000 observaciones cada una con diferentes valores iniciales. Además, se han eliminado las últimas 100 000 iteraciones a modo de *burn-in* en cada cadena.

	Primera cadena	Segunda cadena
$\beta_0^{(1)}$	0	1
$\alpha^{(1)}$	0	1
$\rho^{(1)}$	0	0,2
$\gamma^{(1)}$	0	1
$(\sigma^2)^{(1)}$	1	2

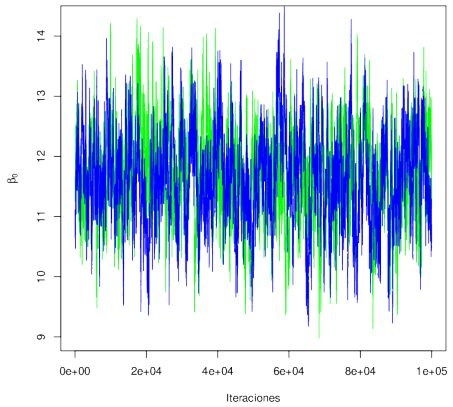
Cuadro A.1: Valores iniciales de los parámetros en las cadenas de Markov para el modelo no jerárquico.

A.2.1. Convergencia de las cadenas de Markov

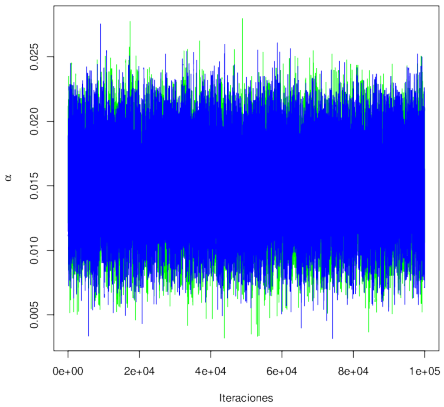
En primer lugar, para garantizar que los resultados que expondremos a continuación son válidos, recogemos algunos diagnósticos de convergencia, que son, los *traceplots*, que representan los valores que ha ido tomando cada parámetro en cada iteración, y el test de diagnóstico basado en el factor de reducción de escala potencial, generalmente denotado por $\hat{R}$.

Traceplots

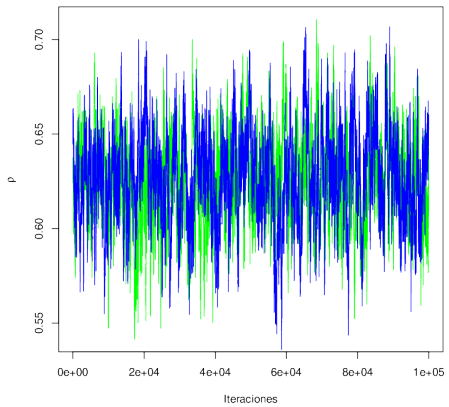
Representando los *traceplots*, se observa claramente como, en todos los casos, los valores de todos los parámetros oscilan en torno a algún valor. Habiendo eliminando las primeras muestras, vemos como las gráficas están muy concentradas y las dos cadenas, que se representan en colores diferentes, se solapan, lo que es signo de convergencia.



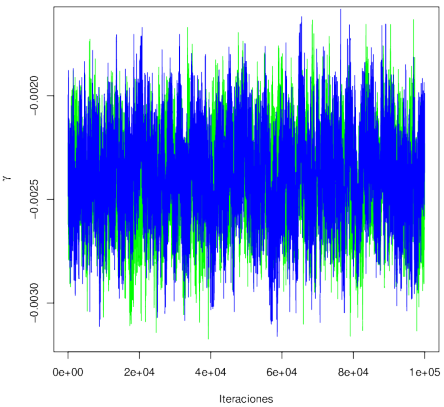
(a) *Traceplot* de β_0



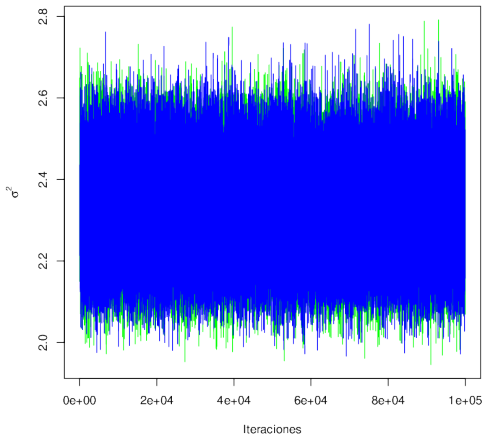
(b) *Traceplot* de α



(c) *Traceplot* de ρ



(d) *Traceplot* de γ



(e) *Traceplot* de σ^2

Figura A.2: *Traceplots* de las últimas 100 000 observaciones para el modelo no jerárquico. La primera cadena se representa en verde y la segunda en azul.

Factor de reducción de escala potencial $\hat{R}$

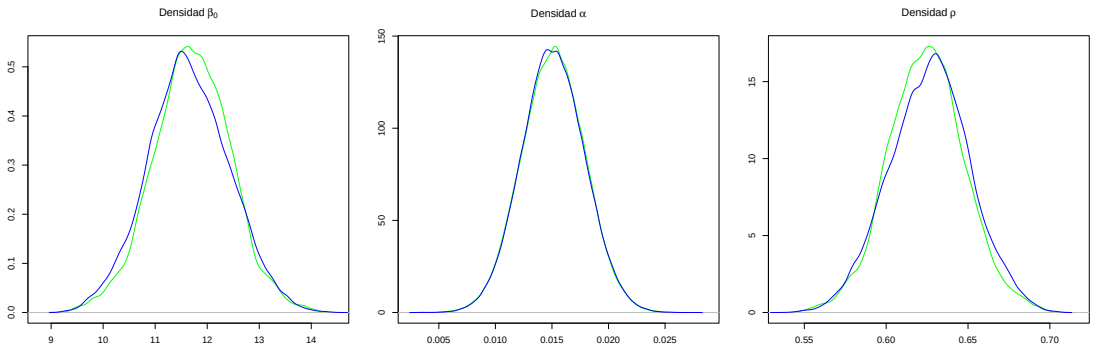
Según Gelman, A. y Rubin, D. B. (1992) [7], $\hat{R}$ es un valor siempre mayor o igual que 1 y, cuanto más próximo es su valor a 1, con seguridad puede afirmarse que la cadena converge. El criterio utilizado más habitual es $\hat{R} < 1{,}1$. Esto, junto a las gráficas anteriores, evidencia la convergencia de ambas cadenas de Markov, ya que, incluso para la cota superior al 95 % del coeficiente se satisface el criterio. Para implementarlo hemos utilizado la función `gelman.diag()` —vid. Apéndice B Sección B.1.

	Estimación $\hat{R}$	Cota superior al 95 %
β_0	1,0040	1,0145
α	1,0000	1,0001
ρ	1,0038	1,0137
γ	1,0024	1,0094
σ^2	1,0000	1,0000

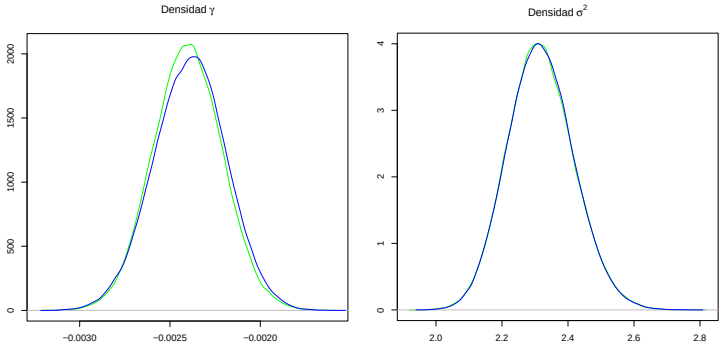
Cuadro A.2: Análisis de convergencia para el modelo no jerárquico. Criterio por factor $\hat{R}$.

A.2.2. Densidades a posteriori marginales

Tras comprobar la convergencia de ambas cadenas, podemos utilizar un estimador *kernel* de densidades dadas las últimas 100 000 observaciones generadas para cada parámetro y en cada cadena. En las gráficas se observa como las distribuciones estacionarias a la que se aproximan las observaciones, que son las distribuciones marginales a posteriori de los parámetros, son asintóticamente normales.



(a) Densidad marginal de β_0. (b) Densidad marginal de α. (c) Densidad marginal de ρ.



(d) Densidad marginal de γ. (e) Densidad marginal de σ^2.

Figura A.3: Densidades a posteriori marginales estimadas para el modelo no jerárquico. En verde se presenta la densidad estimada a partir de la primera cadena de Markov y en azul, la densidad estimada a partir de la segunda.

A.2.3. Resumen númerico. Esperanzas a posteriori e intervalos de credibilidad de los parámetros

Las esperanzas pueden aproximarse por la media de todas las últimas 100 000 observaciones que hemos generado de la primera cadena. Para los intervalos de credibilidad, es suficiente con aproximar los cuantiles poblacionales por los cuantiles muestrales 0,025 y 0,975.

	Esperanzas a posteriori	Extremo inferior IC	Extremo superior IC
β_0	11,7096	10,2025	13,1921
α	0,0151	0,0097	0,0205
ρ	0,6246	0,5780	0,6718
γ	−0,0024	−0,0028	−0,0020
σ^2	2,3176	2,1299	2,5226

Cuadro A.3: Resumen de los resultados para el modelo no jerárquico.

Especialmente interesante resulta que en los intervalos de credibilidad de α y ρ no esté contenido el 0, ya que esto es indicativo de la existencia real de la tendencia creciente de la temperatura respecto del tiempo. En suma, los datos y el modelo ajustado parecen evidenciar la existencia del calentamiento global.

Por otra parte, también es notorio es signo de γ que es negativo al 95 % de confianza. En efecto, en las zonas de mayor altitud, como los Pirineos, las temperaturas medias tienden a ser menores.

A.3. Modelo jerárquico

Para simulación del modelo jerárquico, volvemos a considerar distribuciones a priori débilmente informativas tanto para los parámetros ρ, γ y σ^2 como para los hiperparámetros β_0 y τ^2 de δ_t para $t = 1, \ldots, 60$. En concreto, las elecciones tomadas son las siguientes:

$$\begin{aligned} \delta_t | \beta_0, \tau^2 &\sim \mathrm{N}(\beta_0, \tau^2), \quad t = 1, \ldots, 60, \\ \beta_0 &\sim \mathrm{U}(\mathbb{R}), \\ \tau^2 &\sim \text{Inv-}\chi^2(1,1), \\ \gamma &\sim \mathrm{N}(0, 5000), \\ \rho &\sim \mathrm{U}(-1,1), \\ \sigma^2 &\sim \text{Inv-}\chi^2(1,1). \end{aligned}$$

Además, como en el caso no jerárquico, también se consideran dos cadenas de Markov con diferentes valores iniciales para estudiar la convergencia.

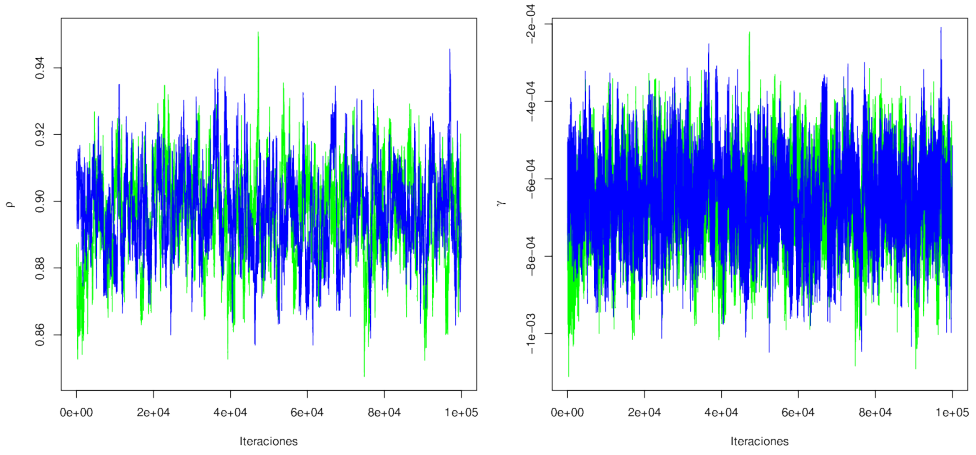
	Primera cadena	Segunda cadena
$\delta_t^{(1)}$	0	1
$\rho^{(1)}$	0	0,2
$\gamma^{(1)}$	0	1
$(\sigma^2)^{(1)}$	1	2
$\beta_0^{(1)}$	0	1
$(\tau^2)^{(1)}$	1	2

Cuadro A.4: Valores iniciales de los parámetros en las cadenas de Markov para el modelo jerárquico.

Los métodos y diagnósticos que usaremos serán los mismos y también expondremos los mismos puntos: análisis de la convergencia, densidades marginales a posteriori y resúmenes numéricos. En este caso, sin embargo, no se incluirán los parámetros $\delta_1, \ldots, \delta_{60}$, puesto que eso supondría un uso innecesario de espacio. En su lugar, recogemos un *boxplot* de todos ellos en la Sección A.3.2.

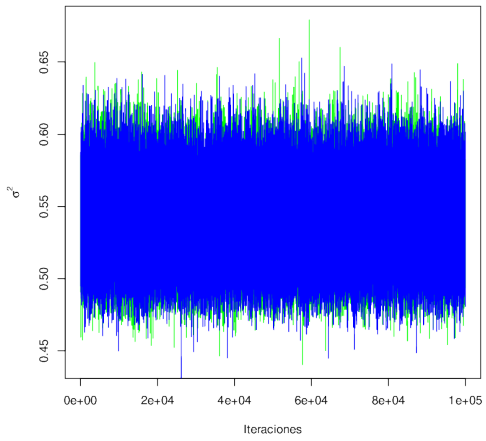
A.3.1. Convergencia de las cadenas de Markov

Traceplots

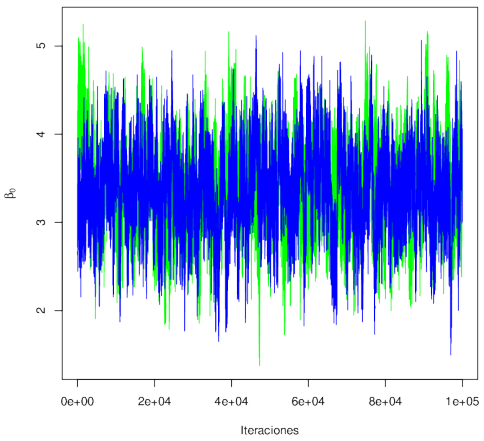


(a) *Traceplot* de ρ

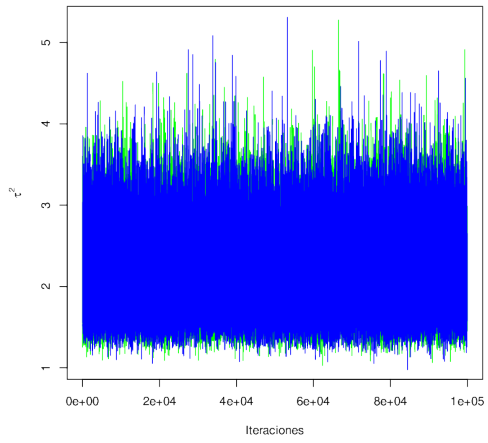
(b) *Traceplot* de γ



(c) *Traceplot* de σ^2



(d) *Traceplot* de β_0



(e) *Traceplot* de τ^2

Figura A.4: *Traceplots* de las últimas 100 000 observaciones para el modelo jerárquico. La primera cadena se representa en verde y la segunda en azul.

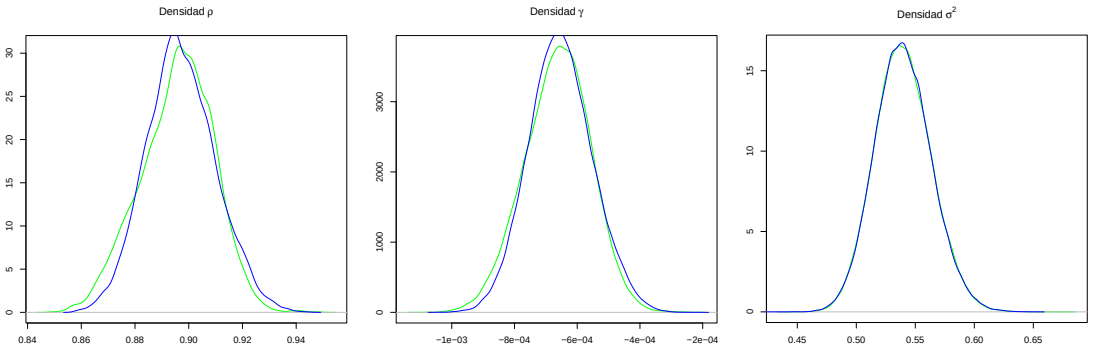
Factor de reducción de escala potencial $\hat{R}$

	Estimación $\hat{R}$	Cota superior al 95 %
ρ	1,0059	1,0238
γ	1,0042	1,0178
σ^2	1,0000	1,0000
β_0	1,0049	1,0198
τ^2	1,0000	1,0001

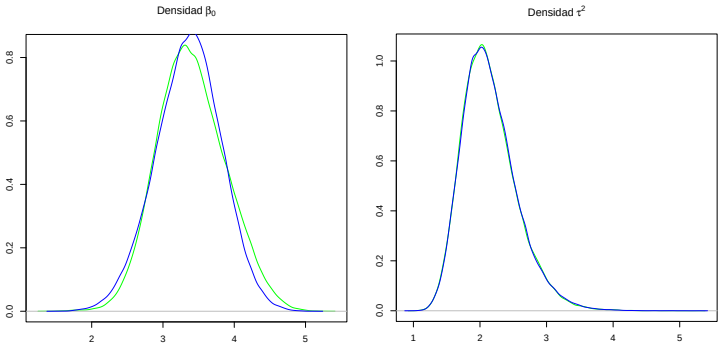
Cuadro A.5: Análisis de convergencia para el modelo jerárquico. Factor $\hat{R}$.

Además, pese a no recogerse por el espacio que ello ocuparía, el diagnóstico también se ha llevado acabo para los parámetros δ_t, $t = 1, \ldots, 60$. Los factores obtenidos han sido también muy próximos a 1, por lo que tenemos evidencia de la convergencia de todos los parámetros.

A.3.2. Densidades a posteriori marginales



(a) Densidad marginal de ρ. (b) Densidad marginal de γ. (c) Densidad marginal de σ^2.



(d) Densidad marginal de β_0. (e) Densidad marginal de τ^2.

Figura A.5: Densidades a posteriori marginales estimadas para el modelo jerárquico. En verde se presenta la densidad estimada a partir de la primera cadena de Markov y en azul, la densidad estimada a partir de la segunda.

A.3.3. Resumen númerico. Esperanzas a posteriori e intervalos de credibilidad de los parámetros

	Esperanzas a posteriori	Extremo inferior IC	Extremo superior IC
ρ	0,8957	0,8708	0,9187
γ	−0,0007	−0,0009	−0,0005
σ^2	0,5398	0,4949	0,5887
β_0	3,4039	2,5180	4,3855
τ^2	2,1446	1,4823	3,0844

Cuadro A.6: Resumen de los resultados para el modelo jerárquico.

Apéndice B

Implementación del código de R

B.1. Modelo no jerárquico

```
### IMPLEMENTACIÓN DE GIBBS SAMPLING. MODELO NO JERÁRQUICO.

library(extraDistr)
library(xtable)
library(coda)

# Datos de temperaturas y altitudes

temp <- readRDS("meanTempAragonJJA19562015.rds.ds")
elev <- readRDS("elev.rds")

# Matriz de datos. Filas por años y columnas por localidades.
# Generamos la fila 1 de y, que se corresponde con y_0,s.
# En general, la fila t refiere a las temperaturas del año t-1.

y <- matrix(nrow = 61, ncol = 18)
y[2:61,] <- temp
y[1,] <- apply(temp, MARGIN = 2, FUN = mean)

# Definimos la matriz que contiene los 60 años (1,...,60) en cada columna.

t <- matrix(data = 1:60, nrow = 60, ncol = 18, byrow = FALSE)

# Definimos la matriz alt que contiene las 18 de las localidades en cada fila.

alt <- matrix(data = elev$altitude, nrow = 60, ncol = 18, byrow = TRUE)

# Parámetros

numberOfSamples <- 200000

beta_0 <- rep(0, numberOfSamples)
alpha <- rep(0, numberOfSamples)
gamma <- rep(0, numberOfSamples)
rho <- rep(0, numberOfSamples)
sigma2 <- rep(0, numberOfSamples)

beta_02 <- rep(0, numberOfSamples)
alpha2 <- rep(0, numberOfSamples)
gamma2 <- rep(0, numberOfSamples)
rho2 <- rep(0, numberOfSamples)
```

```
sigma22 <- rep(0, numberOfSamples)

# Valores iniciales de los parámetros.

beta_0[1] <- 0
alpha[1] <- 0
gamma[1] <- 0
sigma2[1] <- 1
rho[1] <- 0

beta_02[1] <- 1
alpha2[1] <- 1
gamma2[1] <- 1
sigma22[1] <- 2
rho2[1] <- 0.2

# Iteraciones de Gibbs sampling.

# La constante sumSquares es necesaria la función sampleRho
sumSquares <- sum(y[ -61,]^2)
# La constante sumHeightsSq es necesaria la función sampleGamma
sumHeightsSq <- sum(alt[1,]^2)*60

for (i in 2:numberOfSamples) {
  beta_0[i] <- sampleBeta0(0, 5000, alpha[i-1], gamma[i-1], rho[i-1],
                           sigma2[i-1])
  alpha[i] <- sampleAlpha(0, 5000, beta_0[i], gamma[i-1], rho[i-1],
                          sigma2[i-1])
  gamma[i] <- sampleGamma(0, 5000, beta_0[i], alpha[i], rho[i-1],
                          sigma2[i-1])
  rho[i] <- sampleRho(beta_0[i], alpha[i], gamma[i], sigma2[i-1])

  sigma2[i] <- sampleSigma2(1, 1, beta_0[i], alpha[i], rho[i],
                            gamma[i])
}

for (i in 2:numberOfSamples) {
  beta_02[i] <- sampleBeta0(0, 5000, alpha2[i-1], gamma2[i-1], rho2[i-1],
                            sigma22[i-1])
  alpha2[i] <- sampleAlpha(0, 5000, beta_02[i], gamma2[i-1], rho2[i-1],
                           sigma22[i-1])
  gamma2[i] <- sampleGamma(0, 5000, beta_02[i], alpha2[i], rho2[i-1],
                           sigma22[i-1])
  rho2[i] <- sampleRho(beta_02[i], alpha2[i], gamma2[i], sigma22[i-1])

  sigma22[i] <- sampleSigma2(1, 1, beta_02[i], alpha2[i], rho2[i],
                             gamma2[i])
}



# Implementación de las distribuciones condicionales a posteriori y muestreo.

sampleBeta0 <- function(mu, tau2, alpha, gamma, rho, sigma2) {
  suma <- sum(y[-1,] - alpha * t - rho * y[-61,] - gamma * alt) / sigma2
  precision <- (1 / tau2 + 1080 / sigma2)
  media <- (mu / tau2 + suma) / precision
```

```
  return(rnorm(n = 1, mean = media, sd = 1 / sqrt(precision)))
}

sampleAlpha <- function(mu, tau2, beta_0, gamma, rho, sigma2) {
  suma <- sum(t * (y[-1,] - beta_0 - rho * y[-61,] - gamma * alt)) / sigma2
  precision <- (1 / tau2 +  1328580 / sigma2)
  media <- (mu / tau2 + suma) / precision
  return(rnorm(n = 1, mean = media, sd = 1 / sqrt(precision)))
}

sampleGamma <- function(mu, tau2, beta_0, alpha, rho, sigma2) {
  suma <- sum(alt * (y[-1,] - beta_0 - alpha * t - rho * y[-61,])) / sigma2
  precision <- (1 / tau2 + sumHeightsSq / sigma2)
  media <- (mu / tau2 + suma) / precision
  return(rnorm(n = 1, mean = media, sd = 1 / sqrt(precision)))
}

sampleRho <- function(beta_0, alpha, gamma, sigma2) {
  suma <- sum(y[-61,] * (y[-1,] - beta_0 - alpha * t - gamma * alt))
  media <- suma / sumSquares
  varianza <- sigma2 / sumSquares
  return(rtnorm(n = 1, mean = media, sd = sqrt(varianza), -1, 1))
}

sampleSigma2 <- function(n, s2, beta_0, alpha, rho, gamma) {
  suma <- sum((y[-1,] - beta_0 - alpha * t - rho * y[-61,] - gamma * alt)^2)
  scale2 <- n * s2 + suma
  return(scale2 / rchisq(n = 1, df = 1080 + n))
}

# Traceplots de ambas cadenas

plot(beta_0[1:1], type='l', col = 'blue', xlab = 'Iteraciones',
     ylab = expression(beta[0]))
lines(beta_02[100001:200000], col = 'red')
plot(alpha[100001:200000], type='l', col = 'blue', xlab = 'Iteraciones',
     ylab = expression(alpha))
lines(alpha2[100001:200000], col = 'red')
plot(rho[100001:200000], type='l', col = 'blue', xlab = 'Iteraciones',
     ylab = expression(rho))
lines(rho2[100001:200000], col = 'red')
plot(gamma[100001:200000], type='l', col = 'blue', xlab = 'Iteraciones',
     ylab = expression(gamma))
lines(gamma2[100001:200000], col = 'red')
plot(sigma2[100001:200000], type='l', col = 'blue', xlab = 'Iteraciones',
     ylab = expression(sigma^2))
lines(sigma22[100001:200000], col = 'red')

# Factor de reducción de escala pontecial RHat.

gelman.diag(mcmc.list(as.mcmc(beta_0[100001:200000]),
                      as.mcmc(beta_02[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(alpha[100001:200000]),
                      as.mcmc(alpha2[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(rho[100001:200000]),
                      as.mcmc(rho2[100001:200000])),
```

```
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(gamma[100001:200000]),
                      as.mcmc(gamma2[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(sigma2[100001:200000]),
                      as.mcmc(sigma22[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)


# Densidades marginales a posteriori

plot(density(beta_0[100001:200000]), col = 'green')
lines(density(beta_02[100001:200000]), col = 'blue')
plot(density(alpha[100001:200000]), col = 'green')
lines(density(alpha2[100001:200000]), col = 'blue')
plot(density(rho[100001:200000]), col = 'green')
lines(density(rho2[100001:200000]), col = 'blue')
plot(density(gamma[100001:200000]), col = 'green')
lines(density(gamma2[100001:200000]), col = 'blue')
plot(density(sigma2[100001:200000]), col = 'green')
lines(density(sigma22[100001:200000]), col = 'blue')

# Valores aproximados de la esperanza a posteriori

mean(beta_0[100001:200000])
mean(alpha[100001:200000])
mean(rho[100001:200000])
mean(gamma[100001:200000])
mean(sigma2[100001:200000])

#Intervalos de credibilidad

matrix <- cbind(beta_0[100001:200000], alpha[100001:200000],
                rho[100001:200000], gamma[100001:200000],
                sigma2[100001:200000])

lowerBoundsCI <- apply(matrix, MARGIN = 2, FUN = function(x)
  quantile(x, probs = 0.025))

upperBoundsCI <- apply(matrix, MARGIN = 2, FUN = function(x)
  quantile(x, probs = 0.975))
```

B.2. Modelo jerárquico

```
### IMPLEMETANCIÓN MODELO JERÁRQUICO.

# Cargamos las mismas librerías.

# Tomamos los mismos datos y, alt.

numberOfSamples <- 200000

tau2 <- rep(0, numberOfSamples)
beta_0 <- rep(0, numberOfSamples)
gamma <- rep(0, numberOfSamples)
rho <- rep(0, numberOfSamples)
sigma2 <- rep(0, numberOfSamples)
```

```
# La primera fila de deltas corresponde a delta_0. En general, la fila t
# corresponde a delta_{t-1}
deltas <- matrix(data = 0, nrow = numberOfSamples, ncol = 60)

tau22 <- rep(0, numberOfSamples)
beta_02 <- rep(0, numberOfSamples)
gamma2 <- rep(0, numberOfSamples)
rho2 <- rep(0, numberOfSamples)
sigma22 <- rep(0, numberOfSamples)
# La primera fila de deltas2 corresponde a delta_0. En general, la fila t
# corresponde a delta_{t-1}
deltas2 <- matrix(data = 0, nrow = numberOfSamples, ncol = 60)

# Valores iniciales de los parámetros.

tau2[1] <- 1
beta_0[1] <- 0
gamma[1] <- 0
sigma2[1] <- 1
rho[1] <- 0
deltas[1,] <- rep(0, 60)

tau22[1] <- 2
beta_02[1] <- 1
gamma2[1] <- 1
sigma22[1] <- 2
rho2[1] <- 0.2
deltas2[1,] <- rep(1, 60)

# Iteraciones de Gibbs sampling.

sumSquares <- sum(y[ -61,]^2)
sumHeightsSq <- sum(alt[1,]^2)*60

for (i in 2:numberOfSamples) {
  for (j in 1:60) {
    deltas[i,j] <- sampleDeltas(j, beta_0[i-1], tau2[i-1], gamma[i-1],
                                rho[i-1], sigma2[i-1])
    mDeltas <- matrix(data = deltas[i,], nrow = 60, ncol = 18)
  }
  gamma[i] <- sampleGamma(0, 5000, mDeltas, rho[i-1], sigma2[i-1])

  rho[i] <- sampleRho(mDeltas, gamma[i], sigma2[i-1])

  sigma2[i] <- sampleSigma2(1, 1, mDeltas, rho[i], gamma[i])

  beta_0[i] <- sampleBeta_0(deltas[i,], tau2[i-1])

  tau2[i] <- sampleTau2(1, 1, beta_0[i], deltas[i,])
}

for (i in 2:numberOfSamples) {
  for (j in 1:60) {
    deltas2[i,j] <- sampleDeltas(j, beta_02[i-1], tau22[i-1], gamma2[i-1],
                                 rho2[i-1], sigma22[i-1])
    mDeltas <- matrix(data = deltas2[i,], nrow = 60, ncol = 18)
  }
  gamma2[i] <- sampleGamma(0, 5000, mDeltas, rho2[i-1],
```

```
                              sigma22[i-1])
  rho2[i] <- sampleRho(mDeltas, gamma2[i], sigma22[i-1])

  sigma22[i] <- sampleSigma2(1, 1, mDeltas, rho2[i], gamma2[i])

  beta_02[i] <- sampleBeta_0(deltas2[i,], tau22[i-1])

  tau22[i] <- sampleTau2(1, 1, beta_02[i], deltas2[i,])
}

# Implementación de las distribuciones condicionales a posteriori y muestreo.

sampleDeltas <- function(j, beta_0, tau2, gamma, rho, sigma2) {
  suma <- sum((y[j + 1,] - rho * y[j,] - gamma * alt[1,]) / sigma2)
  precision <- (1 / tau2 + 18 / sigma2)
  media <- (beta_0 / tau2 + suma) / precision
  return(rnorm(n = 1, mean = media, sd = 1 / sqrt(precision)))
}

sampleGamma <- function(mu, tau2, deltas, rho, sigma2) {
  suma <- sum(alt * (y[-1,] - deltas - rho * y[-61,])) / sigma2
  precision <- (1 / tau2 + sumHeightsSq / sigma2)
  media <- (mu / tau2 + suma) / precision
  return(rnorm(n = 1, mean = media, sd = 1 / sqrt(precision)))
}

sampleRho <- function(deltas, gamma, sigma2) {
  suma <- sum(y[-61,] * (y[-1,] - deltas - gamma * alt))
  media <- suma / sumSquares
  varianza <- sigma2 / sumSquares
  return(rtnorm(n = 1, mean = media, sd = sqrt(varianza), -1, 1))
}

sampleSigma2 <- function(n, s2, deltas, rho, gamma) {
  suma <- sum((y[-1,] - deltas - rho * y[-61,] - gamma * alt)^2)
  scale2 <- n * s2 + suma
  return(scale2 / rchisq(n = 1, df = 1080 + n))
}

sampleBeta_0 <- function(deltas, tau2) {
  return(rnorm(n = 1, mean = (sum(deltas)) / 60, sd = sqrt(tau2 / 60)))
}

sampleTau2 <- function(n, s2, beta_0, deltas) {
  scale <- sum((deltas - beta_0)^2) + s2
  return(scale / rchisq(n = 1, df = n + 60))
}

# Traceplots de ambas cadenas

plot(rho[100001:200000], type='l', col = 'blue', xlab = 'Iteraciones',
     ylab = expression(rho))
lines(rho2[100001:200000], col = 'green')
plot(gamma[100001:200000], type='l', col = 'green', xlab = 'Iteraciones',
     ylab = expression(gamma))
lines(gamma2[100001:200000], col = 'blue')
plot(sigma2[100001:200000], type='l', col = 'green', xlab = 'Iteraciones',
     ylab = expression(sigma^2))
```

```
lines(sigma22[100001:200000], col = 'blue')
plot(beta_0[100001:200000], type='l', col = 'green', xlab = 'Iteraciones',
     ylab = expression(beta[0]))
lines(beta_02[100001:200000], col = 'blue')
plot(tau2[100001:200000], type='l', col = 'green', xlab = 'Iteraciones',
     ylab = expression(tau^2))
lines(tau22[100001:200000], col = 'blue')

# Factor de reducción de escala pontecial RHat.

gelman.diag(mcmc.list(as.mcmc(rho[100001:200000]),
                      as.mcmc(rho2[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(gamma[100001:200000]),
                      as.mcmc(gamma2[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(sigma2[100001:200000]),
                      as.mcmc(sigma22[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(beta_0[100001:200000]),
                      as.mcmc(beta_02[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
gelman.diag(mcmc.list(as.mcmc(tau2[100001:200000]),
                      as.mcmc(tau22[100001:200000])),
            confidence = 0.95, autoburnin = FALSE)
for (i in 1:60) {
  gelman.diag(mcmc.list(as.mcmc(deltas[100001:200000, i]),
                        as.mcmc(deltas2[100001:200000, i])),
              confidence = 0.95, autoburnin = FALSE)
}

# Densidades marginales a posteriori

plot(density(rho[100001:200000]), col = 'blue')
lines(density(rho2[100001:200000]), col = 'red')
plot(density(gamma[100001:200000]), col = 'blue')
lines(density(gamma2[100001:200000]), col = 'red')
plot(density(sigma2[100001:200000]), col = 'blue')
lines(density(sigma22[100001:200000]), col = 'red')
plot(density(beta_0[100001:200000]), col = 'blue')
lines(density(beta_02[100001:200000]), col = 'red')
plot(density(tau2[100001:200000]), col = 'blue')
lines(density(tau22[100001:200000]), col = 'red')

# Boxplot de los deltas

boxplot(deltas[100001:200000, ], outline = FALSE)

# Estimaciones de las esperanzas a posteriori

mean(rho[100001:200000])
mean(gamma[100001:200000])
mean(sigma2[100001:200000])
mean(beta_0[100001:200000])
mean(tau2[100001:200000])

# Intervalos de credibilidad
```

```
matrix <- cbind(rho[100001:200000], gamma[100001:200000],
                sigma2[100001:200000], beta_0[100001:200000],
                tau2[100001:200000])

lowerBoundsCI <- apply(matrix, MARGIN = 2, FUN = function(x)
  quantile(x, probs = 0.025))

upperBoundsCI <- apply(matrix, MARGIN = 2, FUN = function(x)
  quantile(x, probs = 0.975))
```

Bibliografía

[1] BAYES, T., *An essay towards solving a problem in the doctrine of chances*, Philosophical Transactions of the Royal Society, 1763. Reeditado, con una nota biográfica por G. A. Barnard, en Biometrika (1958), vol. 45.

[2] BROCKWELL, P. J. AND DAVIS, R. A., *Introduction to time series and forecasting*, Springer, 2006.

[3] CASTILLO-MATEO, J., LAFUENTE, M., ASÍN, J., CEBRIÁN, A. C., GELFAND, A. E. AND ABAURREA, J., *Spatial modeling of day-within-year temperature time series: an examination of daily maximum temperatures in Aragón, Spain*, Journal of Agricultural, Biological and Environmental Statistics, 2022.

[4] FISHER, R., *The design of experiments*, Oliver and Boyd, Edinburgh, 1949.

[5] GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. AND RUBIN, D. B., *Bayesian Data Analysis*, Third Edition, Texts in Statistical Science, Chapman & Hall, 2021.

[6] GELMAN, A. AND HILL, J., *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambrigde University Press, 2007.

[7] GELMAN, A. AND RUBIN, D. R., *Inference from Iterative Simulation using Multiple Sequences*, Statistical Science, 7, 1992.

[8] MCSHANE, B. B., DAVID, G., GELMAN, A., ROBERT, B. C. AND TACKETT, J. L., *Abandon Statistical Significance*, The American Statistician, 73, pp. 235-245, 2019.

[9] ROBERT, C. P. AND CASELLA, G., *Monte Carlo Statistical Methods*, disponible en `https://www.researchgate.net/profile/Christian_Robert2/publication/2681158_Monte_Carlo_Statistical_Methods/links/00b49535ccaf6ccc8f000000/Monte-Carlo-Statistical-Methods.pdf`, 1998.

[10] ROBERT, C. P. AND CASELLA, G., *Introducing Monte Carlo methods with R*, Springer, 2010.