Estadística Bayesiana

Métodos Computacionales y Algunas Aplicaciones

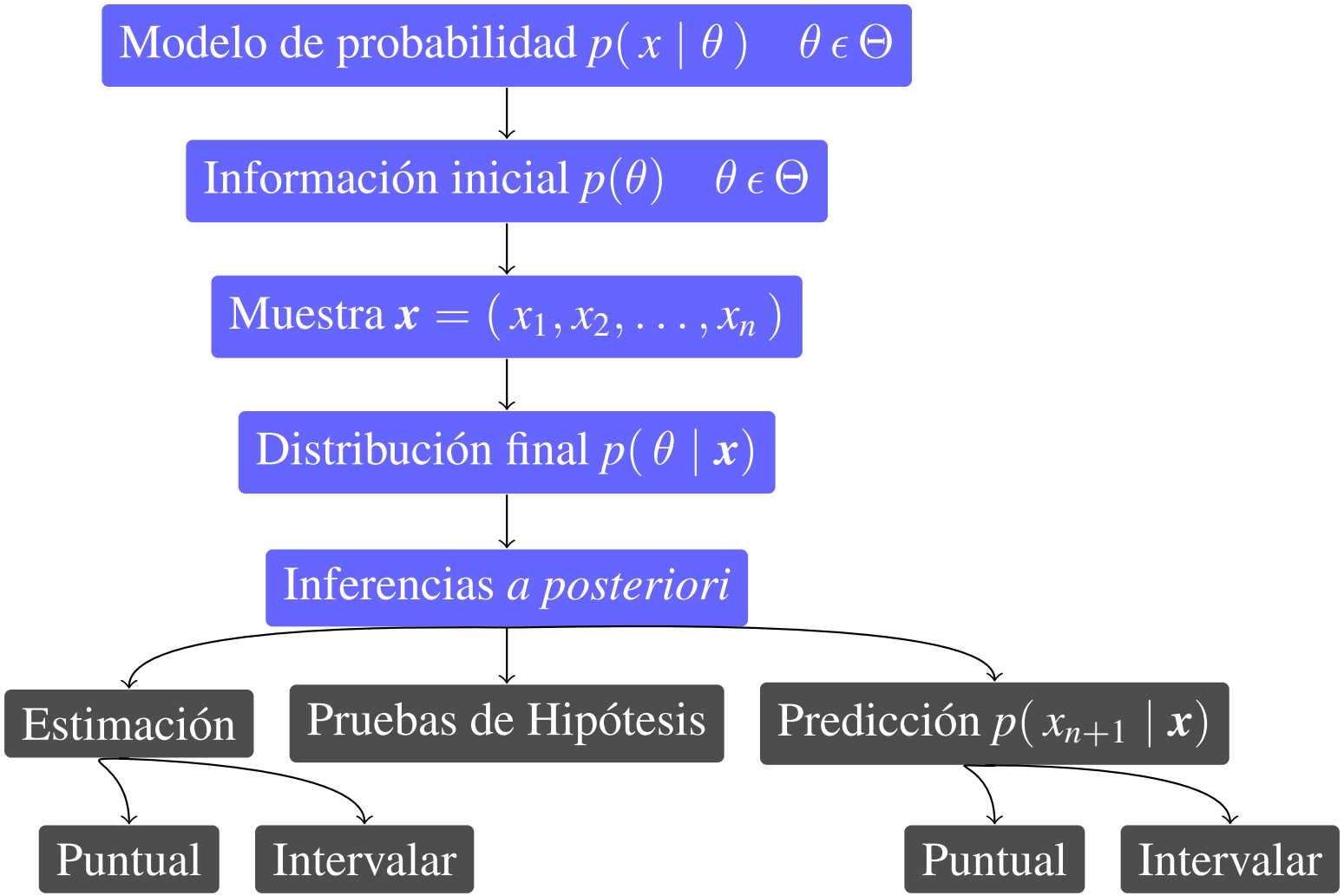
Antonio Soriano Flores

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

asoriano@sigma.iimas.unam.mx

XXXI Foro de Estadística - Universidad Autónoma Chapingo

Estadística Bayesiana



Introducción

En términos generales, en la estadística bayesiana surge la necesidad de llevar a cabo el cálculo de ciertas integrales que pueden llegar a ser analíticamente difíciles de resolver:

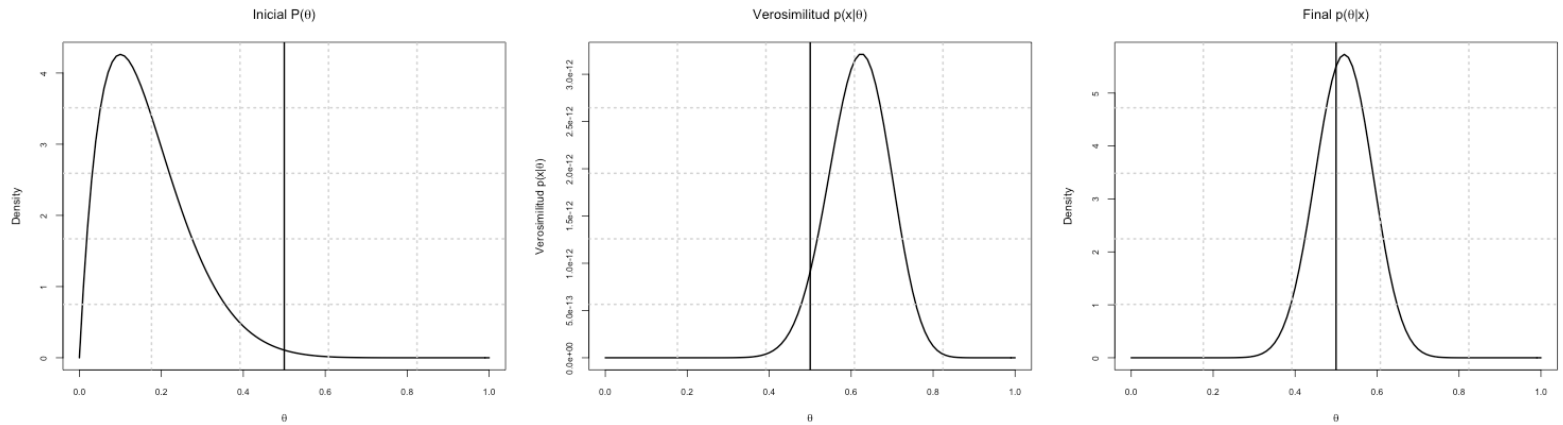
$$\begin{aligned} p\left(\theta|\underline{x}\right) &\propto p\left(\underline{x}|\theta\right) p\left(\theta\right) \\ p\left(x^{*}|\underline{x}\right) &\propto \int_{\Theta} p\left(x^{*}|\theta\right) p\left(\theta|\underline{x}\right) d\theta \\ p\left(\text{Desconocido}|\text{Observado}\right) &\propto \ldots \end{aligned}$$

Algunas veces, la solución analítica es relativamente fácil por medio del uso de las **Familias Conjugadas**

$$x \sim Bernoulli(x|\theta); \quad \theta \sim Beta(\theta|\alpha_0, \beta_0);$$

Entonces:

$$\underbrace{\theta \sim Beta\,(\alpha_0, \beta_0)}_{Inicial} \quad + \quad \underbrace{p\,(\underline{x}|\theta)}_{Muestra} \Rightarrow \underbrace{\theta|\underline{x} \sim Beta\left(\sum_{i=1}^{n} x_i + \alpha_0, n - \sum_{i=1}^{n} x_i + \beta_0\right)}_{Final}$$



Ver video: Final.mp4

En este ejemplo, si estamos interesados en modelar una nueva observación x^*, se puede calcular:

- Densidad predictiva inicial:

$$\begin{aligned}
p(x^*) &= \int p(x^*, \theta)\, d\theta = \int p(x^*|\theta)\, p(\theta)\, d\theta \\
&= \frac{1}{B(\alpha_0, \beta_0)} \int_0^1 \theta^{x^*} (1-\theta)^{1-x^*} \theta^{\alpha_0 - 1} (1-\theta)^{\beta_0 - 1}\, d\theta \\
&= \frac{1}{B(\alpha_0, \beta_0)} \int_0^1 \theta^{x^* + \alpha_0 - 1} (1-\theta)^{\beta_0 - x^*}\, d\theta \\
&= \frac{B(x^* + \alpha_0, \beta_0 - x^* + 1)}{B(\alpha_0, \beta_0)} \\
&= \frac{\Gamma(x^* + \alpha_0)\Gamma(\beta_0 - x^* + 1)}{\Gamma(\alpha_0 + \beta_0 + 1)} \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}
\end{aligned}$$

Entonces:

$$p(0) = \frac{\beta_0}{\alpha_0 + \beta_0}; \qquad p(1) = \frac{\alpha_0}{\alpha_0 + \beta_0}$$

*

Haciendo $\alpha_1 = \sum_{i=1}^{n} x_i + \alpha_0$ y $\beta_1 = n - \sum_{i=1}^{n} x_i + \beta_0$,

- Densidad predictiva final:

$$
\begin{aligned}
p\left(x^*|\underline{x}\right) &= \int p\left(x^*|\theta\right) p\left(\theta|\underline{x}\right) d\theta \\
&= \frac{1}{B\left(\alpha_1, \beta_1\right)} \int_0^1 \theta^{x^*} \left(1-\theta\right)^{1-x^*} \theta^{\alpha_1 - 1} \left(1-\theta\right)^{\beta_1 - 1} d\theta \\
&= \frac{1}{B\left(\alpha_1, \beta_1\right)} \int_0^1 \theta^{x^* + \alpha_1 - 1} \left(1-\theta\right)^{\beta_1 - x^*} d\theta \\
&= \frac{B\left(x^* + \alpha_1, \beta_1 - x^* + 1\right)}{B\left(\alpha_1, \beta_1\right)} \\
&= \frac{\Gamma(x^* + \alpha_1)\Gamma\left(\beta_0 - x^* + 1\right)}{\Gamma\left(\alpha_1 + \beta_1 + 1\right)} \frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\beta_1\right)}
\end{aligned}
$$

Entonces:

$$
p\left(0\right) = \frac{\beta_1}{\alpha_1 + \beta_1} = \frac{n - \sum_{i=1}^{n} x_i + \beta_0}{n + \alpha_0 + \beta_0}; \qquad p\left(1\right) = \frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\sum_{i=1}^{n} x_i + \alpha_0}{n + \alpha_0 + \beta_0}
$$

*

Algunas familias conjugadas:

- $x \sim Poisson\,(x|\lambda) \rightarrow \lambda \sim Gamma\,(\lambda|\alpha_0, \beta_0)$
- $x \sim Binomial\,(x|n^*, \theta) \rightarrow \theta \sim Beta\,(\theta|\alpha_0, \beta_0)$
- $x \sim Multinomial\,(x|p_1, \ldots, p_k, n^*) \rightarrow (p_1, \ldots, p_k) \sim Dir\,(p_1, \ldots, p_k|\alpha_1, \ldots, \alpha_k)$
- $x \sim Geometric\,(x|\theta) \rightarrow \theta \sim Beta\,(\theta|\alpha_0, \beta_0)$
- $x \sim U\,(x|0, \theta) \rightarrow \theta \sim Pareto\,(\theta|\alpha_0, \beta_0)$
- $x \sim Gamma\,(x|\alpha^*, \beta) \rightarrow \beta \sim Gamma\,(\beta|\alpha_0, \beta_0)$
- $x \sim Normal\,(x|\mu, \tau^*) \rightarrow \mu \sim Normal\,(\mu|\mu_0, \tau_0)$, donde: $\tau = \frac{1}{\sigma^2}$ se conoce como la precisión del modelo.
- $x \sim Normal\left(x|\mu^{*\,1}, \tau\right) \rightarrow \tau \sim Gamma\,(\tau|\alpha_0, \beta_0)$
- $x \sim Normal\,(x|\mu, \tau) \rightarrow (\mu, \tau) \sim Normal - Gamma\,(\mu, \tau|\mu_0, \tau_0, \alpha_0, \beta_0)$

[1] * Parámetros que se consideran conocidos

Sin embargo, en muchas ocasiones existen problemas que involucran el cálculo de integrales analíticamente complicadas.

Ejemplo 1:

$$x \sim Gamma(x|\alpha, 2); \qquad \alpha \sim Gamma(\alpha|\alpha_0, \beta_0)$$

Dada una muestra $\boldsymbol{x} = (x_1, \ldots, x_n)$, hacer inferencias sobre α así como para una nueva observación (x_F).

$$\begin{aligned}
p(\alpha|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\alpha)\, p(\alpha) \\
&\propto \left(\prod_{i=1}^{n} \frac{2^{\alpha}}{\Gamma(\alpha)} x_i^{(\alpha-1)} e^{-2x_i} \right) \alpha^{(\alpha_0-1)} e^{-\beta_0 \alpha} \\
&\propto \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i \right)^{(\alpha-1)} e^{-2\sum_{i=1}^{n} x_i} \alpha^{(\alpha_0-1)} e^{-\beta_0 \alpha} \\
&\propto \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i \right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0 \alpha}
\end{aligned}$$

$$\int_0^{\infty} \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i \right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0 \alpha} d\alpha = ?????$$

Por otro lado, para una observación futura:

$$
\begin{aligned}
p\left(x_F|\boldsymbol{x}\right) &= \int_0^\infty p\left(x_F|\alpha\right) p\left(\alpha|\boldsymbol{x}\right) d\alpha \\
&\propto \int_0^\infty \frac{2^\alpha}{\Gamma\left(\alpha\right)} x_F^{\alpha-1} e^{-2x_F} \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^n x_i\right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0\alpha} d\alpha
\end{aligned}
$$

$$
\int_0^\infty \frac{2^\alpha}{\Gamma\left(\alpha\right)} x_F^{\alpha-1} \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^n x_i\right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0\alpha} d\alpha = ?????
$$

Ejemplo 2: Suponga que se tienen x_1 y x_2 observaciones independientes del modelo Cauchy:

$$p\left(x|\theta\right) \propto \frac{1}{1+(x-\theta)^2} \qquad (x \in \mathbb{R}; \theta \in \mathbb{R})$$

Se desea hacer inferencias sobre el valor de θ suponiendo una inicial no informativa de la forma:

$$p\left(\theta\right) \propto 1$$

La regla de Bayes nos dice entonces que:

$$p\left(\theta|x_1, x_2\right) \propto \left(\frac{1}{1+(x_1-\theta)^2}\right)\left(\frac{1}{1+(x_2-\theta)^2}\right)$$

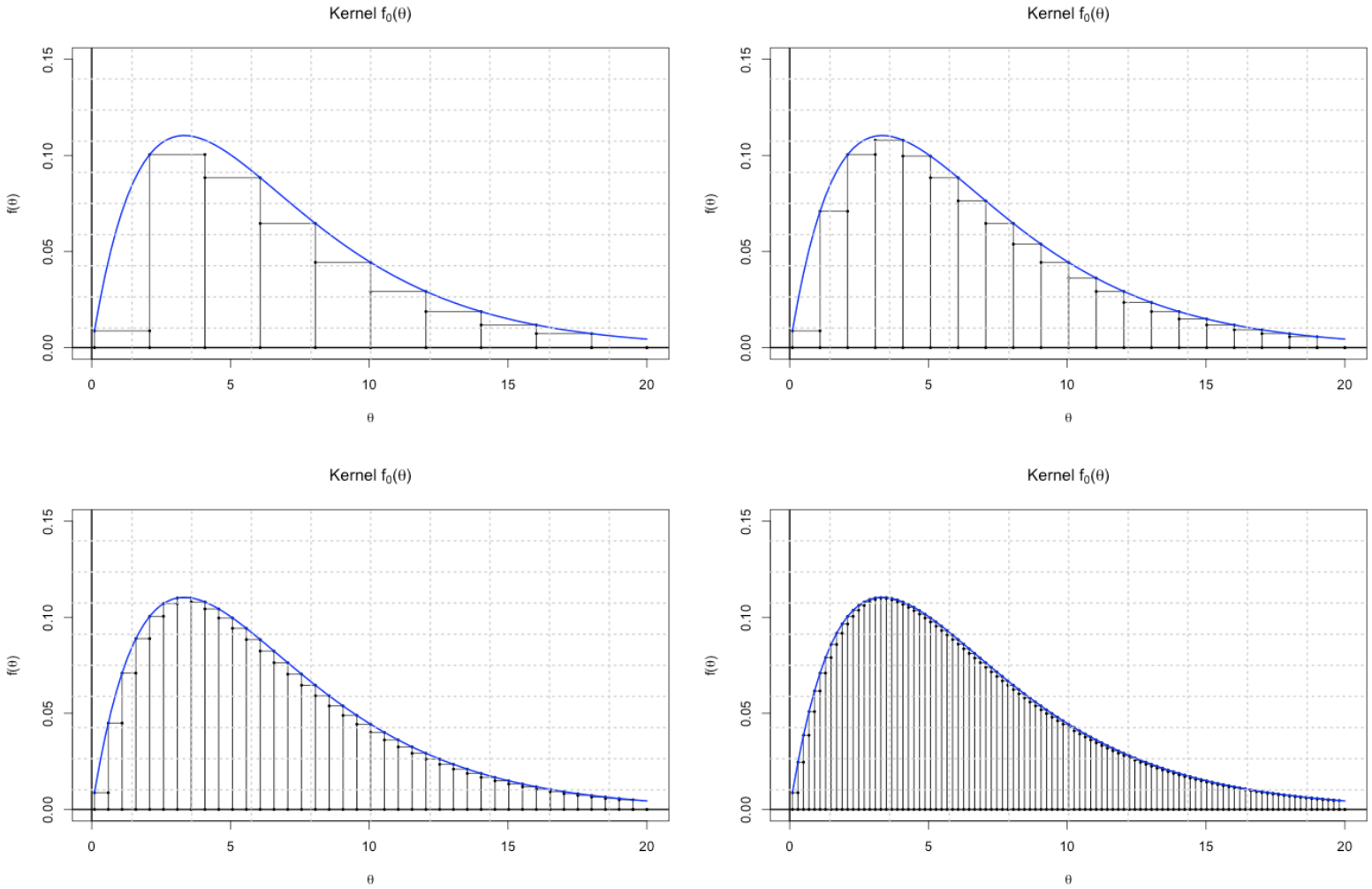
El kernel de la distribución final toma una forma complicada. Es conveniente notar, sin embargo, que $p\left(x|\theta\right)$ corresponde a la densidad marginal de x respecto a la densidad conjunta:

$$p\left(x, \lambda|\theta\right) = N\left(x\left|\theta, \frac{1}{\lambda}\right.\right) Ga\left(\lambda\left|\frac{1}{2}, \frac{1}{2}\right.\right)$$

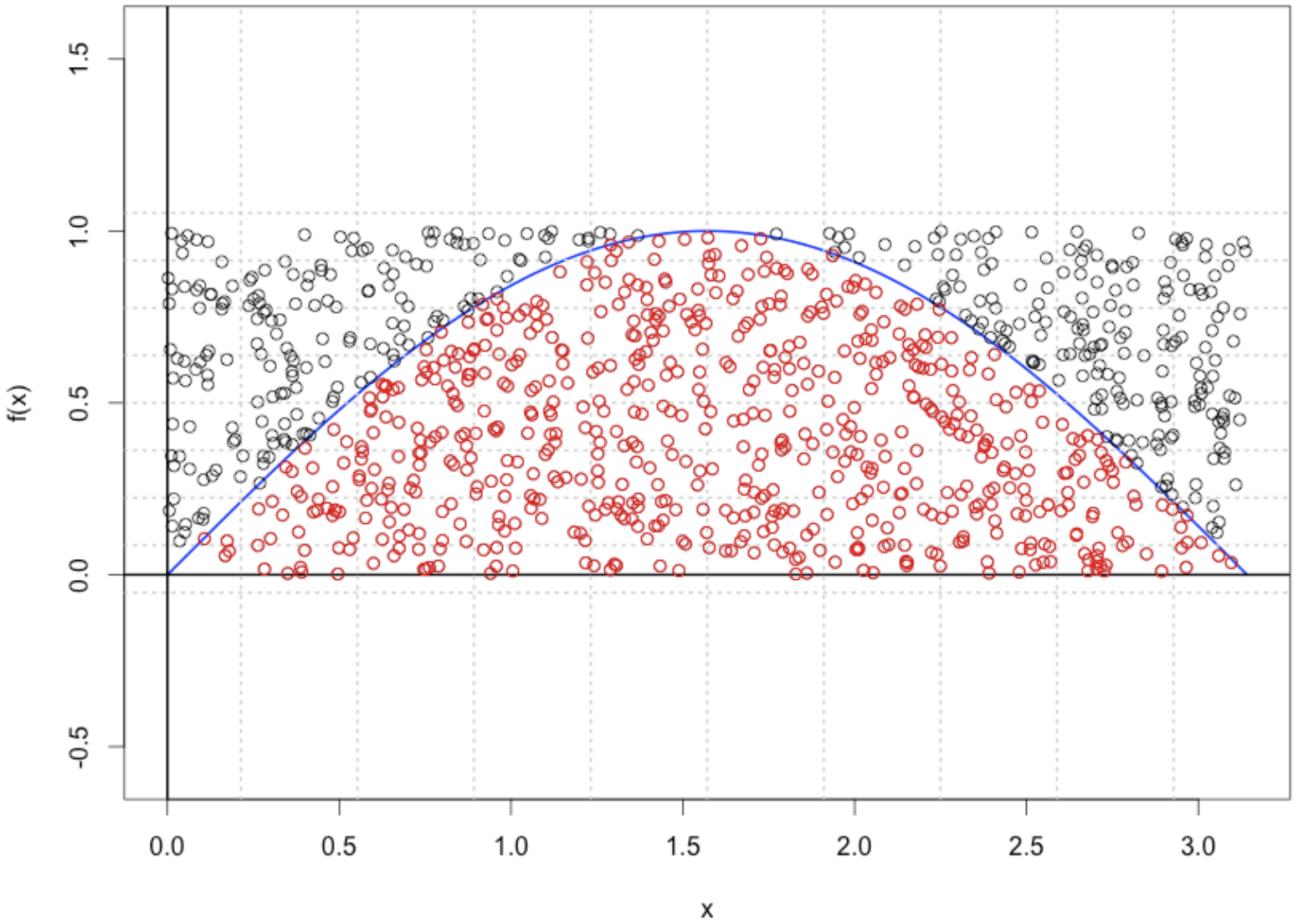
Nota: A veces incluir variables que inicialmente parece complicar el problema, en realidad lo hace más fácil de resolver!!!

Soluciones Numéricas:

1. Cuadratura: Regla del Punto Medio, Regla Trapezoidal
2. Cuadratura de Gauss-Hermite

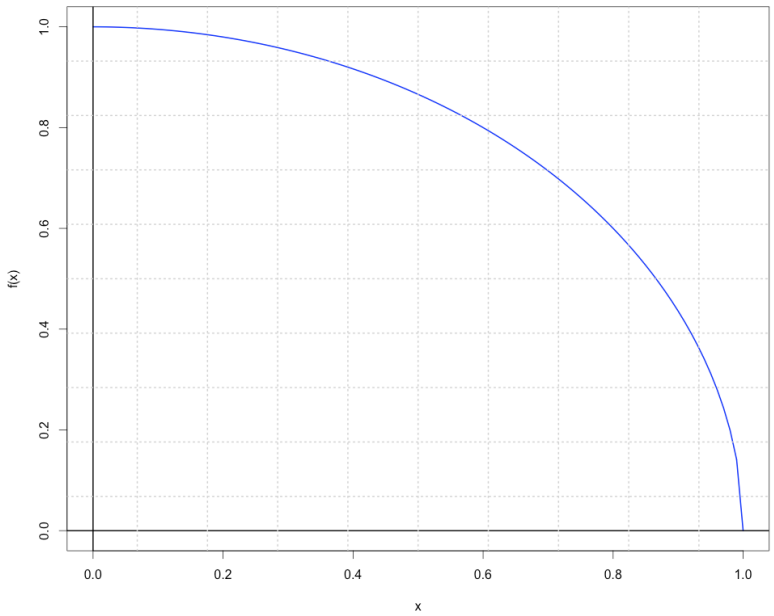


Aproximación vía simulación



Ideas básicas de la aproximación

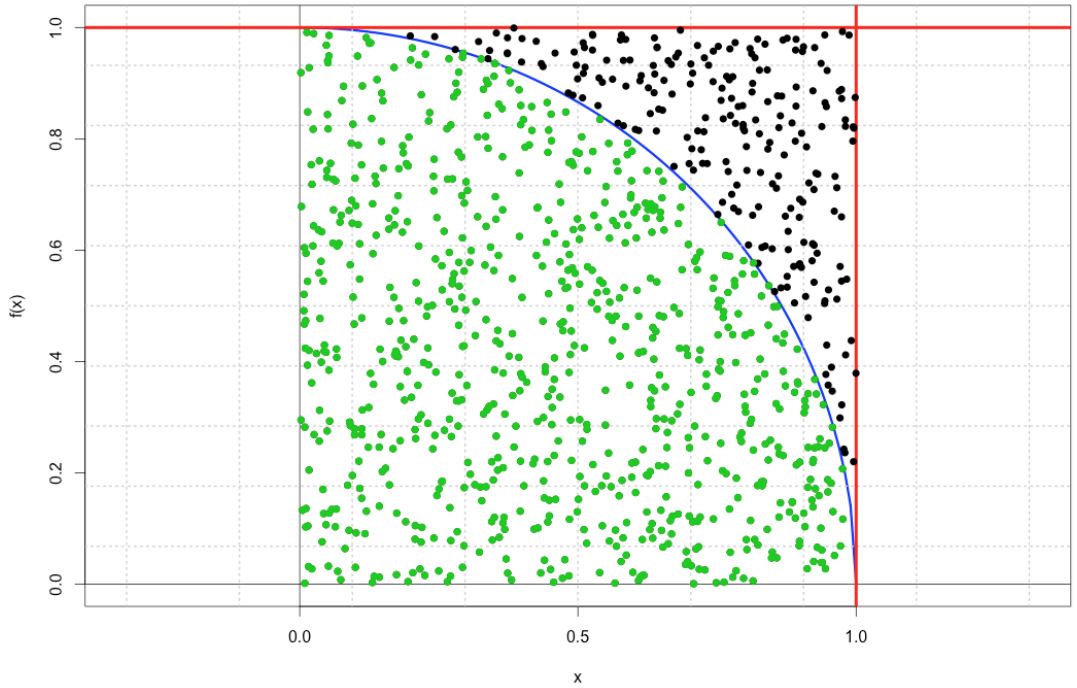
Problema 1: Se desea encontrar el área bajo la curva $f(x) = \sqrt{1 - x^2}$ en el intervalo $(0, 1)$



Esta integral se puede resolver vía cambio de variable y el resultado es:

$$I = \int_0^1 \sqrt{1 - x^2} dx = \frac{\pi}{4} \approx 0.7853982$$

La idea es construir un rectángulo de área A que cubra completamente a la función, luego simular N observaciones de forma uniforme y contar la proporción de puntos que caen debajo de la curva.



Definiendo N_f como el número de observaciones dentro de la curva, se tendría que el área estimada bajo la curva es:

$$\hat{I} = A\frac{N_f}{N}$$

Intuitivamente, entre más simulaciones hagamos mejor será nuestra aproximación.

N	$\hat{I}$
1,000	0.8050000
10,000	0.7857000
100,000	0.7848500
1,000,000	0.7856720
10,000,000	0.7851957

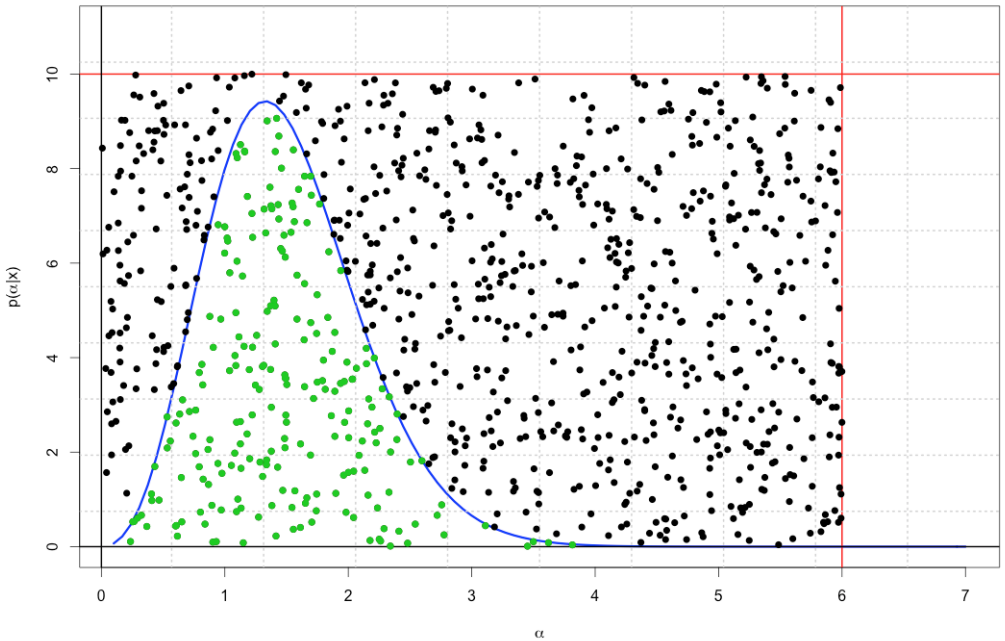
$$I = \int_0^1 \sqrt{1-x^2}dx = \frac{\pi}{4} \approx 0.7853982$$

Ver código: Programa01.r

Problema 2:

$$\int_0^{\infty} \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i \right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0 \alpha} d\alpha = ?????$$

Asumiendo $n = 3$ y que $\boldsymbol{x} = (0.7615236, 0.6414124, 0.3593526)$, con hiperparámetros $(\alpha_0 = 0.001, \beta_0 = 0.001)$



En este caso la aproximación es:

$$\hat{I} = (10)(6)\frac{N_f}{N}$$

Las aproximaciones para distintos valores de N son:

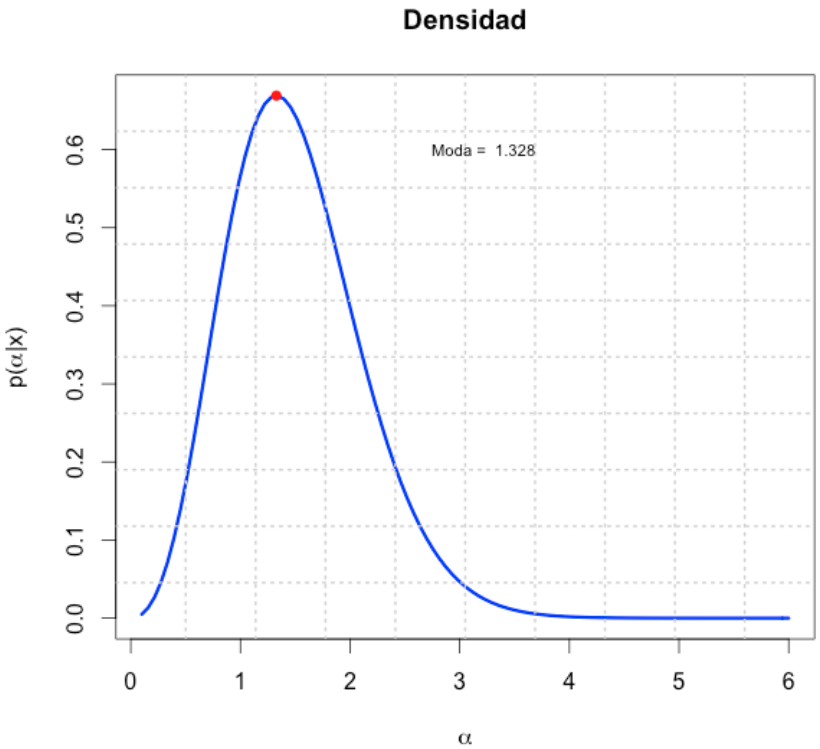
N	$\hat{I}$
1,000	13.38000
10,000	13.56600
100,000	14.15460
1,000,000	14.09142
10,000,000	14.08628
100,000,000	14.08861

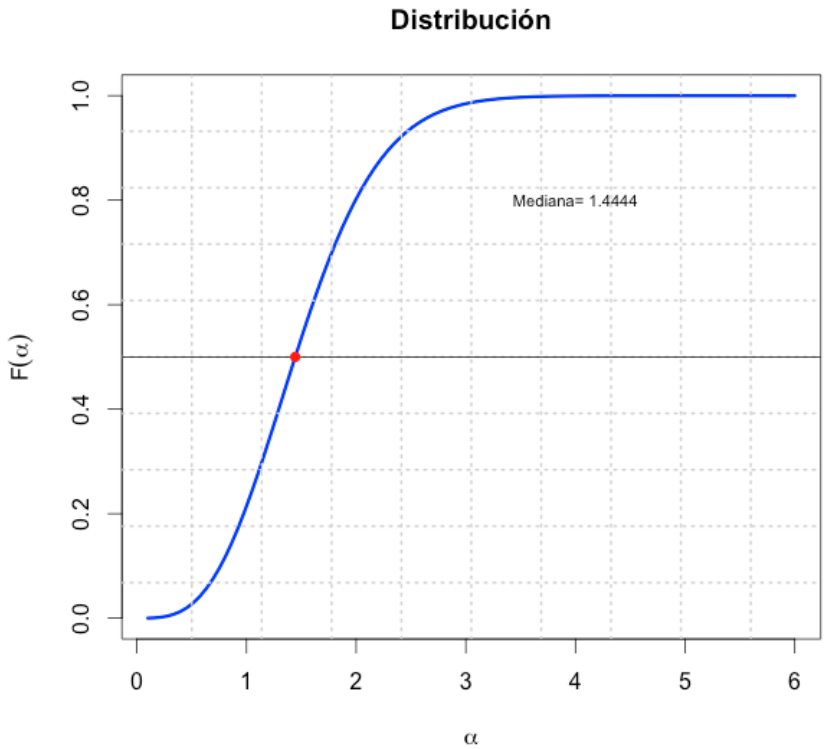
Por lo tanto la aproximación a la constante de proporcionalidad es $1/14.08861 \approx 0.07097932$, de donde concluimos entonces que:

$$p\left(\alpha|\underline{x}\right) \approx 0.07097932 \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i\right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0\alpha}$$

Ver código: Programa02.r

Calculada la constante de proporcionalidad, podemos encontrar numéricamente algunas cantidades de interés.





Función: maxLik → Moda → 1.327961
Función: uniroot → Mediana → 1.444442
Ver código: Programa02.r

Si queremos calcular la media, nuevamente necesitamos una integral!!!

$$\mathbb{E}(\alpha|\underline{x}) = \int_0^{\infty} 0.07097932\alpha \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i\right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0\alpha} d\alpha$$

Aplicando nuevamente las simulaciones correspondientes:

N	$\hat{I}$
1,000	1.452000
10,000	1.492800
100,000	1.502220
1,000,000	1.500240
10,000,000	1.500289
100,000,000	1.500306

Concluimos entonces que:

$$\mathbb{E}(\alpha|\underline{x}) \approx 1.500306$$

Ver código: Programa02.r

Muestreo por importancia

Esta idea se basa en la Ley Fuerte de los Grandes Números, que nos garantiza que si se tiene $x_1, \ldots, x_n$ m.a. de un modelo $s(x)$ (puede ser multivariado) y si se asume que

$$\mathbb{E}(g(x)) = \int_{\mathbb{R}^d} g(x)s(x)\,dx < \infty$$

donde $g : \mathbb{R}^d \to \mathbb{R}$, entonces

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \to \mathbb{E}(g(x)) = \int_{-\infty}^{\infty} g(x)s(x)\,dx$$

Ahora suponga que se está interesado en encontrar la integral:

$$I = \int_{\mathbb{R}^d} f(x)dx$$

Podemos rescribir dicha integral como:

$$I = \int_{\mathbb{R}^d} \left(\frac{f(x)}{s(x)}\right) s(x)dx$$

Si podemos simular $x_1, \ldots, x_n$ de la densidad $s(.)$, entonces:

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{f(x_i)}{s(x_i)}\right) \to \mathbb{E}\left(\frac{f(x_i)}{s(x_i)}\right) = I$$

A $s(.)$ se le conoce como la distribución de **muestreo por importancia** y debe tener las siguientes caracterísiticas:

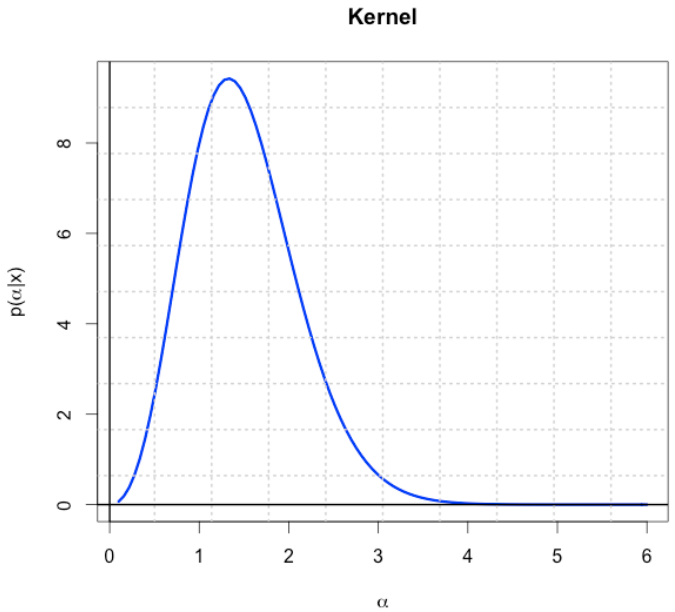
- Debe de ser fácil de simular (Normal, Gamma, Beta, Uniforme,...)
- Debe de tener una forma similiar a la de $f(x)$ (la función que se desea integrar)
- Debe de tener el mismo soporte que $f(x)$
- En la práctica es común trabajar en términos de alguna reparametrización de manera que la integral esté definida en todo $\mathbb{R}^d$ y luego utilizar como distribución de muestreo a una Normal o t de Student multivariada.

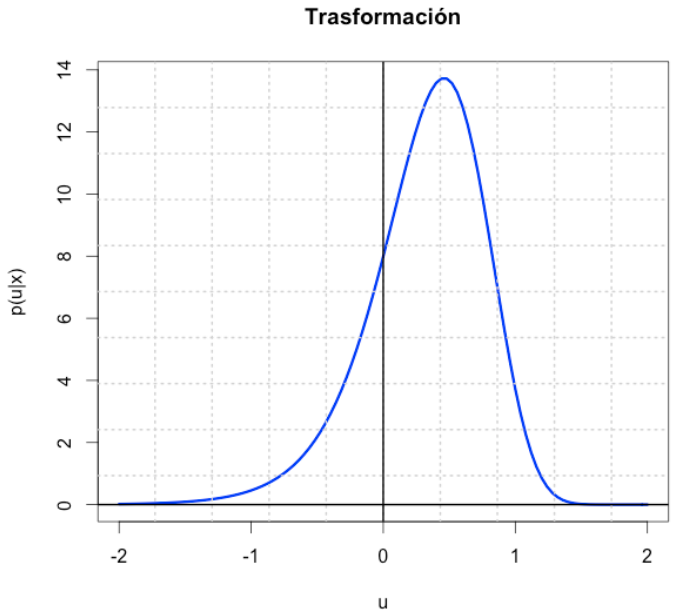
Ejemplo 1: Consideremos nuevamente el problema relacionado con la constante de proporcionalidad:

$$\int_0^\infty \frac{2^{n\alpha}}{\Gamma(\alpha)^n}\left(\prod_{i=1}^n x_i\right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0\alpha} d\alpha = ????$$

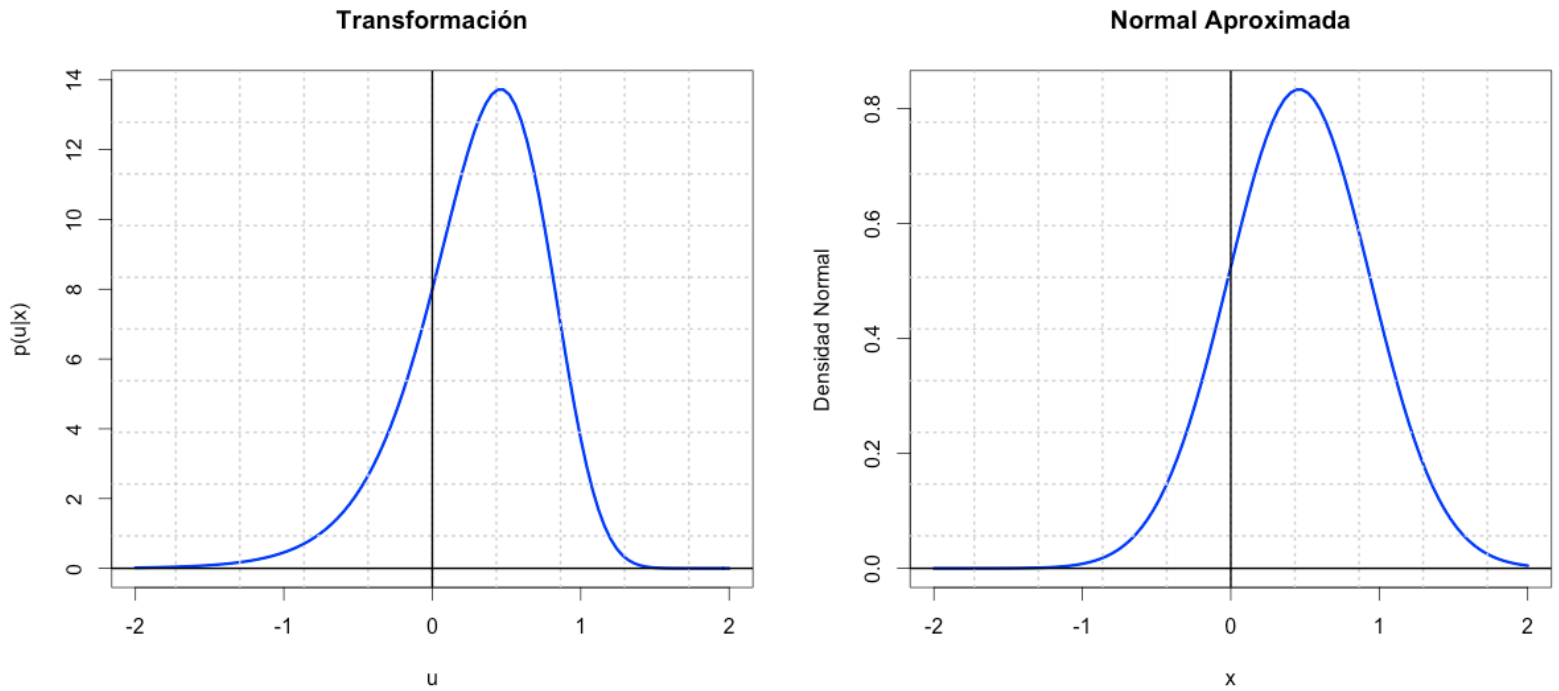
- Reparametrizamos el integrando para que esté definido en todo $\mathbb{R}$, mediante el cambio $u = \log(\alpha)$

$$\int_0^\infty f(\alpha)\, d\alpha = \int_{-\infty}^\infty f(e^u) e^u du$$





- Buscamos la distribución normal que mejor aproxime la "forma" de la función transformada.



Los parámetros de la normal que se obtuvieron son:

$$\mu = 0.4605435; \quad \sigma^2 = 0.228919$$

Se procede a simular observaciones de esta densidad.
Ver código: Programa03.r

Los resultados obtenidos son:

N	$\hat{I}$
1,000	14.23363
10,000	14.13107
100,000	14.08652
1,000,000	14.07526
10,000,000	14.07714
100,000,000	14.08597

En este ejercicio, la constante de proporcionalidad toma el valor de

$$1/14.08326 \approx 0.07099263$$

$$0.07097932 \qquad vs \qquad 0.07099263$$

Con esta constante se pueden realizar nuevamente las estimaciones de α.
Ver código: Programa03.r

Los resultados obtenidos son:

Estadística	Simulación basada en Uniformes	Muestreo por importancia
Constante	0.070979	0.070993
Moda	1.327961	1.327961
Mediana	1.444442	1.444299
Media	1.500306	1.500663

Ver código: Programa03.r

Simulación de una distribución

Los métodos anteriores permiten calcular resúmenes inferenciales a partir del cálculo de integrales complicadas, por ejemplo:

- La constante de normalización
- Valores esperados

Los métodos que a continuación se presentan se basan en **simular muestras de la distribución final**, las cuales permiten, en principio, aproximar cualquier característica de interés, como son cuantiles, medias, varianza, etc.

Existe una dualidad interesante aquí: por un lado, dada la distribución podemos simular observaciones de ella; por otro lado, dada una muestra grande es posible recrear la distribución que la generó (**Teorema de Glivenko-Cantelli**).

$$\lim_{n\to\infty} \sup\left(\left|F_n(x) - F(x)\right|\right) = 0$$

La idea entonces es: Supongamos que tenemos $f_0(\theta)$ el kernel de una densidad, de tal forma que:

$$f(\theta) = \frac{f_0(\theta)}{\int f_0(\theta')\, d\theta'}$$

El problema es generar un algoritmo que pueda simular observaciones de la densidad $f(\theta)$ a través de simulaciones de otra densidad $s(\cdot)$. En este curso veremos dos casos:

- Muestreo por importancia
- Simulación vía cadenas de Markov (Uso de JAGS)

Muestreo por importancia

Este método asume que podemos encontrar una constante M tal que $\frac{f_0(\theta)}{s(\theta)} \leq M$ para todo θ y que $s(.)$ es una función fácil de simular. El algoritmo que se propone es el siguiente:

Algoritmo

1. Generar una observación $\tilde{\theta}$ de $s(\theta)$
2. Generar una variable $u \sim U(0,1)$
3. si $u \leq \frac{f_0(\tilde{\theta})}{M\, s(\tilde{\theta})}$, aceptar $\tilde{\theta}$ como una observación de la densidad $f(\theta)$; en caso contrario, repetir los pasos 1 a 3.

Observación: Suponiendo que se desea una muestra de tamaño N de $f(\theta)$, el valor esperado para el tamaño de la correspondiente muestra de $s(\theta)$ es:

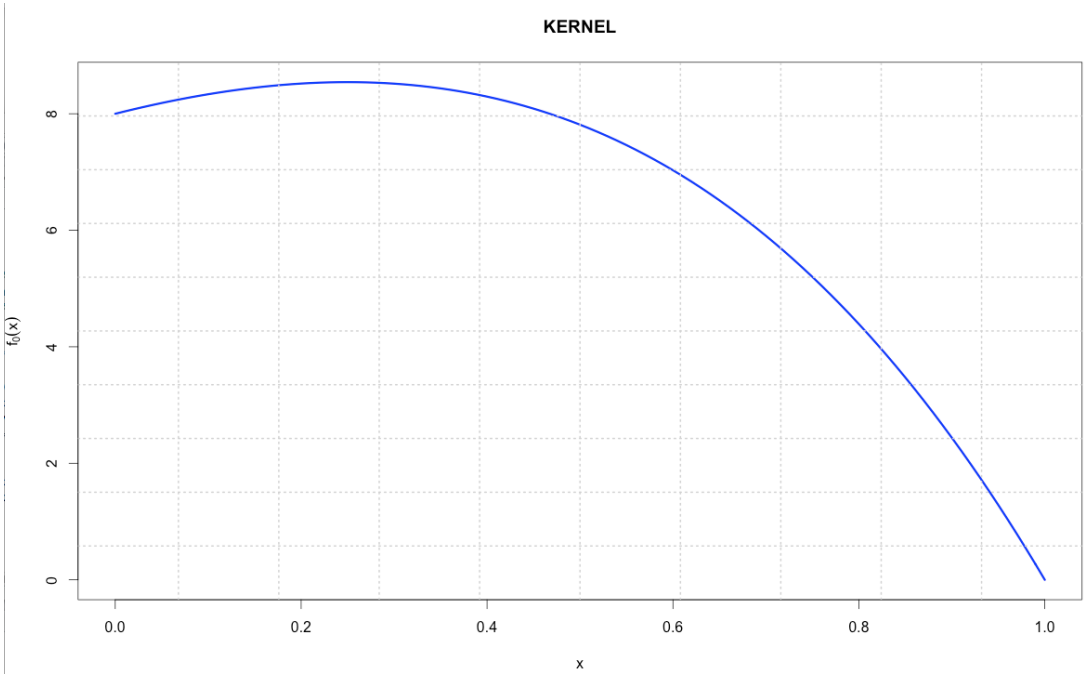
$$N_0 = \frac{M\, N}{\int f_0(\theta)\, d\theta}$$

Ejemplo 3 : Suponga que se tiene el siguiente kernel:

$$f_0(x) = (x+2)^3(1-x) \quad x \in (0,1)$$

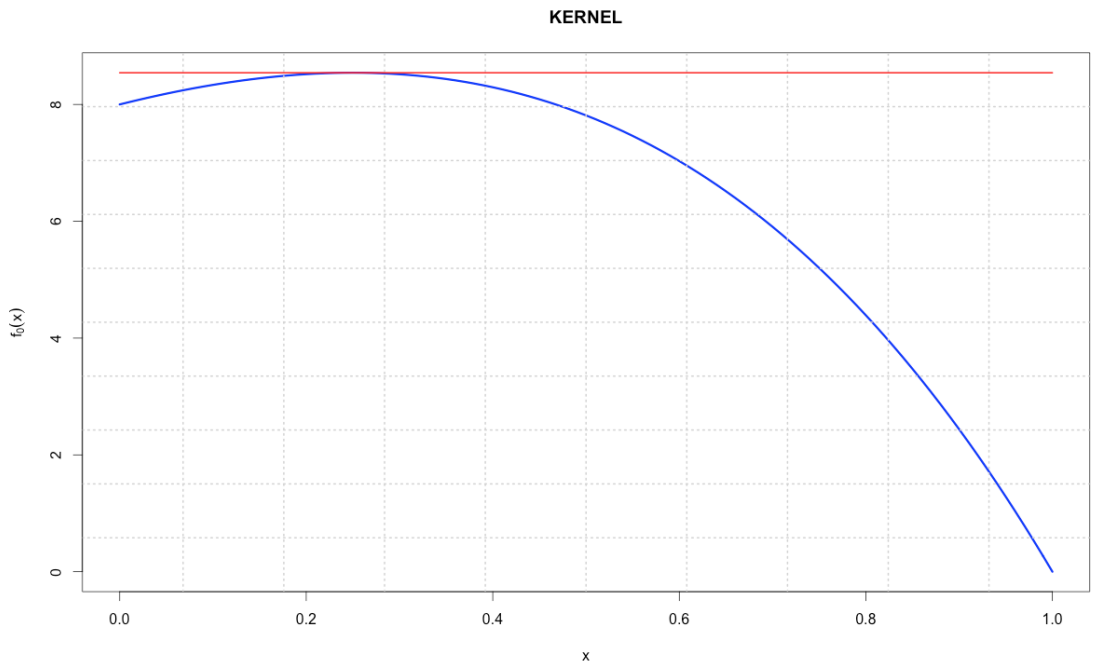
Se desea simular observaciones de dicho kernel para aproximar la media $\mathbb{E}(X)$ y la mediana $(q_{0.5})$.
Se puede probar que, en este caso, la constante de proporcionalidad así como las cantidades de interés son:

$$k = \frac{20}{131} \approx 0.1527; \qquad \mathbb{E}(x) = \frac{160}{393} \approx 0.4071; \qquad q_{0.5} = 0.3900$$



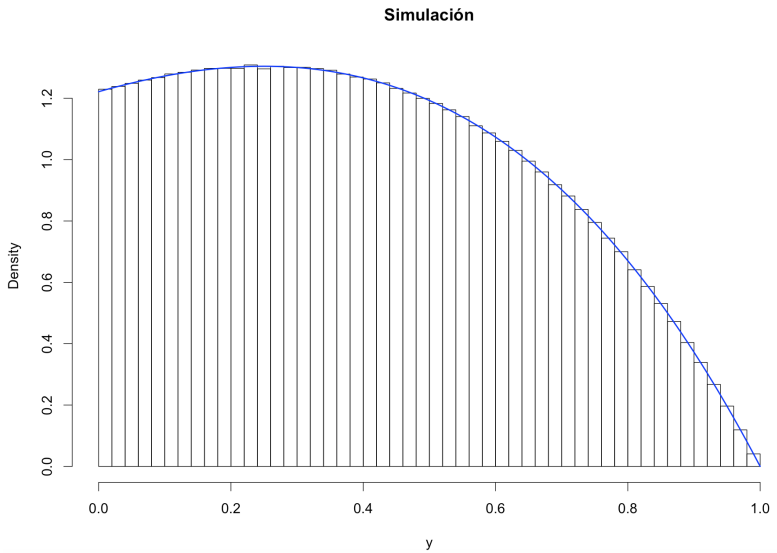
Proponemos $s \sim U(x|0,1)$ y $M = \sup\{f_0(x)\}$ El algoritmo es:

1. Generar una observación $\tilde{x}$ de $s(x) = 1$, es decir $\tilde{x} \sim U(0,1)$
2. Generar una variable $u \sim U(0,1)$
3. si $u \leq \frac{f_0(\tilde{x})}{M\, s(\tilde{x})}$, aceptar $\tilde{x}$ como una observación de la densidad $f\,(x)$; en caso contrario, repetir los pasos 1 a 3. En este caso $s\,(\tilde{x}) = 1$ y $M = 8.542969$



Corriendo el algoritmo anterior para $N_0 = 100,000$, se obtuvo que:

$$N = 76,771; \qquad \frac{N}{N_0} = 0.76776; \qquad k \approx \frac{N_0}{MN} = 0.152463$$

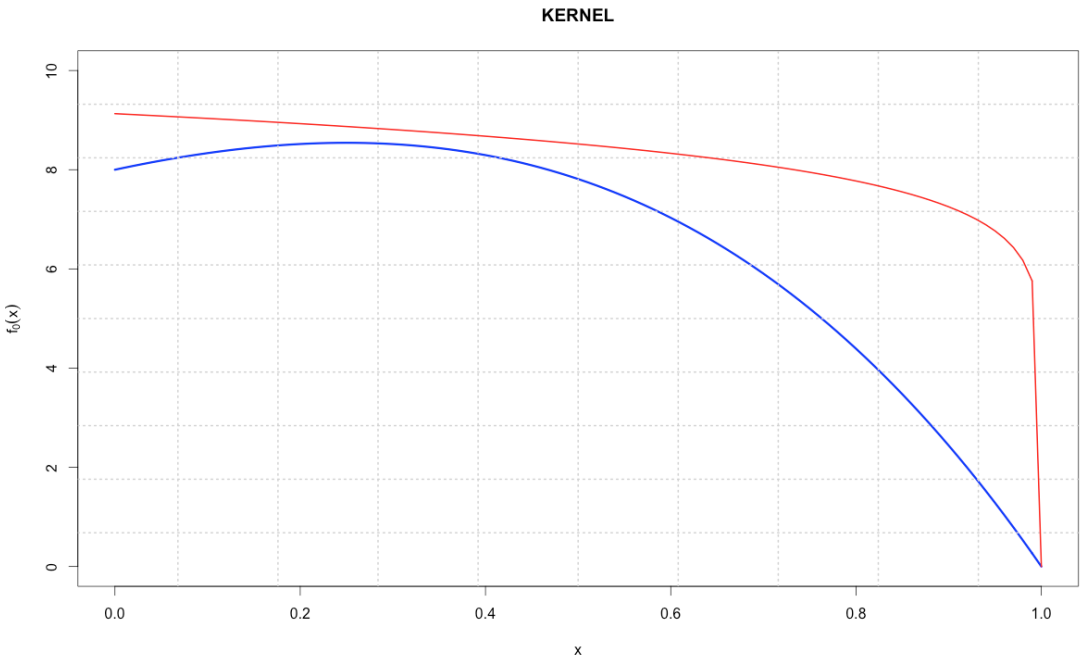


En este caso vía simulación las carecteristicas distribucionales son aproximadas como:

$$\mathbb{E}(X) \approx \frac{1}{N}\sum_{i=1}^{N} y_i = 0.4071616 \qquad q_{0.5} \approx \text{mediana}(y_1, \ldots, y_N) = 0.3897104$$

Ver código: Programa04.r

Es deseable mejorar la tasa de aceptación. Para ello se puede proponer una densidad $s(.)$ más parecida al kernel $f_0(x)$. En el ejemplo anterior, supongamos entonces que $s(x) = Beta(x|1, 1.1)$ y que se propone $M = 8.1$



El algoritmo es:

1. Generar una observación $\tilde{x}$ de $s(x) = Beta(x|1, 1.1)$
2. Generar una variable $u \sim U(0, 1)$
3. si $u \leq \frac{f_0(\tilde{x})}{M\, s(\tilde{x})}$, aceptar $\tilde{x}$ como una observación de la densidad $f\,(x)$; en caso contrario, repetir los pasos 1 a 3. Corriendo el algoritmo anterior para $N_0 = 100,000$ se obtuvo que:

$$N = 80653; \qquad \frac{N}{N_0} = 0.80653; \qquad k \approx \frac{N_0}{M\,N} = 0.1530715$$

Mientras que las caracteristicas distribucionales son aproximadas como:

$$\mathbb{E}(X) \approx \frac{1}{N}\sum_{i=1}^{N} y_i = 0.4066467 \qquad q_{0.5} \approx \text{mediana}(y_1, \ldots, y_N) = 0.3884548$$

Ver código: Programa05.r

En resumen tenemos que las simulaciones arrojan lo siguiente:

Estadística	Exacto	Uniforme	Beta
Constante	$\frac{20}{131} \approx 0.1526718$	0.1524630	0.1530715
Mediana	0.3900254	0.3897104	0.3884548
Media	$\frac{160}{393} \approx 0.4071247$	0.4071616	0.4066467
Tasa de Aceptación		0.76776	0.80653

Ver código: Programa05.r

Ejercicio: Simular observaciones de la distribución $N(0,1)$utilizando $s(x) = Cauchy(x|0,1)$
Ver código: Programa06.r

*

Ejemplo Bayesiano:

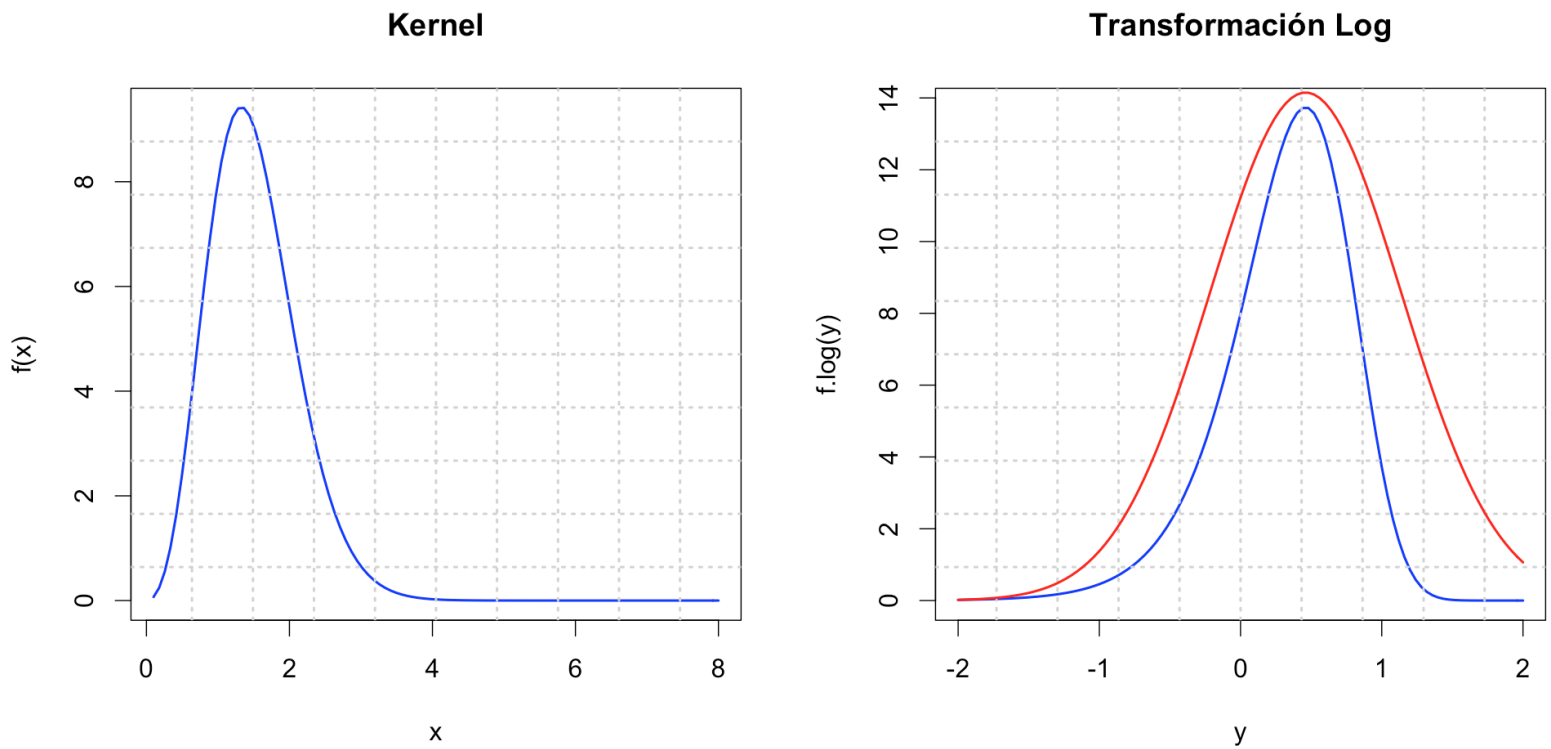
$$x \sim Gamma(x|\alpha, 2); \qquad \alpha \sim Gamma(\alpha|\alpha_0, \beta_0)$$

Dada $\boldsymbol{x} = (x_1, \ldots, x_n)$ una muestra, hacer inferencia sobre α

$$p\left(\alpha|\underline{x}\right) \quad \propto \quad \frac{2^{n\alpha}}{\Gamma(\alpha)^n}\left(\prod_{i=1}^{n} x_i\right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0\alpha}$$

Asumiendo $n = 3$ y que $\boldsymbol{x} = (0.7615236, 0.6414124, 0.3593526)$, con hiperparámetros $(\alpha_0 = 0.001, \beta_0 = 0.001)$, simularemos observaciones de la distribución $p\left(\alpha|\boldsymbol{x}\right)$ para posteriormente hacer la inferencia correspondiente

El proceso consiste en transformar el kernel para tomar valores en todos los reales ($Y = \log(\alpha)$) y proponer a la densidad normal que ajuste lo mejor posible a dicha transformación



Se procede entonces a simular observaciones de la v.a. Y.
Se propone $s(x) = Normal(x|0.4605435, 0.457838)$, $M = 24$
Ver código: Programa07.r

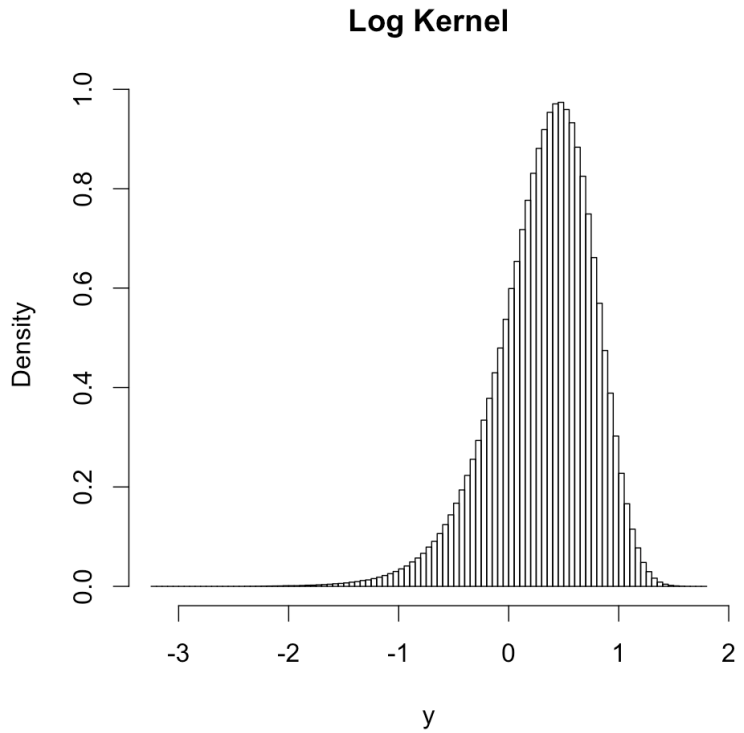
El algoritmo es:

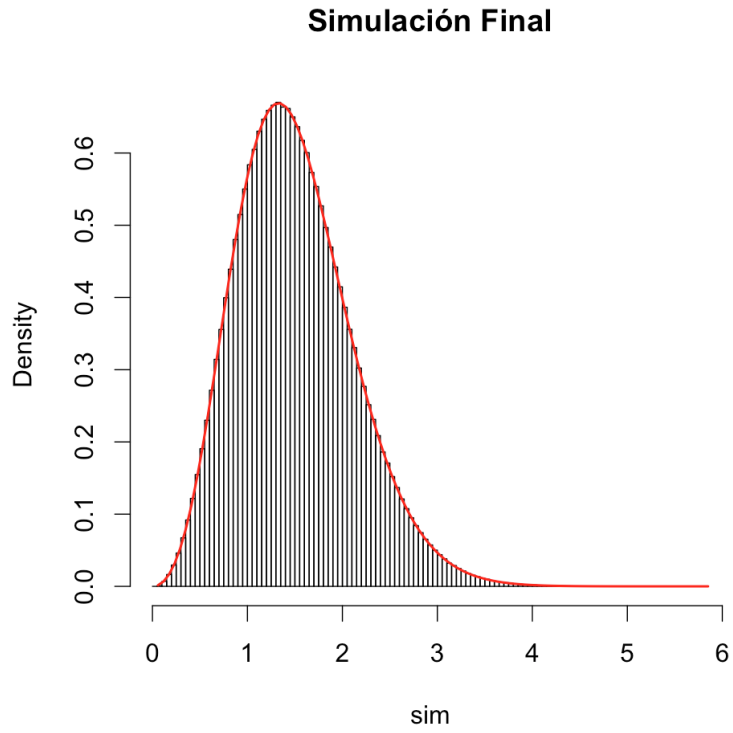
1. Generar una observación $\tilde{x}$ de $s(x) = Normal(x|0.4605435, 0.457838)$
2. Generar una variable $u \sim U(0,1)$
3. si $u \leq \frac{f_0(\tilde{x})}{M\, s(\tilde{x})}$, aceptar $\tilde{x}$ como una observación de la densidad $f\,(x)$; en caso contrario, repetir los pasos 1 a 3.

Corriendo el algoritmo anterior para $N_0 = 10,000,000$ se obtuvo que:

$$N_0 = 5,869,603; \qquad \frac{N}{N_0} = 0.5869603;$$

Lo anterior genera simulaciones del kernel transformado
Ver código: Programa07.r





La esperanza a posterior es:

$$\mathbb{E}(\alpha \mid \underline{x}) \approx 1.501118;$$

El intervalo de máxima densidad al 95 % es:

$$(0.3937097, 2.6891916)$$

Ver código: Programa07.r

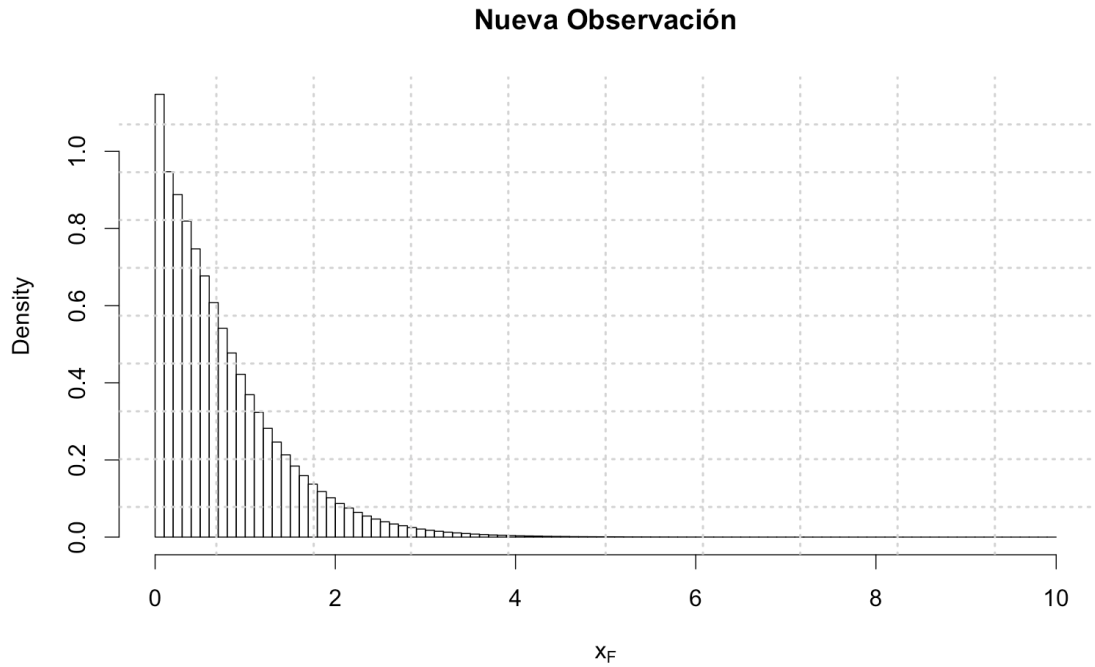
Una de las ventajas de tener simulaciones de la distribución final es que podemos simular fácilmente obervaciones de la distribución de x_F (una nueva observación). Para ello recordemos que:

$$\begin{aligned} p\left(x_F, \alpha|\boldsymbol{x}\right) &= \frac{p\left(x_F, \alpha, \boldsymbol{x}\right)}{p\left(\boldsymbol{x}\right)} \\ &= \frac{p\left(x_F|\alpha, \boldsymbol{x}\right) p\left(\alpha, \boldsymbol{x}\right)}{p\left(\boldsymbol{x}\right)} \\ &= p\left(x_F|\alpha\right) p\left(\alpha|\boldsymbol{x}\right) \end{aligned}$$

Es decir, simular observaciones del vector (x_F, α) dada la muestra $\underline{x}$ es simple mediante el siguiente algoritmo (muestreo condicional):

1. Generar una observación $\tilde{\alpha}$ de $p\left(\alpha|\boldsymbol{x}\right)$
2. Generar una observación $\tilde{x}$ de la densidad $Gamma(x|\tilde{\alpha}, 2)$
3. Repetir paso 1 y 2. Entonces $(\tilde{x}, \tilde{\alpha})$ son simulaciones del vector aleatorio (x_F, α) dada la muestra $\boldsymbol{x}$
   Ver código: Programa07.r

Simulación para una nueva observación del modelo:



En este caso:

$$\mathbb{E}(x_F \mid \boldsymbol{x}) = 0.75045$$

y un intervalo de credibilidad al 95 % es:

$$(0, 2.104152)$$

Ver código: Programa07.r

Ejercicio: Sea $(0, 0, 0, 1, 0)$ una muestra observada del modelo $Bernoulli(x|\theta)$, asumiendo que la inicial es $p(\theta) = -\log(\theta)$, simular observaciones de la densidad final $p(\theta|\underline{x})$ y encontrar un intervalo de credibilidad al 95 % para θ.

Ver código: Programa08.r

*

Monte Carlo vía cadenas de Markov

Esta técnica permite generar, de manera iterativa, observaciones de distribuciones multivariadas que díficilmente podrían simularse utilizando métodos directos. La idea es simple: **Construir una cadena de Markov que sea fácil de simular y cuya distribución de equilibrio corresponda a la distribución que nos interesa**.

Teorema

Sea $\theta^{(1)}, \theta^{(2)}, \ldots,$ *una cadena de Markov homogénea, irreducible y aperiódica, con espacio de estados* Θ *y distribución de equilibrio* $p(\theta|x)$. *Entonces, conforme* $t \to \infty$ *se tiene que:*

- $\theta^{(t)} \to \theta \sim p(\theta|x)$

Es decir, si dejamos correr por mucho tiempo a la cadena, eventualmente estaremos simulando observaciones de la distribución $p(\theta|x)$

Consideremos un ejemplo para el caso discreto. Se tiene sólo dos posibles estados $\{0, 1\}$, con las siguientes probabilidades de transición:

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{pmatrix} \end{array}$$

Se demuestra que una forma de obtener la distribución estacionaria es resolver el sistema

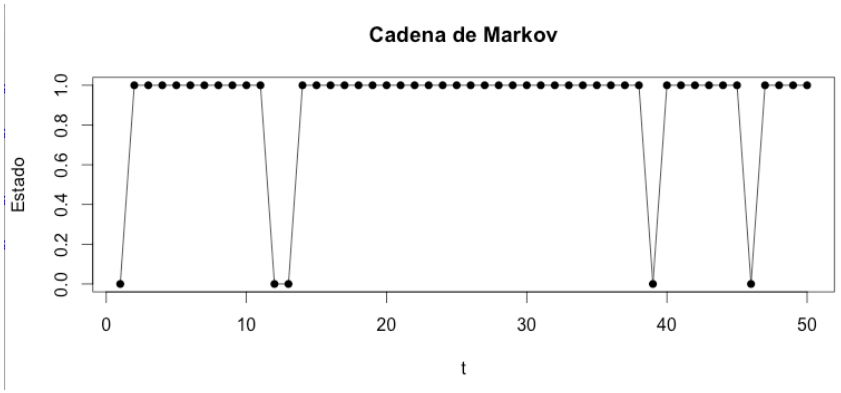
$$\boldsymbol{\pi} = \boldsymbol{\pi} P,$$

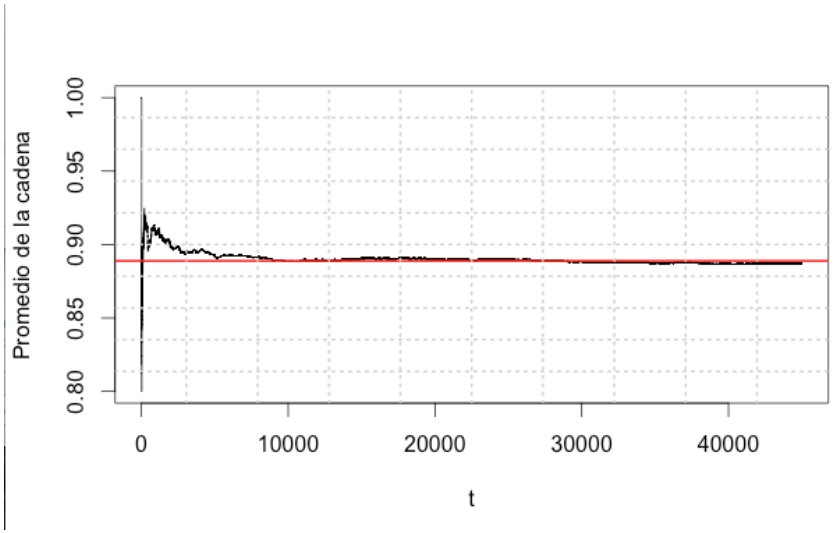
de donde se encuentra que la distribución de equilibrio es $\boldsymbol{\pi} = \left(\frac{0.1}{0.9}, \frac{0.8}{0.9}\right)$.
Teóricamente sabemos entonces que no importa dónde inicie la cadena, después de varias iteraciones la cadena estará simulando observaciones de la distribución estacionaria dada por:

$$\mathbb{P}(X = 0) = \frac{1}{9} \qquad \mathbb{P}(X = 1) = \frac{8}{9}$$

Es decir, en este caso, la distribución estacionaria es una $Bernoulli(\frac{8}{9})$

Simulación de la cadena:





Ver código: Programa09.r

Ejercicio: Suponga que se tiene $\{X_n : n \in \mathbb{N}\}$, un proceso estocástico con espacio de estados $\{0, 1, 2\}$ y con matriz de transición dada por:

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/3 & 2/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}$$

Realizar un programa en R que simule la cadena y verificar que la cadena converge a la distribución estacionaria dada por:

$$\pi_0 = 1/4 : \quad \pi_1 = 3/8; \quad \pi_2 = 3/8$$

Ver código: Programa10.r

Problema: Dada una densidad $p(\theta|x)$, ¿Cómo construir una cadena que tenga como distribución de equilibrio precisamente a $p(\theta|x)$?

*

Algoritmo de Metropolis-Hasting (1953)
El algoritmo construye una cadena de Markov **apropiada** definiendo probabilidades de transición tal que se cumplan las condiciones para tener una cadena estacionaria con distribución límite $p(\theta|x)$.
El algoritmo requiere una distribución de transición (en principio arbitraria) $Q(\theta^*|\theta)$ que sea fácil de simular. Se define:

$$\alpha\left(\theta^*, \theta\right) = \text{mín}\left\{\frac{p(\theta^*|x)Q\left(\theta|\theta^*\right)}{p\left(\theta|x\right)Q\left(\theta^*|\theta\right)}, 1\right\}$$

1. Inicializar la cadena $\theta^{(0)}$ (en un principo de forma arbitraria)
2. Para la t-ésima iteración. Simular θ^* de $Q\left(\theta^* \middle| \theta^{(t-1)}\right)$;
3. Genera una variable $u \sim U(0,1)$
4. Si $u \leq \alpha\left(\theta^*, \theta^{(t-1)}\right)$, hacer $\theta^{(t)} = \theta^*$; en caso contrario, hacer $\theta^{(t)} = \theta^{(t-1)}$.

Obs: $\alpha\left(\theta^*, \theta^{(t-1)}\right)$, en este contexto, es la probabilidad de que la cadena se mueva de $\theta^{(t-1)}$ a θ^* y para su cálculo **no es necesario tener la constante de proporcionalidad** de la densidad $p\left(\theta|x\right)$.

Técnicamente, la distribución de transición $Q(\theta^*|\theta)$ puede ser arbitraria. Sin embargo, una forma práctica es suponer independencia y asumir que:

$$Q(\theta^*|\theta) = Q_0(\theta^*)$$

Donde para evitar que la cadena se quede estancada, se sugiere que Q_0 tengo una forma similar al kernel de $p(\theta|x)$ En este caso:

$$\alpha(\theta^*, \theta) = \text{mín}\left\{\frac{p(\theta^*|x)Q_0(\theta)}{p(\theta|x)\,Q_0(\theta^*)}, 1\right\}$$

En la páctica es común utilizar, después de una reparametrización apropiada, distribuciones de transición normales o t de Student sobredispersas.
Por ejemplo:

$$Q_0(\theta^*) = N_d\left(\theta^* \left| \hat{\theta}, kV(\hat{\theta})\right.\right)$$

donde $\hat{\theta}$ y $V(\hat{\theta})$ denotan la media y la matriz de varianzas y covarianzas de la aproximación normal asintótica para $p(\theta|x)$ y k es un factor de dispersión para lograr explorar mejor el soporte de la densidad que se desea simular.

Por construcción, después de un determinado número de iteraciones, la cadena debe empezar a estabilizarse y comenzar a simular observaciones de la distribución estacionaria, en este caso de la final $p(\theta|x)$. Una cuestión interesante es el momento en el cual la cadena se ha estabilizado. El objetivo es simular N observaciones de $p(\theta|x)$; para ello dos posibles opciones son:

- Fijar T suficiente grande, luego inicializar N cadenas $\theta_1^{(0)}, \ldots, \theta_N^{(0)}$ y correrlas durante T pasos. Finalmente, considerar los valores $\theta_1^{(T)}, \theta_2^{(T)}, \ldots, \theta_N^{(T)}$ como una muestra de f_Y. (Computacionalmente es demandante)
- Correr una sola cadena, luego fijar T suficientemente grande y tomar $\theta^{(T+K)}, \theta^{(T+2K)}, \ldots, \theta^{(T+NK)}$ como m.a. de $p(\theta|x)$, donde K se elige de manera que la correlación entre las observaciones sea pequeña.

Nota: No es fácil determinar en qué momento la cadena converge, por lo que comúnmente se hacen pruebas empíricas, por ejemplo sobre los promedios ergódicos.

Ejemplo 1 : (Consideremos nuevamente el caso Gamma)

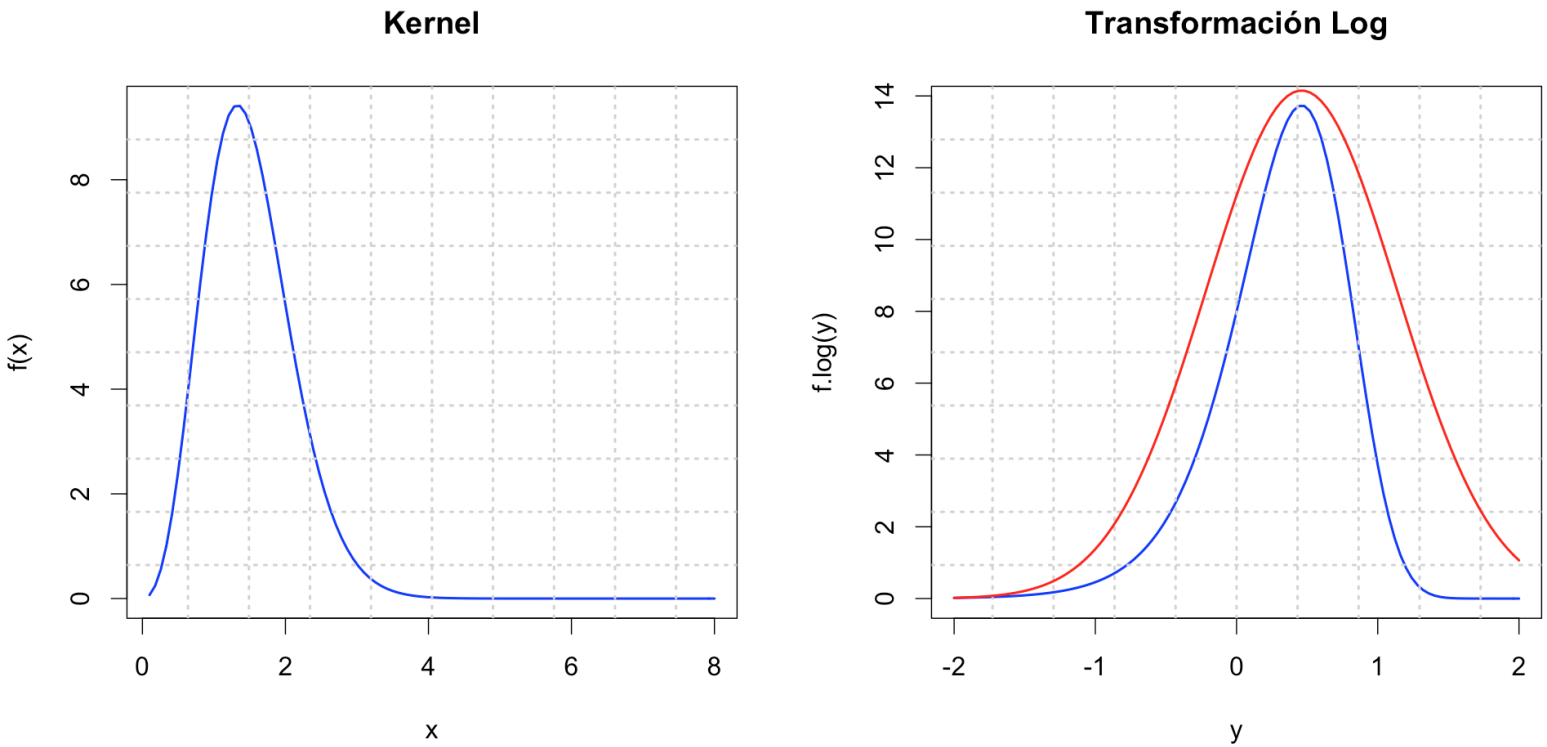
$$x \sim Gamma(x|\alpha, 2); \qquad \alpha \sim Gamma(\alpha|\alpha_0, \beta_0)$$

Dada $\boldsymbol{x} = (x_1, \ldots, x_n)$ una muestra, hacer inferencia sobre α

$$p\left(\alpha|\underline{x}\right) \quad \propto \quad \frac{2^{n\alpha}}{\Gamma(\alpha)^n}\left(\prod_{i=1}^{n} x_i\right)^{(\alpha-1)} \alpha^{(\alpha_0-1)} e^{-\beta_0\alpha}$$

Asumiendo $n = 3$ y que $\boldsymbol{x} = (0.7615236, 0.6414124, 0.3593526)$, con hiperparámetros $(\alpha_0 = 0.001, \beta_0 = 0.001)$, simularemos observaciones de la distribución $p\left(\alpha|\underline{x}\right)$ utilizando el Algoritmo de Metropolis-Hasting para posteriormente hacer la inferencia correspondiente.
Ver código: Programa11.r

El proceso consiste en transformar el kernel para tomar valores en todos los reales ($Y = \log(\alpha)$) y proponer a la densidad normal que ajuste lo mejor posible a dicha transformación



Se procede entonces a simular observaciones de la v.a. Y.
Se propone $Q_0(\theta) = Normal(\theta|0.4605435, 0.457838)$
Ver código: Programa11.r

El algoritmo de Metropolis-Hasting es:

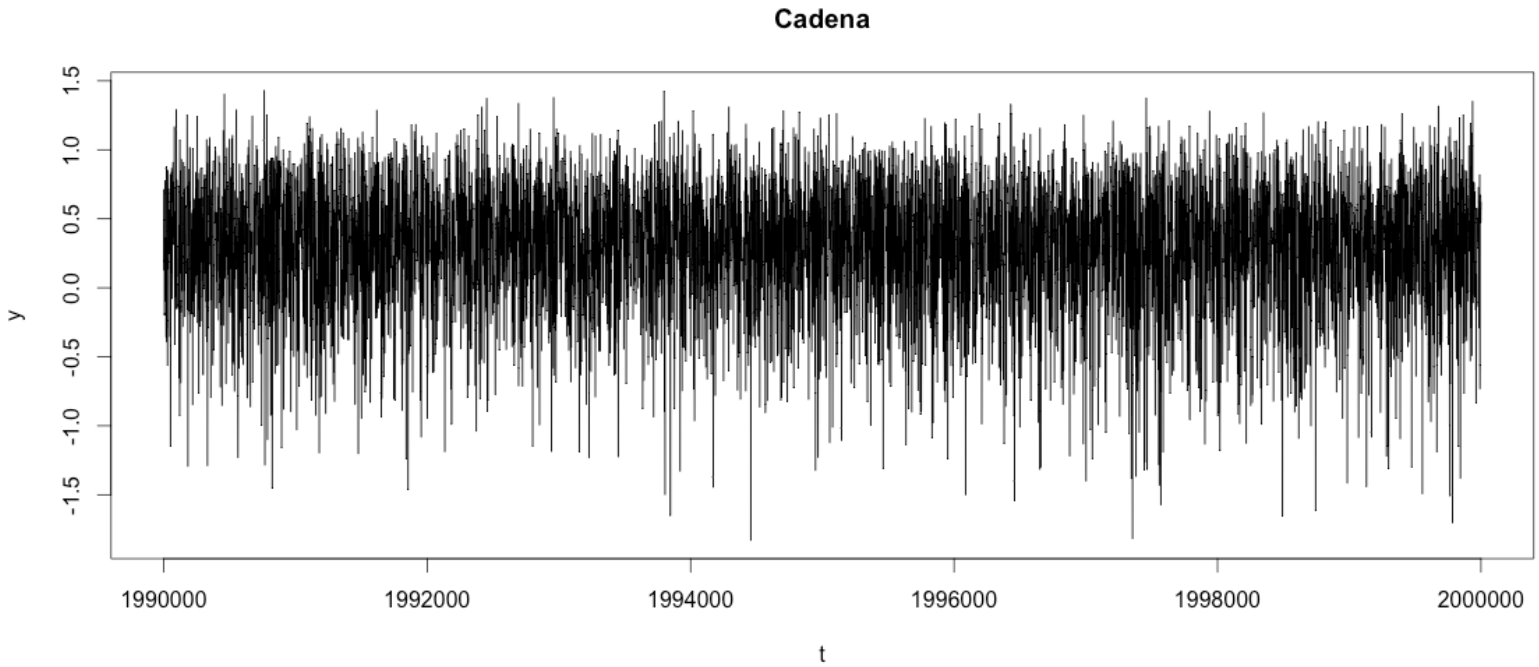
1. Inicializar la cadena $y^{(0)}$ (en un principo de forma arbitraria, pero cuidando evitar errores numéricos)
2. Para la t-ésima iteración. Simular y^* de $N(\theta|0.4605435, 0.457838)$;
3. Genera una variable $u \sim U(0,1)$
4. Si $u \leq \alpha\left(y^*, y^{(t-1)}\right)$, hacer $y^{(t)} = y^*$; en caso contrario, hacer $y^{(t)} = y^{(t-1)}$.

En este caso:

$$\alpha\left(y^*, y^{(t-1)}\right) = \text{mín}\left\{\frac{f_Y(y^*)Normal(y^{(t-1)}|0.4605435, 0.457838)}{f_Y(y^{(t-1)})Normal(y^*|0.4605435, 0.457838)}, 1\right\}$$

Ver código: Programa11.r

Se corrió el algoritmo durante 2,000,000 iteraciones. Calentamiento: 100,000.



Ver código: Programa11.r

Ejemplo 3: Supongamos ahora 2 parámetros desconocidos:

$$x \sim Gamma\left(x|\alpha,\beta\right); \qquad p\left(\alpha,\beta\right) = Gamma(\alpha|\alpha_0,\beta_0)Gamma(\beta|\alpha_1,\beta_1)$$

Asumiendo que:

$$\alpha_0 = \beta_0 = \alpha_1 = \beta_1 = 0.001$$

y que se observa la muestra de tamaño 5:

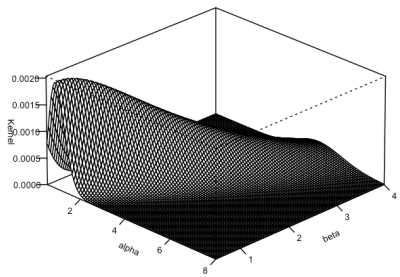
$$\boldsymbol{x} = (0.4154325, 1.7853782, 1.7315852, 1.0254059, 1.9427045)$$

Hacer inferencia sobre α, β y x_F

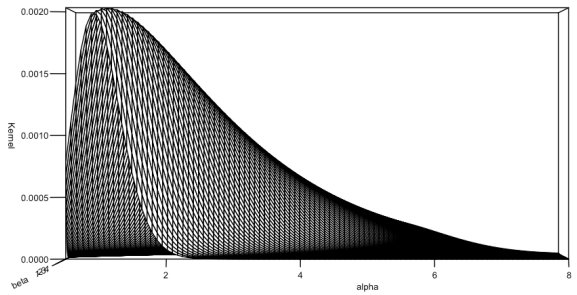
Solución clásica: Mediante métodos numéricos

$$\hat{\alpha}_{MV} = 3.882556; \qquad \hat{\beta}_{MV} = 2.813241; \qquad x_F \sim ?$$

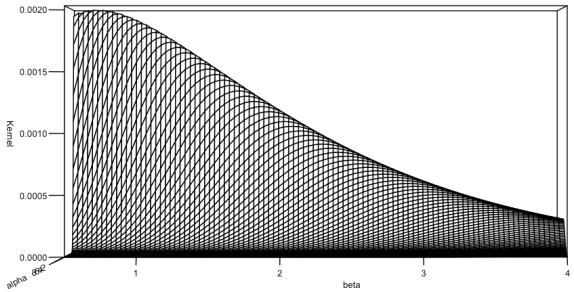
$f_0(\alpha,\beta|x)$

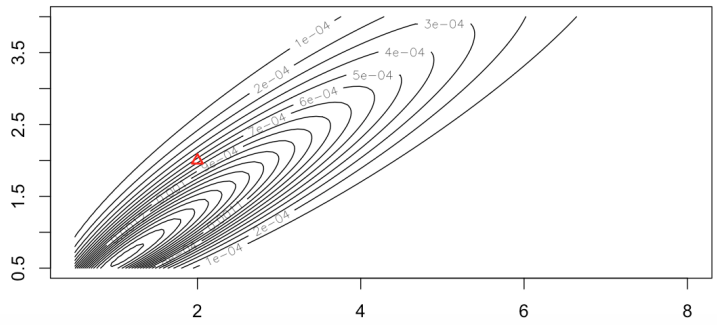


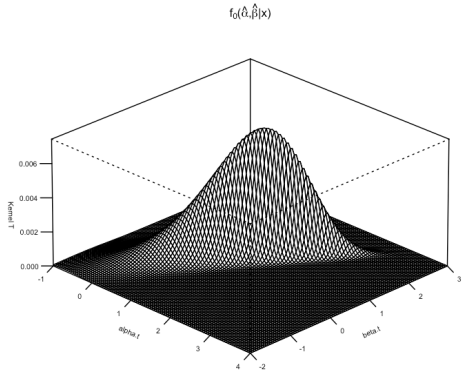
$f_0(\alpha,\beta|x)$

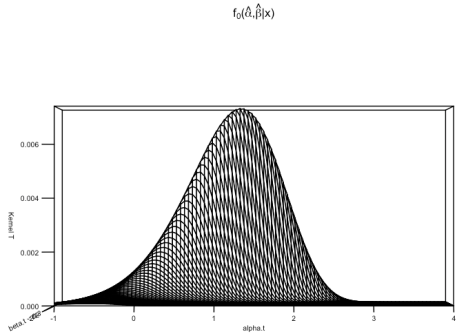


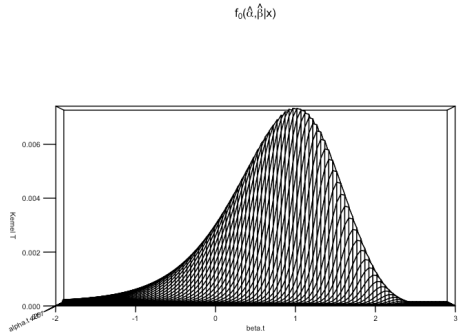
$f_0(\alpha,\beta|x)$

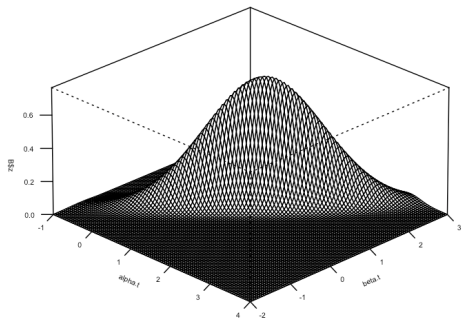


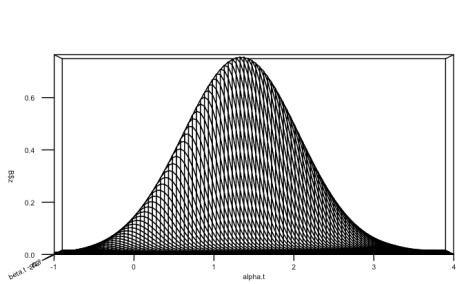


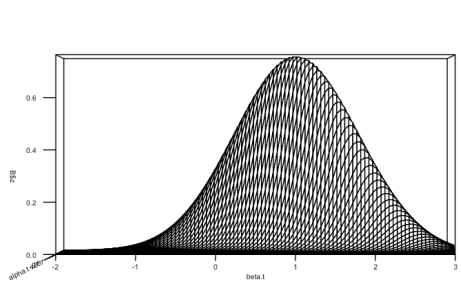

$f_0(\hat{\alpha},\hat{\beta}|x)$
Kernel
alpha.t
beta.t

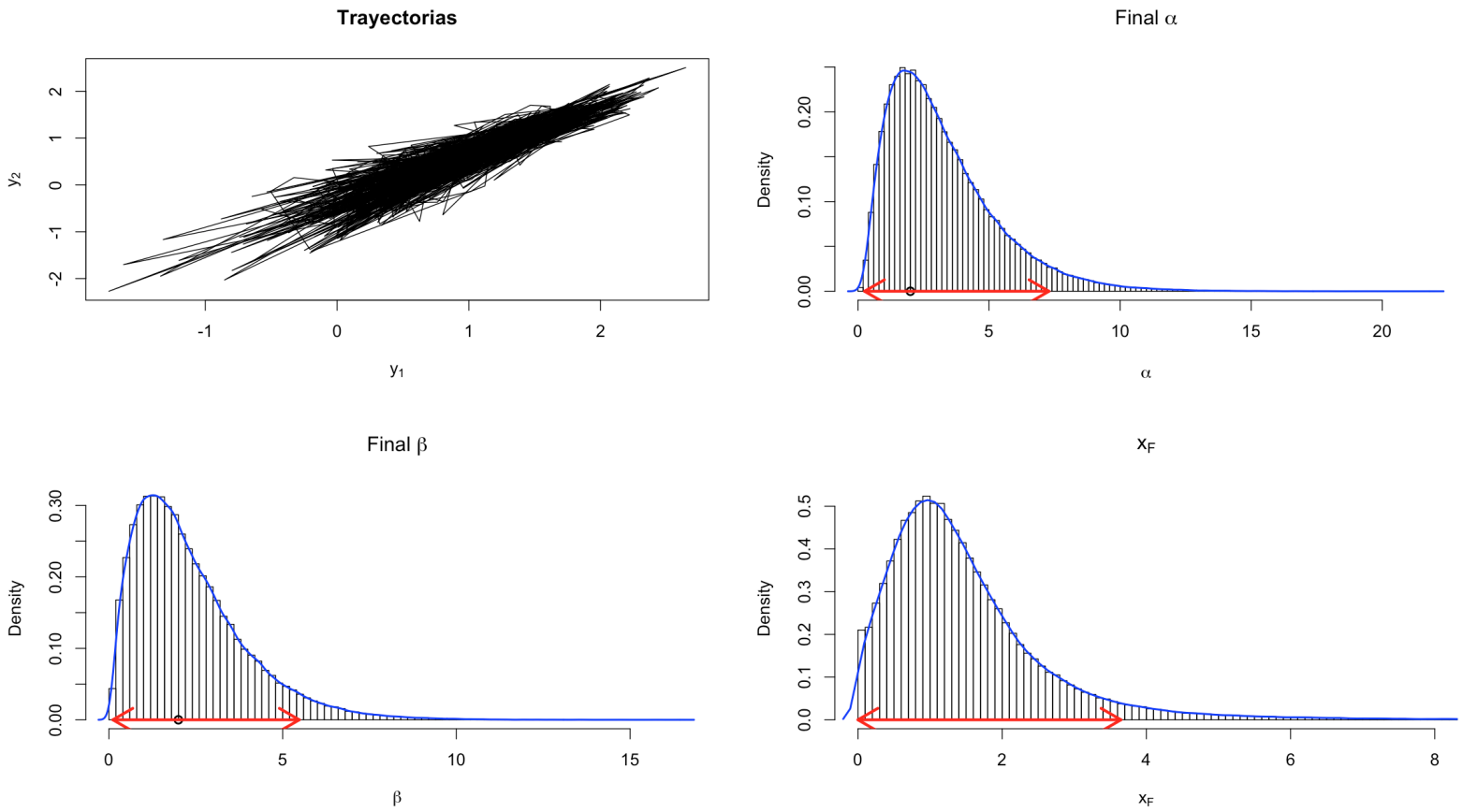

$f_0(\hat{\alpha},\hat{\beta}|x)$
Kernel
alpha.t


$f_0(\hat{\alpha},\hat{\beta}|x)$
Kernel
beta.t


Aprox N(μ,Σ)
BSE
alpha.t
beta.t


Aprox N(μ,Σ)
BSE
alpha.t


Aprox N(μ,Σ)
BSE
beta.t



Ver video: CADENA MARKOV.mp4
Ver código: Programa12.r

Algoritmo de Gibbs

En el Algoritmo MH, el valor $\alpha\left(\theta^*, \theta\right)$ se interpreta como la probabilidad de que la cadena se mueva de θ a θ^*. Lo ideal es construir un algoritmo en el cual $\alpha\left(\theta^*, \theta\right) = 1$, es decir, que el movimiento esté siempre garantizado para favorecer posteriormente a la convergencia de la cadena. El Algoritmo de Gibbs no es más que una caso especial del Algoritmo MH que precisamente garantiza lo mencionado anteriormente. La principal característica de este algoritmo es:

- Se requiere poder simular de las condicionales completas

$$p\left(\theta_i | \theta_{[i]}, \boldsymbol{x}\right); \quad \theta_{[i]} = \left(\theta_1, \ldots, \theta_{[i-1]}, \theta_{[i+1]}, \ldots, \theta_k\right)$$

- Cada valor de la cadena se obtiene a través de un proceso iterativo que utiliza a las condicionales completas.

Algoritmo de Gibbs

El algoritmo propuesto es el siguiente:

0 Inicializar la cadena en un valor inicial $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_k^{(0)}\right)$
Para $t \in \{0, 1, 2, \ldots, \}$

1 Generar una muestra $\theta_1^{(t+1)}$ de $p\left(\theta_1^{(t+1)} \middle| \theta_2^{(t)}, \theta_3^{(t)} \ldots, \theta_k^{(t)}, \boldsymbol{x}\right)$

2 Generar una muestra $\theta_2^{(t+1)}$ de $p\left(\theta_2^{(t+1)} \middle| \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}, \boldsymbol{x}\right)$

3 Generar una muestra $\theta_3^{(t+1)}$ de $p\left(\theta_3^{(t+1)} \middle| \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_k^{(t)}, \boldsymbol{x}\right)$

$\vdots$

k Generar una muestra $\theta_k^{(t+1)}$ de $p\left(\theta_k^{(t+1)} \middle| \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)}, \boldsymbol{x}\right)$

k+1 Con los pasos 1 a k construir $\theta^{(t+1)} = \left(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_k^{(t+1)}\right)$

k+2 Repetir los pasos 1 a $k+1$, y generar la cadena $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots,$

Algoritmo de Gibbs

Ejemplo 2: Recordando el modelo Cauchy con inicial no informativa.

$$p\left(x|\theta\right) \propto \frac{1}{1+(x-\theta)^2}; \quad p\left(\theta\right) \propto 1; \quad (x \in \mathbb{R}; \theta \in \mathbb{R})$$

Suponiendo que se observan 2 muestras independientes $\boldsymbol{x} = (x_1, x_2)$, la densidad final es:

$$p\left(\theta|\boldsymbol{x}\right) \propto \left(\frac{1}{1+(x_1-\theta)^2}\right)\left(\frac{1}{1+(x_2-\theta)^2}\right)$$

Nos interesa simular observaciones de este kernel para aproximar algunas características de interés (media, mediana, moda, IC al 95 %.)

Algoritmo de Gibbs

Para solucionar este problema se introducen variables latentes para faciltar el problema de simulación.
Al inicio de la sesión se mencionó que $p\left(x_1|\theta\right)$ se obtiene al marginalizar la siguiente densidad conjunta

$$p\left(x_1, \lambda_1|\theta\right) = N\left(x_1 \left| \theta, \frac{1}{\lambda_1}\right.\right) Ga\left(\lambda_1 \left| \frac{1}{2}, \frac{1}{2}\right.\right)$$

Lo anterior se puede generalizar para dos observaciones del modelo. En este caso se prueba que $p\left(x_1, x_2|\theta\right)$ se obtiene al marginalizar la siguiente densidad conjunta:

$$p\left(x_1, x_2, \lambda_1, \lambda_2|\theta\right) = p\left(\boldsymbol{x}, \lambda_1, \lambda_2|\theta\right) = \prod_{i=1}^{2} N\left(x_i \left| \theta, \frac{1}{\lambda_i}\right.\right) Ga\left(\lambda_i \left| \frac{1}{2}, \frac{1}{2}\right.\right)$$

Algoritmo de Gibbs

El problema parece haberse **triplicado**, pues de tener sólo a θ como parámetro de interés, ahora han aparecido dos nuevas variables λ_1 y λ_2.
Sin embargo notemos que gracias a las variables introducidas, es posible simular observaciones del vector $(\theta, \lambda_1, \lambda_2)$ dada la muestra x_1, x_2 mediante el Algoritmo de Gibbs. Para ello necesitaremos las distribuciones condicionales completas:

- Para λ_1:

$$
\begin{aligned}
p(\lambda_1|\lambda_2,\theta,\boldsymbol{x}) &= \frac{p(\lambda_1,\lambda_2,\theta,\boldsymbol{x})}{p(\lambda_2,\theta,\boldsymbol{x})} = \frac{p(\boldsymbol{x},\lambda_1,\lambda_2|\theta)\,p(\theta)}{p(\lambda_2,\theta,\boldsymbol{x})} \\
&\propto p(\boldsymbol{x},\lambda_1,\lambda_2|\theta) = \prod_{i=1}^{2} N\left(x_i\left|\theta,\frac{1}{\lambda_i}\right.\right) Ga\left(\lambda_i\left|\frac{1}{2},\frac{1}{2}\right.\right) \\
&\propto N\left(x_1\left|\theta,\frac{1}{\lambda_1}\right.\right) Ga\left(\lambda_1\left|\frac{1}{2},\frac{1}{2}\right.\right) \\
&\propto \lambda_1^{\frac{1}{2}} e^{\frac{-\lambda_1}{2}(x_1-\theta)^2} \lambda_1^{\frac{1}{2}-1} e^{-\frac{1}{2}\lambda_1} \\
&\propto Ga\left(\lambda_1\left|1,\frac{1+(x_1-\theta)^2}{2}\right.\right)
\end{aligned}
$$

Lo anterior es conocido como el **Método de Variables Latentes**

Algoritmo de Gibbs

- Para λ_2, de forma similar se prueba que:

$$p\left(\lambda_2|\lambda_1,\theta,\boldsymbol{x}\right) = Ga\left(\lambda_2\left|1,\frac{1+(x_2-\theta)^2}{2}\right.\right)$$

- Para θ:

$$\begin{aligned}
p\left(\theta|\lambda_1,\lambda_2,\boldsymbol{x}\right) &= \frac{p\left(\theta,\lambda_1,\lambda_2,\boldsymbol{x}\right)}{p\left(\lambda_1,\lambda_2,\boldsymbol{x}\right)} = \frac{p\left(\boldsymbol{x},\lambda_1,\lambda_2|\theta\right)p\left(\theta\right)}{p\left(\lambda_1,\lambda_2,\boldsymbol{x}\right)} \\
&\propto p\left(\boldsymbol{x},\lambda_1,\lambda_2|\theta\right) = \prod_{i=1}^{2} N\left(x_i\left|\theta,\frac{1}{\lambda_i}\right.\right) Ga\left(\lambda_i\left|\frac{1}{2},\frac{1}{2}\right.\right) \\
&\propto N\left(x_1\left|\theta,\frac{1}{\lambda_1}\right.\right) N\left(x_2\left|\theta,\frac{1}{\lambda_2}\right.\right) \\
&\propto e^{-\frac{\lambda_1}{2}(x_1-\theta)^2} e^{-\frac{\lambda_2}{2}(x_2-\theta)^2} \\
&\propto N\left(\theta\left|\mu_0,\sigma_0^2\right.\right)
\end{aligned}$$

donde:

$$\mu_0 = \frac{\lambda_1 x_1 + \lambda_2 x_2}{\lambda_1+\lambda_2}; \qquad \sigma_0^2 = \frac{1}{\lambda_1+\lambda_2}$$

Algoritmo de Gibbs

En resumen, las condicionales completas son

$$
\begin{aligned}
p\left(\lambda_1|\lambda_2,\theta,\boldsymbol{x}\right) &= Ga\left(\lambda_1\left|1,\frac{1+(x_1-\theta)^2}{2}\right.\right)\\
p\left(\lambda_2|\lambda_2,\theta,\boldsymbol{x}\right) &= Ga\left(\lambda_2\left|1,\frac{1+(x_2-\theta)^2}{2}\right.\right)\\
p\left(\theta|\lambda_1,\lambda_2,\boldsymbol{x}\right) &= N\left(\theta\left|\frac{\lambda_1 x_1+\lambda_2 x_2}{\lambda_1+\lambda_2},\frac{1}{\lambda_1+\lambda_2}\right.\right)
\end{aligned}
$$

¡Todas las condicionales son fáciles de simular!!! Por lo que se puede aplicar Gibbs.

Algoritmo de Gibbs

El Algoritmo de Gibbs para simular de $p\left(\theta, \lambda_1, \lambda_2 | \boldsymbol{x}\right)$ es el siguiente:

0 Inicializar la cadena en un valor inicial $\boldsymbol{\theta}^{(0)} = \left(\theta^{(0)}, \lambda_1^{(0)}, \lambda_2^{(0)}\right)$
Para $t \in \{0, 1, 2, \ldots, \}$

1 Generar una muestra $\lambda_1^{(t+1)}$ de
$$p\left(\lambda_1^{(t)} \middle| \lambda_2^{(t)}, \theta^{(t)}, \boldsymbol{x}\right) = Ga\left(\lambda_1^{(t)} \middle| 1, \frac{1+\left(x_1-\theta^{(t)}\right)^2}{2}\right)$$

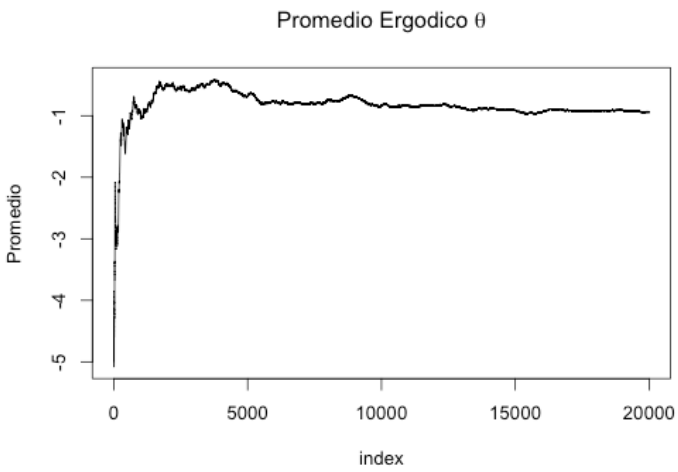
2 Generar una muestra $\lambda_2^{(t+1)}$ de
$$p\left(\lambda_2^{(t)} \middle| \lambda_1^{(t)}, \theta^{(t)}, \boldsymbol{x}\right) = Ga\left(\lambda_2^{(t)} \middle| 1, \frac{1+\left(x_2-\theta^{(t)}\right)^2}{2}\right)$$

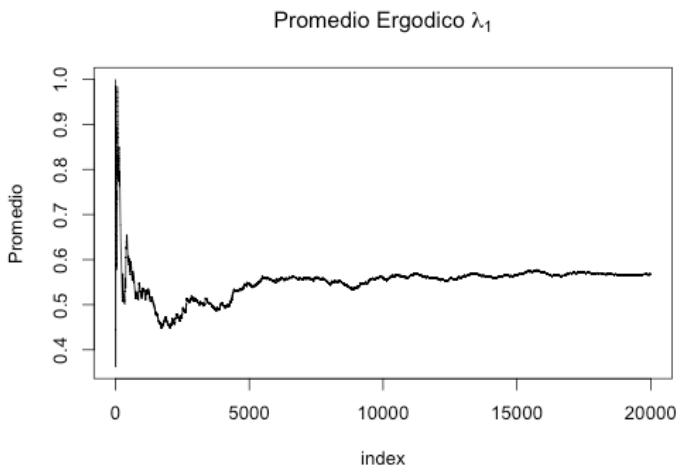
3 Generar una muestra $\theta^{(t+1)}$ de
$$p\left(\theta^{(t+1)} \middle| \lambda_1^{(t+1)}, \lambda_2^{(t+1)}, \boldsymbol{x}\right) = N\left(\theta \middle| \frac{\lambda_1^{(t+1)} x_1 + \lambda_2^{(t+1)} x_2}{\lambda_1^{(t+1)} + \lambda_2^{(t+1)}}, \frac{1}{\lambda_1^{(t+1)} + \lambda_2^{(t+1)}}\right)$$

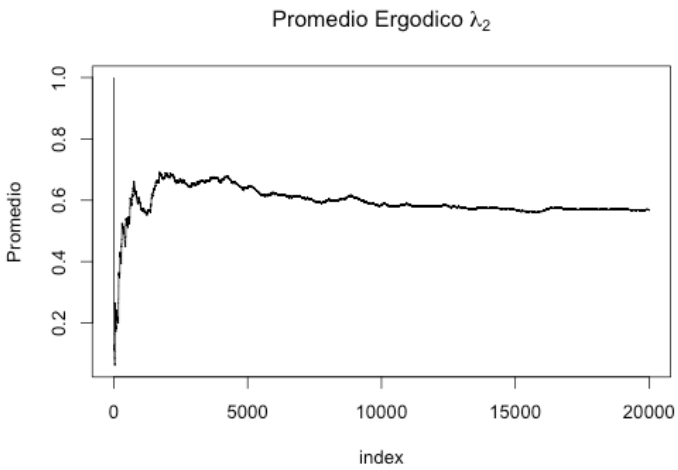
4 Con los pasos 1 a 3 construir $\boldsymbol{\theta}^{(t+1)} = \left(\theta^{(t+1)}, \lambda_1^{(t+1)}, \lambda_2^{(t+1)}\right)$

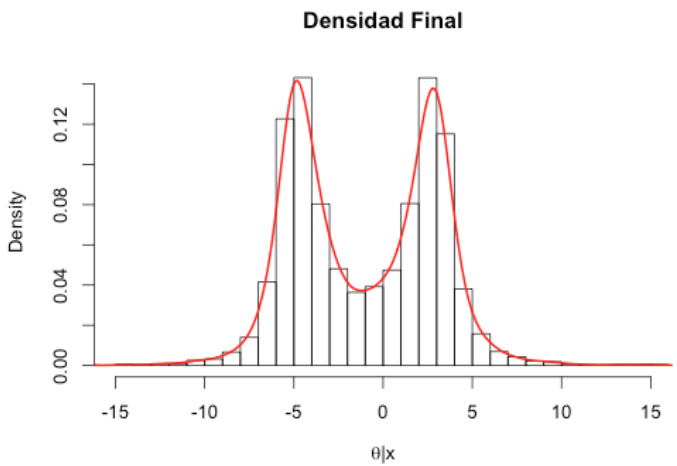
5 Repetir los pasos 1 a 4, y generar la cadena $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}, \ldots,$

Algoritmo de Gibbs







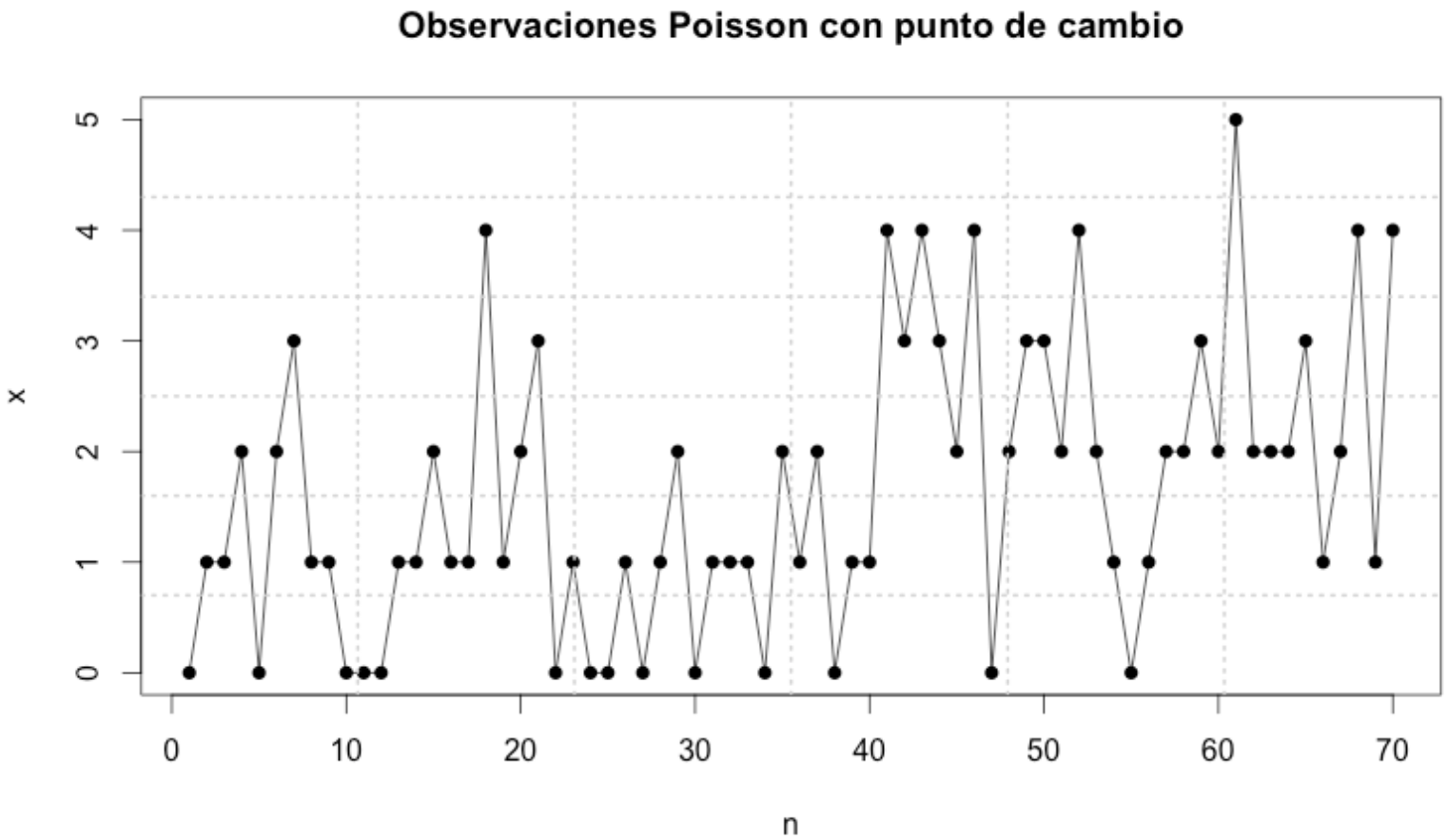


Ver código: Programa13.r

Algoritmo de Gibbs

Ejemplo 4: Se tienen $\{x_1, \ldots, x_n\}$ observaciones de una distribución Poisson en la que existe un punto de cambio en el proceso de observación, digamos en algún $m \in \{1, \ldots, n\}$. Es decir, condicional en el valor de m tenemos que:

$$x_i \sim Poi\,(x|\lambda)\,, \quad i \in \{1, \ldots, m\}\,; \qquad x_i \sim Poi\,(x|\phi)\,, \quad i \in \{m+1, \ldots, n\}$$



Algoritmo de Gibbs

El problema anterior cuenta con tres parámetros a estimar: λ, ϕ y m. Desde el punto de vista Bayesiano debemos asignar probabilidades iniciales:

- $p(\lambda|\alpha,\beta) = Ga(\lambda|\alpha_\lambda,\beta_\lambda)$
- $p(\phi|\alpha,\beta) = Ga(\phi|\alpha_\phi,\beta_\phi)$
- $p(m) = \frac{1}{n}; \quad m \in \{1,2,\ldots,n\}$

La regla de Bayes nos lleva a la siguiente distribución final:

$$p(\lambda,\phi,m|\boldsymbol{x}) \propto \lambda^{\alpha_\lambda t_m - 1}\phi^{\alpha_\phi + u_m - 1}\exp\{-(\beta_\lambda + m)\lambda\}\exp\{-(\beta_\phi + n - m)\phi\}$$

Donde $t_m = \sum_{i=1}^{m} x_i$ y $u_m = \sum_{i=m+1}^{n} x_i$.
En este caso $\theta = (\lambda,\phi,m)$, por lo que si queremos aplicar el algoritmo requeriremos las condicionales completas correspondientes:

$$p(\lambda|\phi,m,\boldsymbol{x}); \qquad p(\phi|\lambda,m,\boldsymbol{x}); \qquad p(m|\lambda,\phi,\boldsymbol{x})$$

*

Algoritmo de Gibbs

Las densidades condicionales completas tienen una forma fácil de simular:

- $p\left(\lambda|\phi,m,\boldsymbol{x}\right)=Ga\left(\lambda|\alpha_\lambda+t_m,\beta_\lambda+m\right)$
- $p\left(\phi|\lambda,m,\boldsymbol{x}\right)=Ga\left(\phi|\alpha_\phi+u_m,\beta_\phi+n-m\right)$
- $p\left(m|\lambda,\phi,\boldsymbol{x}\right)=\frac{\lambda^{\alpha_\lambda t_m-1}\phi^{\alpha_\phi+u_m-1}\exp\{-(\beta_\lambda+m)\lambda\}\exp\left\{-\left(\beta_\phi+n-m\right)\phi\right\}}{\sum_{l=1}^{n}\lambda^{\alpha_\lambda t_l-1}\phi^{\alpha_\phi+u_l-1}\exp\{-(\beta_\lambda+l)\lambda\}\exp\left\{-\left(\beta_\phi+n-l\right)\phi\right\}}$

*

Algoritmo de Gibbs

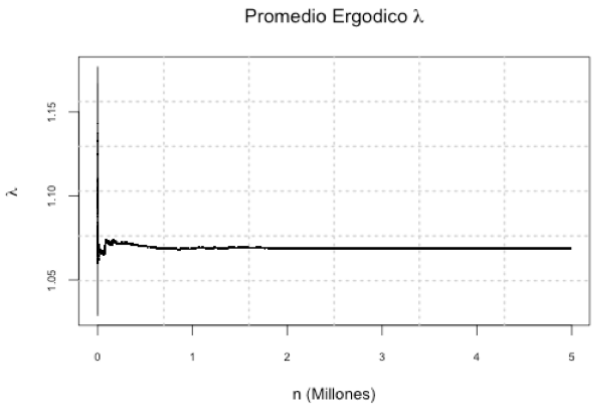
El Algoritmo de Gibbs para simular de $p(\lambda, \phi, m|\boldsymbol{x})$ es el siguiente:

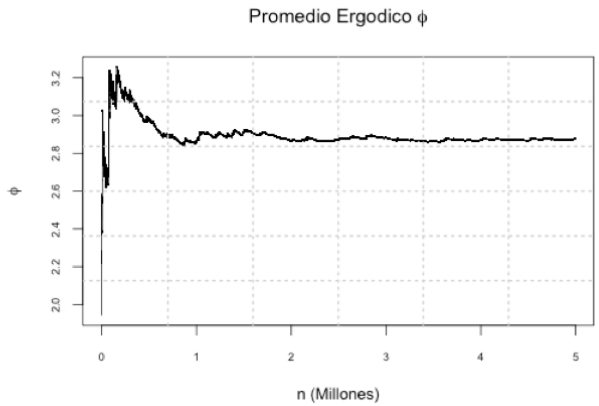
0. Inicializar la cadena en un valor inicial $\theta^{(0)} = \left(\lambda^{(0)}, \phi^{(0)}, m^{(0)}\right)$
   Para $t \in \{0, 1, 2, \ldots, \}$
1. Generar una muestra $\lambda^{(t+1)}$ de
   $p\left(\lambda^{(t)}|\phi^{(t)}, m^{(t)}, \boldsymbol{x}\right) = Ga\left(\lambda^{(t)}|\alpha_\lambda + t_{m^{(t)}, \beta_\lambda + m^{(t)}}\right)$
2. Generar una muestra $\phi^{(t+1)}$ de
   $p\left(\phi^{(t)}|m^{(t)}, \boldsymbol{x}\right) = Ga\left(\phi^{(t)}|\alpha_\phi + u_{m^{(t)}}, \beta_\phi + n - m^{(t)}\right)$
3. Generar una muestra $m^{(t+1)}$ de $p\left(m^{(t+1)}|\lambda^{(t+1)}, \phi^{(t+1)}, \boldsymbol{x}\right)$
4. Con los pasos 1 a 3 construir $\theta^{(t+1)} = \left(\lambda^{(t+1)}, \phi^{(t+1)}, m^{(t+1)}\right)$
5. Repetir los pasos 1 a 4, y generar la cadena $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots,$

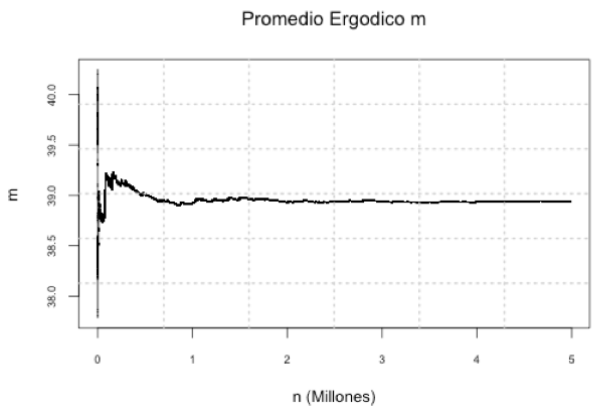
*

Algoritmo de Gibbs

Para ejemplificar: Supongamos $n = 70$, $\alpha_\lambda = \beta_\lambda = \alpha_\phi = \beta_\phi = 0.1$. Número de iteraciones: 5 millones.



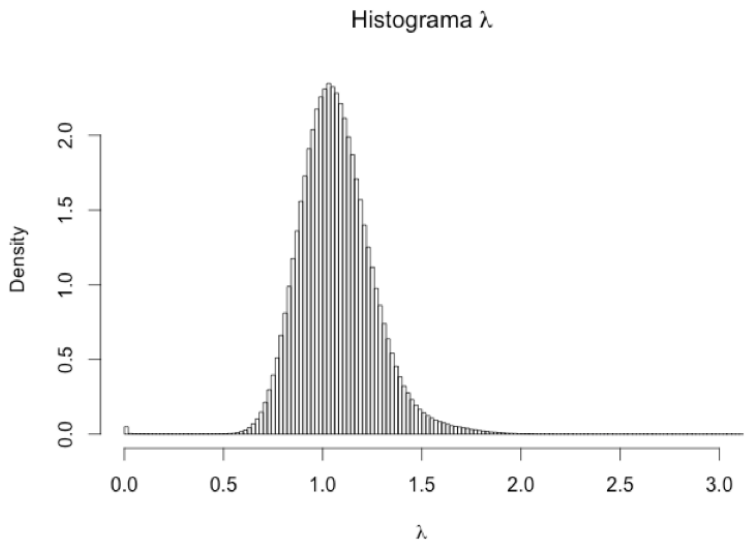


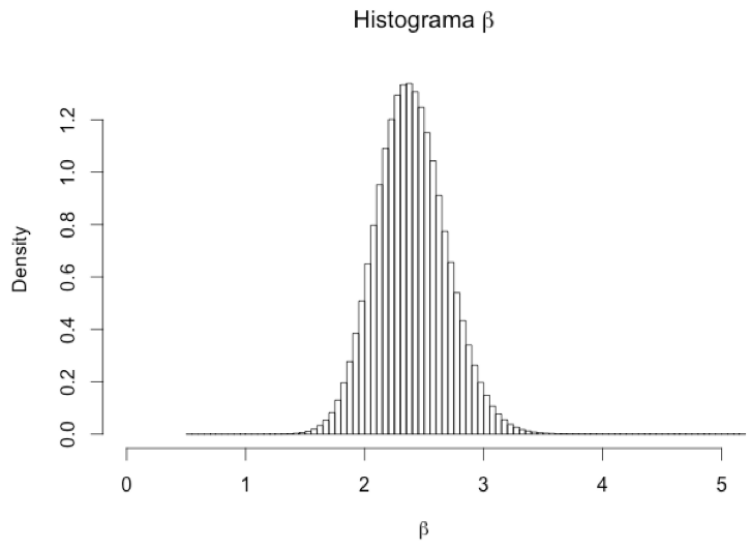


Ver código: Programa14.r

Algoritmo de Gibbs

Distribuciones finales marginales para λ y ϕ





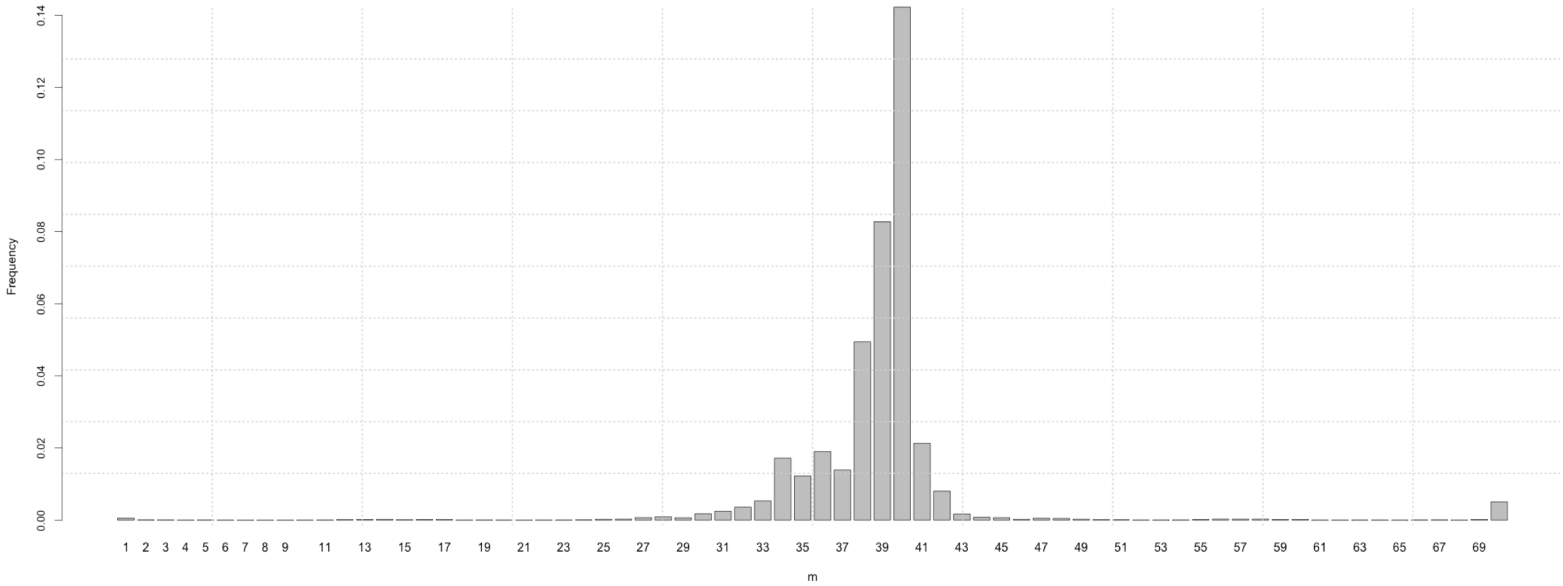
Intervalo de credibilidad de máxima densidad:

$$\mathbb{P}\left(0.722 < \lambda < 1.438\right) = 0.95 \quad \mathbb{P}\left(1.793 < \phi < 3.029\right) = 0.95$$

Ver código: Programa14.r

Algoritmo de Gibbs

Distribución final marginal para el punto de cambio:



Intervalo de credibilidad de máxima densidad:

$$\mathbb{P}\left(33 < m < 42\right) = 0.9500026$$

$$\text{Moda} = 40$$

Ver código: Programa14.r

¿Qué es JAGS?

"Just Another Gibbs Sampler"

JAGS (Plummer, 2013)

Es un programa para el análisis de modelos Bayesianos usando Monte Carlo vía Cadenas de Markov

¿Que es JAGS?

JAGS fue escrito para:

- Tener un motor para el lenguaje BUGS que corra en Unix, Mac y Windows
- Ser extendible
- Proporcionar una plataforma para experimentos con ideas de modelación Bayesiana

Corriendo un modelo en JAGS

Para obtener muestras de las distribuciones finales de los parámetros, JAGS realiza 5 pasos:

1. Definición del modelo
2. Compilación
3. Inicialización
4. Adaptación y *burn-in*
5. Monitoreo

Otras etapas del análisis se realizan fuera de JAGS; por ejemplo, diagnósticos de convergencia.

1. Definición del Modelo

Existen dos partes en la definición del modelo en JAGS: El modelo y los datos.

Descripción del modelo. El modelo se define en un archivo de texto usando el lenguaje BUGS.

Ej. Modelo Gamma (ambos parámetros desconocidos)

```
model.jags <- function()  {
   alpha ~ dgamma(0.01, 0.01)
   beta ~ dgamma(0.01, 0.01)
   for (i in 1:n){
     x[i] ~ dgamma(alpha, beta)
   }
  }
```

Datos

Los datos pueden ser dados en un archivo por separado o directamente en R.
Por ejemplo:

- En .txt

```
"x"<- c(1, 2, 3, 4, 5)
"Y" <- c(1, 3, 3, 3, 5)
"N" <- 5
```

- En R:

```
x<- c(1, 2, 3, 4, 5)
Y <- c(1, 3, 3, 3, 5)
N <- 5
datos<-list("Y","x","N")
```

cont..

- **2. Compilación**: Verifica si no hay errores de sintaxis
- **3. Inicialización:**
  - El usuario puede fijar los valores iniciales.
  - Si no se especifican los valores iniciales, un "valor típico" es obtenido de la distribución inicial (media, mediana o moda)
- **4. Burn-in**
- **5. Monitoreo:** Un objeto que registra los valores de los parámetros en cada iteración. (p.e. Trace monitor)

Ejemplo 3 : Gamma con 2 parámetros desconocidos.

$$x \sim Gamma\,(x|\alpha,\beta)\,; \qquad p\,(\alpha,\beta) = Gamma(\alpha|\alpha_0,\beta_0)Gamma(\beta|\alpha_1,\beta_1)$$

Asumiendo que:

$$\alpha_0 = \beta_0 = \alpha_1 = \beta_1 = 0.01$$

y que se observa la muestra de tamaño 5:

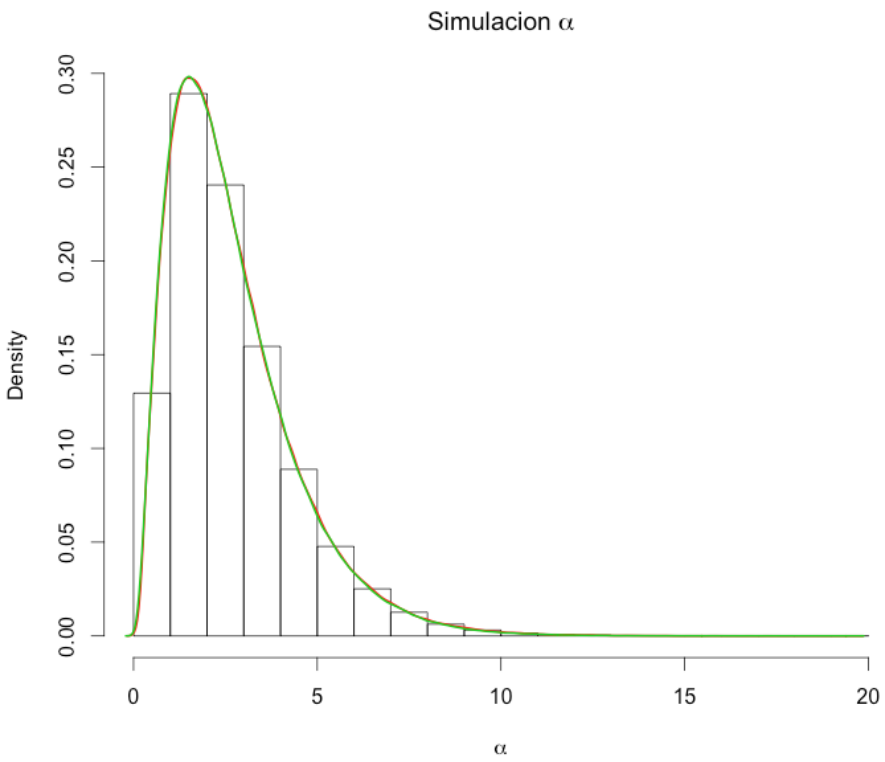
$$(0.4154325, 1.7853782, 1.7315852, 1.0254059, 1.9427045)$$

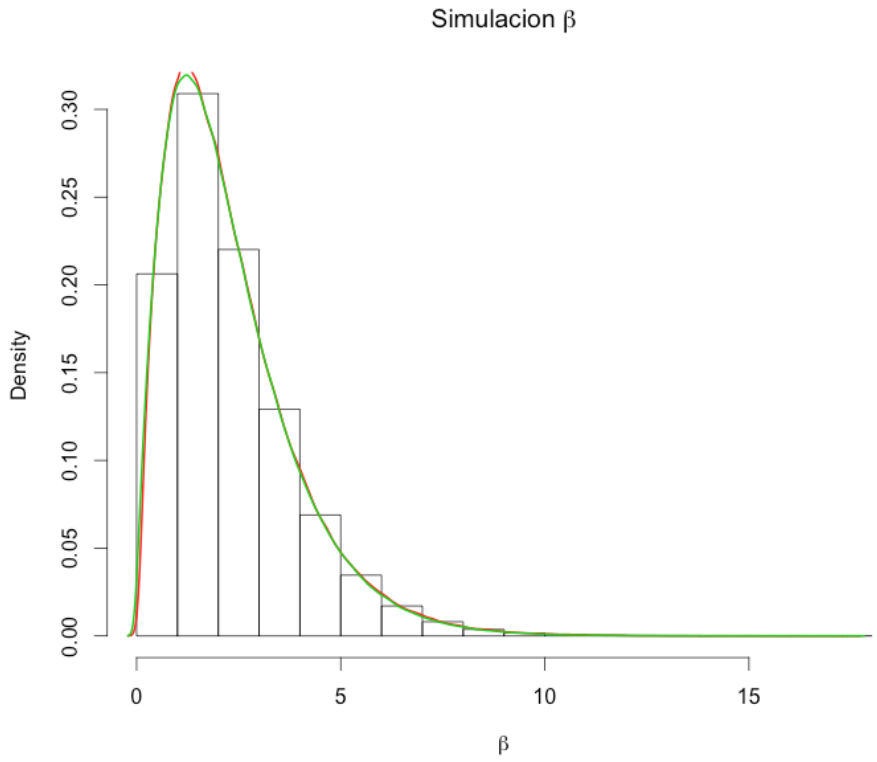
Haremos inferencias sobre α, β y x_F

```
model.jags <- function()  {
  for (i in 1:n){
    sample[i] ~ dgamma(alpha.jags, beta.jags)
  }
  alpha.jags ~ dgamma(0.01, 0.01)
  beta.jags ~ dgamma(0.01, 0.01)
  x.f ~ dgamma(alpha.jags, beta.jags)
}
```

Ver código: Programa15.r

Comparación de las distribuciones finales (MH programado en R vs JAGS)

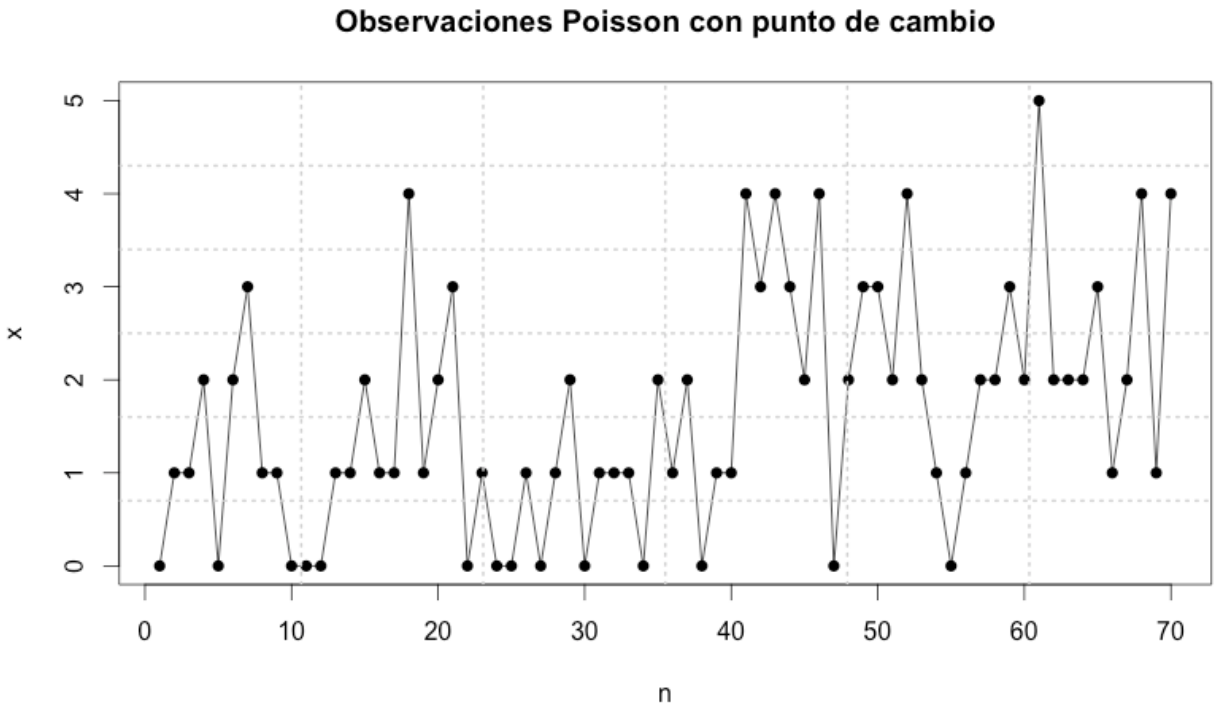




Ver código: Programa15.r

Ejemplo 4: Se tienen observaciones $\{x_1, \ldots, x_n\}$ de una distribución Poisson en la que existe un punto de cambio, digamos en $m \in \{1, \ldots, n\}$. Es decir, condicional en el valor de m tenemos que:

$$x_i \sim Poi\,(x|\lambda)\,, \quad i \in \{1, \ldots, m\}\,; \qquad x_i \sim Poi\,(x|\phi)\,, \quad i \in \{m+1, \ldots, n\}$$



Ver código: Programa16.r

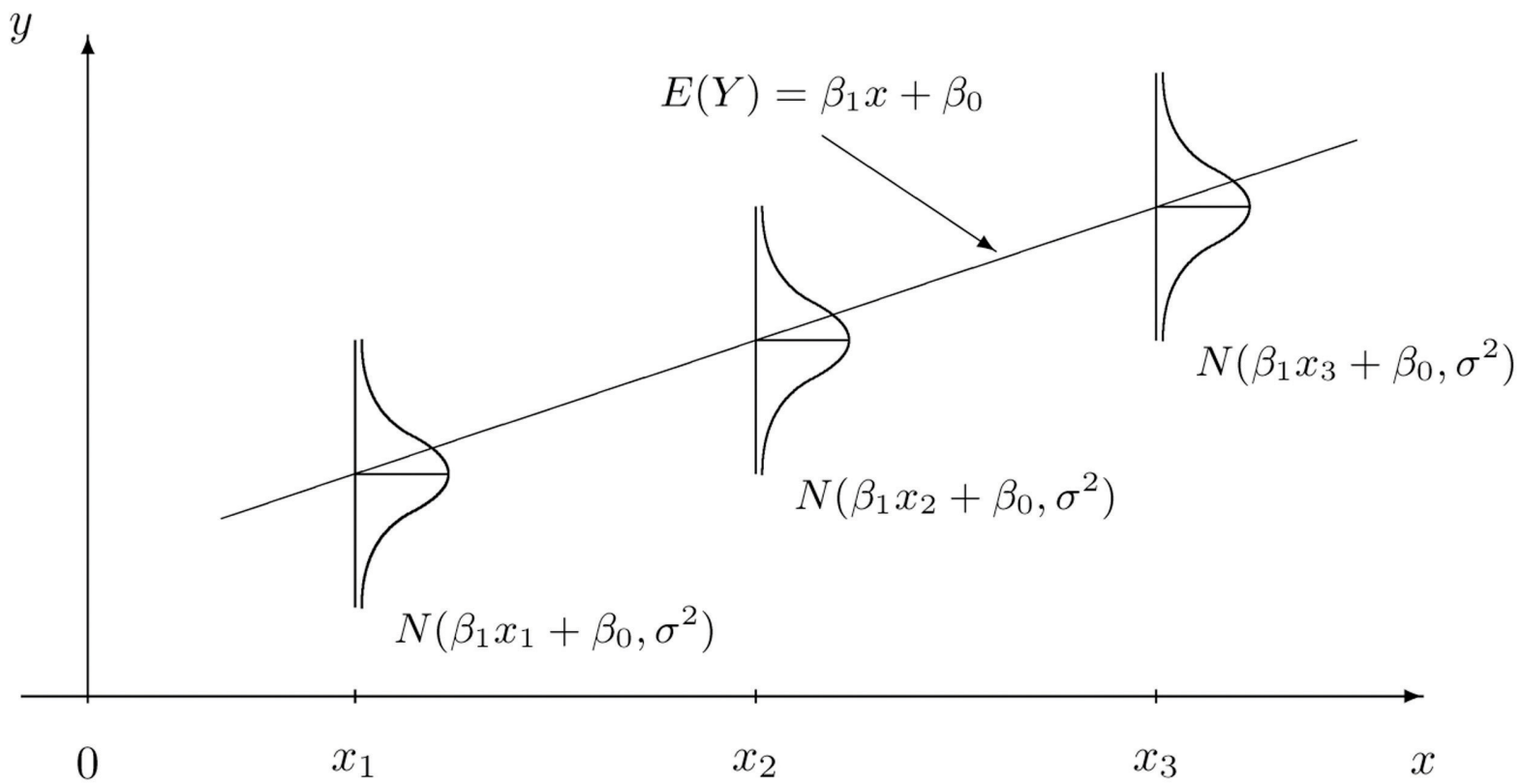
*

Artículos, libros, software


Plummer, Martyn. (2013).
JAGS Version 3.4.0 User Manual


Plummer, Martyn y Northcott, Bill (2013).
JAGS Version 3.4.0 Installation Manual

Modelos Lineales con JAGS



Modelos Lineales Generalizados

Dada una variable repuesta $\boldsymbol{y}$ con un conjunto de covariables $\boldsymbol{z}$, surge de manera natural preguntarnos cuál podría ser la relación funcional entre ellas. Una forma de modelarla podría ser:

$$\mathbb{E}(y \,|\, \boldsymbol{z}) = \mu(\boldsymbol{z})$$

donde, en general, $\mu(.)$ es una función desconocida. En la práctica es común aproximar a $\mu(.)$ a través de una función más simple (paramétrica):

$$\mu(\boldsymbol{z}) = \psi(\boldsymbol{z}; \boldsymbol{\beta})$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^t$ denota a un vector de parámetros desconocidos.

La forma mas simple para modelar la relación es suponer una función lineal de $\boldsymbol{\beta}$, es decir:

$$\psi\left(\boldsymbol{z};\boldsymbol{\beta}\right)=h\left(\beta_0+\beta_1 s_1\left(\boldsymbol{z}\right)+\ldots+\beta_k s_k\left(\boldsymbol{z}\right)\right)$$

donde s_i son funciones conocidas.

Finalmente esta función $\psi\left(\boldsymbol{z};\boldsymbol{\beta}\right)$ es tratada como si fuera la verdadera función que modelará el valor esperado de la variable respuesta y, por lo que el problema se reduce a hacer inferencias sobre el valor del vector de parámetros $\boldsymbol{\beta}$.

Es decir:

$$\mathbb{E}(y\,|\,\boldsymbol{z})=h\left(\beta_0+\beta_1 s_1\left(\boldsymbol{z}\right)+\ldots+\beta_k s_k\left(\boldsymbol{z}\right)\right)$$

o bien:

$$g\left(\mathbb{E}(y\,|\,\boldsymbol{z})\right)=\beta_0+\beta_1 s_1\left(\boldsymbol{z}\right)+\ldots+\beta_k s_k\left(\boldsymbol{z}\right)$$

De forma general, si se tienen n obsevaciones $(y_1, \boldsymbol{z}_1), \ldots, (y_n, \boldsymbol{z}_n)$, definimos el **modelo lineal generalizado** como:

$$g(\mu_i) = \eta_i$$

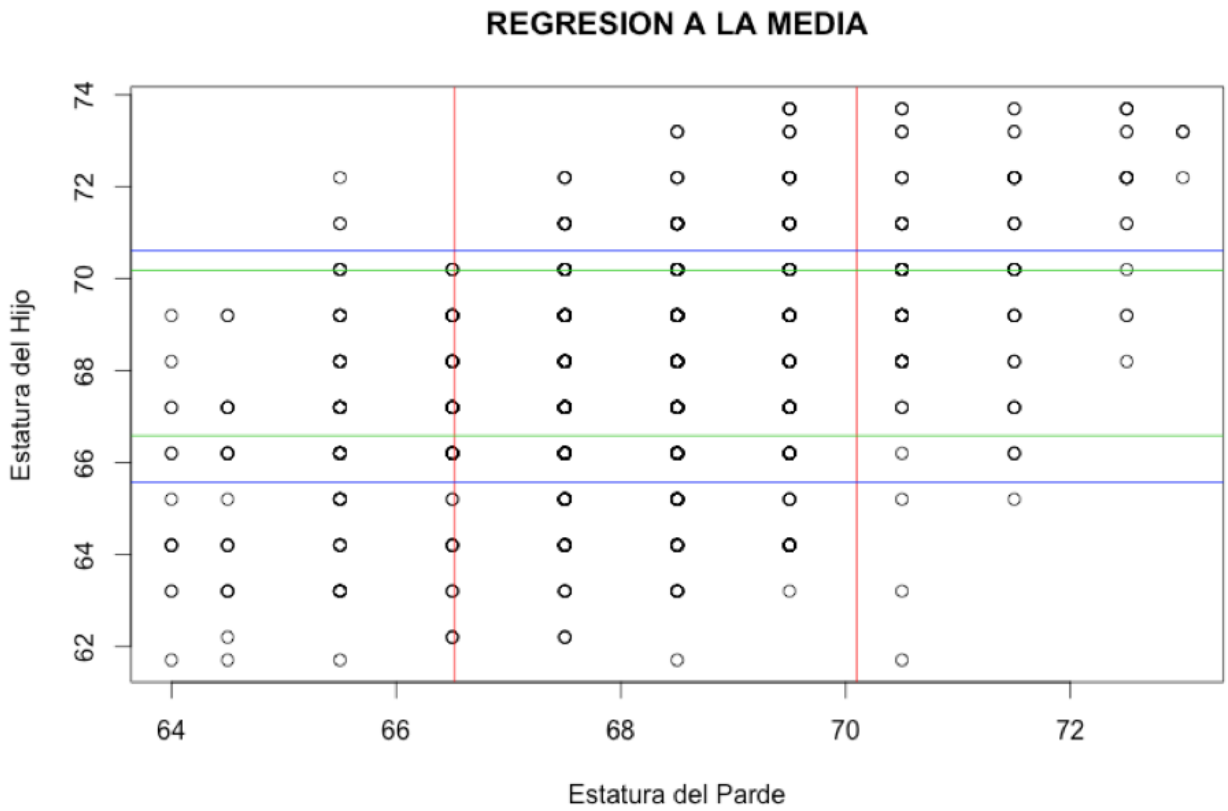
donde

- $(y_1, \ldots, y_n)$ se conoce como la **componente aleatoria** y las obsrevaciones se asumen independientes
- $\mu_i = \mathbb{E}(y_i \mid \boldsymbol{z}_i)$ es el valor esperado de y_i condicionado en los valores de las covariables $\boldsymbol{z}_i$
- $\eta_i = \beta_0 + \beta_1 s_1(\boldsymbol{z}_i) + \ldots + \beta_k s_k(\boldsymbol{z}_i)$, se conoce como la **componente sistemática** (predictor lineal)
- $g(.)$ función liga (o función vínculo), la cual relaciona a las componentes aleatoria y sistemática

Un caso muy utilizado es cuando g es la función identidad y y sigue una distribución Normal, dando origen al modelo de **Regresión Lineal**

Un poco de historia....

El término **regresión** fue acuñando por Francis Galton en el Siglo XIX en su artículo **Regression towards mediocrity in hereditary stature**, en donde observó que las alturas de los descendientes de ancestros altos tienden a regresar hacia abajo, hacia un promedio normal (un fenómeno conocido como regresión a la media).



Ver: data(Galton), del paquete HistData

Supongamos que se tiene n observaciones independientes $(y_1, \boldsymbol{z}_1), \ldots, (y_n, \boldsymbol{z}_n)$ del modelo:

$$y_i|\boldsymbol{z}_i \sim N\left(y_i|\, \mu\left(\boldsymbol{z}_i\right), \sigma^2\right) \left(\sigma^2 > 0, \text{desconocida}\right)$$

donde

$$\mu\left(\boldsymbol{z}_i\right) = \beta_0 + \beta_1 s_1\left(\boldsymbol{z}_i\right) + \ldots + \beta_k s_k\left(\boldsymbol{z}_i\right)$$

Definamos:

$$x_{ij} = s_j\left(\underline{z}_i\right) \quad i \in \{1, \ldots, n\}; \quad j \in \{1, \ldots, k\}$$

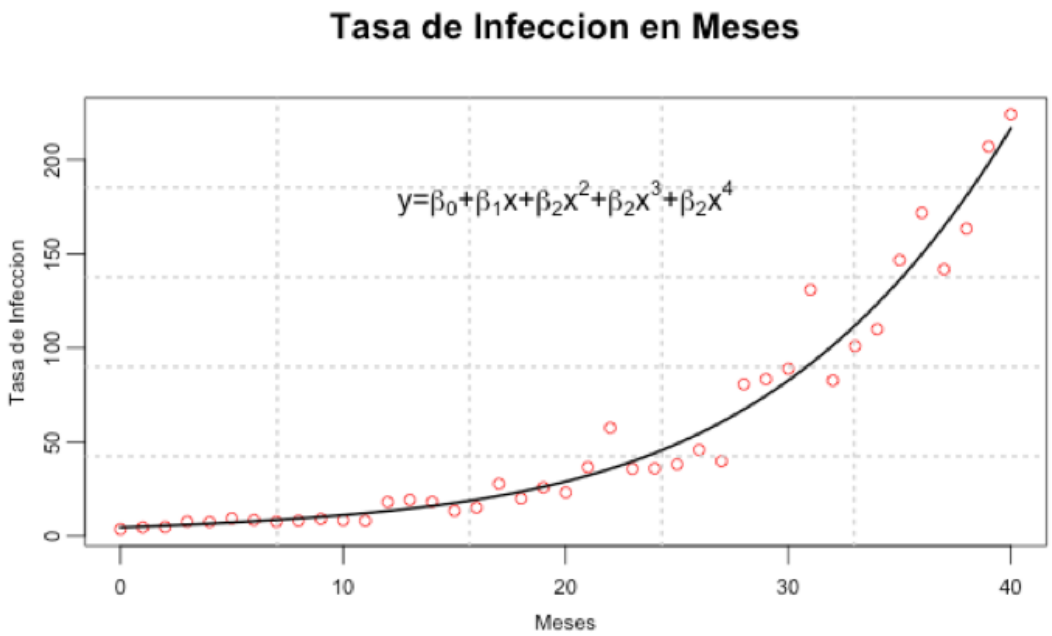
Entonces el modelo lo podemos escribir como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right) \tag{1}$$

Ejemplo:
Suponiendo que sólo tenemos una covariable $z \in \mathbb{R}$, y haciendo $s_j(z) = z^j$ entonces (1) toma la forma:

$$y_i = \beta_0 + \beta_1 z + \ldots + \beta_k z^k + \varepsilon_i \quad \varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right) \tag{2}$$

El modelo anterior pretende modelar el valor de y a través de una función polinomial de la covariable z.



Dada la relación lineal que hemos impuesto resulta conveniente utilizar una notación matricial y escribir (1) como: (haciendo $p = k + 1$)

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \qquad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right) \Rightarrow \boldsymbol{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$$

donde

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}_{n\times p} \qquad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

y

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{p\times 1}$$

Solución clásica del problema de estimación:

Dado que ahora conocemos la forma de distribución de $\boldsymbol{Y}$, podemos encontrar la función de verosimilitud

$$f\left(\boldsymbol{Y};\boldsymbol{\beta},\sigma^2\right)=\frac{1}{(2\pi)^{\frac{n}{2}}\left|\sigma^2\mathbf{I}_n\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})^t\left(\sigma^2\mathbf{I}_n\right)^{-1}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})}$$

Como $\left|\sigma^2\mathbf{I}_n\right|=\sigma^{2n}$ y $\left(\sigma^2\mathbf{I}_n\right)^{-1}=\frac{1}{\sigma^2}\mathbf{I}_n$, entonces la verosimilitud es:

$$\mathscr{L}\left(\boldsymbol{\beta},\sigma^2;\boldsymbol{Y}\right)=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})^t(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})}$$

Sacamos el logaritmo de la verosimilitud:

$$\log\mathscr{L}\left(\boldsymbol{\beta},\sigma^2;\boldsymbol{Y}\right)=-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\right)^t\left(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\right)$$

Maximizamos con respecto a $\left(\boldsymbol{\beta},\sigma^2\right)$; para ello derivamos e igualamos a cero

$$\frac{d}{d\boldsymbol{\beta}}\log\mathscr{L}\left(\boldsymbol{\beta},\sigma^2;\boldsymbol{Y}\right)=-\frac{1}{2\sigma^2}\left(-2\mathbf{X}^t\boldsymbol{Y}+2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}\right)\tag{3}$$

$$\frac{d}{d\sigma^2}\log\mathscr{L}\left(\boldsymbol{\beta},\sigma^2;\boldsymbol{Y}\right)=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\left(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\right)^t\left(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\right)\tag{4}$$

*

De la ecuación (3) obtenemos las ecuaciones normales, es decir:

$$\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\boldsymbol{Y} \tag{5}$$

Notamos que (5) tiene solución única si y sólo si la matriz $\mathbf{X}^t\mathbf{X}$ es invertible ($\mathbf{X}$ de rango completo) en cuyo caso el estimador máximo verosímil para $\boldsymbol{\beta}$ es:

$$\hat{\boldsymbol{\beta}}_{MV} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\boldsymbol{Y} \tag{6}$$

Para σ^2, de la ecuación (4) obtenemos:

$$-n + \frac{1}{\sigma^2}\left(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\right)^t\left(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\right) = 0 \Rightarrow \sigma^2 = \frac{\left(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\right)^t\left(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\right)}{n}$$

Por lo tanto, al sustituir en la última igualdad lo que obtuvimos en la ecuación (6) obtenemos que estimador máximo verosímil para σ^2 es:

$$\hat{\sigma}^2_{MV} = \frac{\left(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{MV}\right)^t\left(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{MV}\right)}{n}$$

Definiendo $\hat{\boldsymbol{Y}} := \mathbf{X}\hat{\boldsymbol{\beta}}_{MV}$, tenemos:

$$\hat{\sigma}^2_{MV} = \frac{\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)^t\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Los estimadores máximo verosímiles gozan de las siguientes propiedades:

- $\mathbb{E}\left(\hat{\boldsymbol{\beta}}_{MV}\right) = \boldsymbol{\beta}$ (insesgamiento)
- $\text{Var}\left(\hat{\boldsymbol{\beta}}_{MV}\right) = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$
- $\hat{\boldsymbol{\beta}}_{MV} \sim N_p\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right)$
- $\mathbb{E}\left(\hat{\sigma}^2_{MV}\right) = \sigma^2 \frac{n-p}{n}$ (sesgado)
- $\text{Var}\left(\hat{\sigma}^2_{MV}\right) = 2\frac{n-p}{n^2}\sigma^4$
- $\hat{\sigma}^2_{MV} \sim Gamma\left(\frac{n-p}{2}, \frac{n}{2\sigma^2}\right)$
- $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ (insesgamiento)
- Como $\hat{\boldsymbol{\beta}}_{MV} \sim N_p\left(\underline{\beta}, \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right)$, entonces haciendo $\mathbf{C} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ y definiendo C_{ij} al elemento (i,j) de la matriz $\mathbf{C}$ se tiene que:

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 C_{(i+1)(i+1)}\right) \Rightarrow \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(i+1)}}} \sim t_{(n-p)} \tag{7}$$

De la última expresión de (7), la inferencia clásica desprende las pruebas de hipótesis e intervalos de confianza correspondientes para el parámetro β_i con $i \in \{0, \ldots, k\}$

Regresión Lineal : Enfoque Bayesiano

Sea $\boldsymbol{\beta} \in \mathbb{R}^p$ un vector de parámetros, y $\mathbf{X} \in \mathbb{R}^{n \times p}$ una matriz de diseño conocida. Definamos el modelo lineal:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8}$$

donde $\boldsymbol{\varepsilon} \sim N_n\left(\underline{\varepsilon}|\mathbf{0}, \tau\mathbf{I}_n\right)$ con $\tau = \sigma^{-2} > 0$. Suponemos una matriz de precisión con correlación 0, entre la variables lo que implica independencia entre las ε_i. De la ecuación (8) concluimos que:

$$\boldsymbol{y} \sim N_n\left(\boldsymbol{y}|\mathbf{X}\boldsymbol{\beta}, \tau\mathbf{I}_n\right)$$

Objetivo: suponiendo que observamos $\boldsymbol{y}$, **inferir** sobre los parámetros $\boldsymbol{\beta}$ y $\tau = \sigma^{-2}$
Solución: (La receta) Encontrar las distribuciones finales:

$$p\left(\boldsymbol{\beta}, \tau|\boldsymbol{y}\right) \propto p\left(\boldsymbol{y}|\boldsymbol{\beta}, \tau\right) p\left(\boldsymbol{\beta}, \tau\right)$$

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \int_0^\infty p\left(\boldsymbol{\beta}, \tau|\boldsymbol{y}\right) d\tau; \qquad p\left(\tau|\boldsymbol{y}\right) = \int_{\mathbb{R}^p} p\left(\boldsymbol{\beta}, \tau|\boldsymbol{y}\right) d\boldsymbol{\beta};$$

Bajo los supuestos que tiene este modelo, es posible construir distribuciones conjugadas que permiten encontrar distribuciones finales exactas que no requieren uso de herramientas de simulación para su estudio.

Para proponer una distribución conjugada, se estudia la verosimilitud.

$$p(\boldsymbol{y}|\boldsymbol{\beta},\tau) \propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\left((\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}-\underline{\hat{\beta}})+(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right)} \quad (9)$$

$$\propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\left((\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+\tilde{\beta}_0\right)} \quad (10)$$

Vista como función de $(\boldsymbol{\beta},\tau)$ aparece el denominado **kernel** de una distribución **Normal-Multivariada - Gamma** !!

*

Regresión Lineal : Enfoque Bayesiano

En la literatura se propone como inicial una distribución Normal-Multivariada - Gamma

$$p\left(\boldsymbol{\beta}, \tau | \boldsymbol{\mu}_0, \mathbf{P}_0, \alpha_0, \delta_0\right) \propto N_p\left(\boldsymbol{\beta} | \boldsymbol{\mu}_0, \tau \mathbf{P}_0\right) Ga\left(\tau | \alpha_0, \delta_0\right)$$

donde $\mathbf{P}_0 \in \mathbb{R}^{p \times p}$, $\boldsymbol{\mu}_0 \in \mathbb{R}^p$, $\alpha_0, \delta_0 \in \mathbb{R}$ son hiperparámetros.

De las propiedades de esta densidad se obtiene que:

$$\boldsymbol{\beta} \sim T_p\left(\boldsymbol{\beta} \middle| 2\alpha_0, \boldsymbol{\mu}_0, \frac{\alpha_0}{\delta_0}\mathbf{P}_0\right) \quad \Rightarrow \quad \mathbb{E}(\boldsymbol{\beta}) = \boldsymbol{\mu}_0; \quad \text{Var}(\boldsymbol{\beta}) = \frac{\delta_0}{\alpha_0 - 1}\mathbf{P}_0^{-1}$$

$$\tau \sim Ga\left(\tau | \alpha_0, \delta_0\right) \quad \Rightarrow \quad \mathbb{E}(\tau) = \frac{\alpha_0}{\delta_0}; \quad \text{Var}(\tau) = \frac{\alpha_0}{\delta_0^2}$$

*

Regresión Lineal : Enfoque Bayesiano

Definida la distribución inicial tenemos todos los ingredientes para obtener la final, usando nuestra receta!!
Tras un poco de algebra se obtiene que:

$$p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) = N_p(\boldsymbol{\beta}|\boldsymbol{\mu}_1, \tau\mathbf{P}_1)\, Ga(\tau|\alpha_1, \delta_1)$$

donde:

- $\mathbf{P}_1 = \mathbf{X}^T\mathbf{X} + \mathbf{P}_0$
- $\mu_1 = \left(\mathbf{X}^T\mathbf{X} + \mathbf{P}_0\right)^{-1}\left(\mathbf{X}^T\underline{y} + \mathbf{P}_0\mu_0\right)$
- $\alpha_1 = \frac{n}{2} + \alpha_0$
- $\delta_1 = \delta_0 + \frac{1}{2}\left(\left(\underline{y} - \mathbf{X}\mu_1\right)^T\left(\underline{y} - \mathbf{X}\mu_1\right) + (\mu_1 - \mu_0)^T\mathbf{P}_0(\mu_1 - \mu_0)\right)$

Finalmente, sabemos que, por propiedades de esta distribución,

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = T_p\left(\boldsymbol{\beta}\left|2\alpha_1, \mu_1, \frac{\alpha_1}{\delta_1}\mathbf{P}_1\right.\right)$$

$$p(\tau|\boldsymbol{y}) = Ga(\alpha_1, \delta_1) \Rightarrow p\left(\tau^{-1}|\underline{y}\right) = IGa(\alpha_1, \delta_1)$$

Regresión Lineal: Inferencia sobre y^*

Si se desea hacer inferencia sobre nuevas observaciones del modelo, también se obtiene fórmulas cerradas:

$$\mathbf{w} = \mathbf{Z}\underline{\beta} + \mathbf{e}$$

donde

- $\mathbf{Z} \in \mathbb{R}^{k \times p}$ es una nueva matriz de covariables.
- $\mathbf{e} \sim N_k\left(\mathbf{e}|\underline{0}, \tau\mathbf{I}\right)$
- $\mathbf{w} \sim N_k\left(\mathbf{w}|\mathbf{Z}\underline{\beta}, \tau\mathbf{I}\right)$

Desde el punto de vista Bayesiano, el objetivo es determinar $p\left(\mathbf{w}|\boldsymbol{y}\right)$

$$\begin{aligned} p\left(\mathbf{w}|\boldsymbol{y}\right) &= \int_0^\infty \int_{\mathbb{R}^p} p\left(\mathbf{w}, \boldsymbol{\beta}, \tau|\boldsymbol{y}\right) d\boldsymbol{\beta} d\tau \\ &= .... \\ &= T_k\left(\mathbf{w}\left|2\alpha_1, \mathbf{Z}\mu_1, \frac{\alpha_1}{\delta_1}\left(\mathbf{I} + \mathbf{Z}\mathbf{P}_1\mathbf{Z}^T\right)^{-1}\right.\right) \end{aligned}$$

*

Regresión Lineal: Distribuciones no Informativas

En estos casos es posible, además, construir distribuciones no informativas.
- **Jeffreys**:

$$p(\boldsymbol{\beta}, \tau) \quad \propto \quad \tau^{\frac{p-2}{2}}$$

La cual da origen a la final

$$p\left(\boldsymbol{\beta}, \tau | \underline{y}\right) = NG\left(\boldsymbol{\beta}, \tau \left| \hat{\boldsymbol{\beta}}, \mathbf{X}^T\mathbf{X}, \frac{n}{2}, \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right.\right)$$

- Distribución de **Referencia**:

$$p(\boldsymbol{\beta}, \tau) \propto \tau^{-1}$$

La cual da origen a la final

$$p(\boldsymbol{\beta}, \tau | \mathbf{y}) = NG\left(\boldsymbol{\beta}, \tau \left| \hat{\boldsymbol{\beta}}, \mathbf{X}^T\mathbf{X}, \frac{n-p}{2}, \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right.\right)$$

*

Regresión Lineal con JAGS

Para ilustrar el uso de la simulación en este modelo, supongamos que se propone una inicial de la siguiente forma:

$$p\left(\boldsymbol{\beta}, \tau | \boldsymbol{\mu}_0, \alpha_0, \delta_0\right) = N_p\left(\boldsymbol{\beta} | \boldsymbol{\mu}_0, \mathbf{I}\right) Ga\left(\tau | \alpha_0, \delta_0\right)$$

(Asume independencia entre los parámetros de forma inicial)
En este caso no hay conjugación con la verosimilitud, por lo que no necesariamente se puede llegar a formas cerradas para la distribución final

$$p\left(\boldsymbol{\beta}, \tau | \mathbf{y}\right) \propto p\left(\mathbf{y} | \boldsymbol{\beta}, \tau\right) N_p\left(\boldsymbol{\beta} | \boldsymbol{\mu}_0, \mathbf{I}\right) Ga\left(\tau | \alpha_0, \delta_0\right)$$

El objetivo es, entonces, construir una cadena de Markov cuya distribución final sea precisamente $p\left(\boldsymbol{\beta}, \tau | \mathbf{y}\right)$.

¿Cómo describimos un modelo lineal simple en JAGS?
El modelo lineal simple es:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \qquad \varepsilon_i \sim N(\varepsilon_i|0,\tau); \qquad y_i \sim N(y_i|\beta_0 + \beta_1 x_i, \tau)$$

Asumiendo las iniciales

$$\tau \sim Ga(\tau|0.01, 0.01); \quad \beta_0 \sim N(\beta_0|0, 0.001); \quad \beta_1 \sim N(\beta_1|0, 0.001).$$

(Obs: Asumimos independencia de forma inicial entre los parámetros y una precisión pequeña, lo que refleja poca información inicial)

Asumiendo que tenemos los vectores x, y de longitud n, entonces el modelo es:

```
model.jags <- function()  {
   beta0 ~ dnorm(0, 0.001)
   beta1 ~ dnorm(0, 0.001)
   tau ~ dgamma(0.01,0.01)
   for (i in 1:n){
     y[i] ~ dnorm(beta0+beta1*x[i], tau)
   }
  }
```

¿Cómo describimos un modelo lineal múltiple en JAGS?
El modelo lineal múltiple es:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad \varepsilon_i \sim N(\varepsilon_i|0, \tau)$$

En este caso:

$$\tau \sim Ga(\tau|0.01, 0.01); \quad \beta_i \sim N(\beta_0|0, 0.001); \quad i \in \{0, 1, \ldots k\}$$

Asumiendo que tenemos la matriz de diseño X, y el vector y, entonces el modelo es:

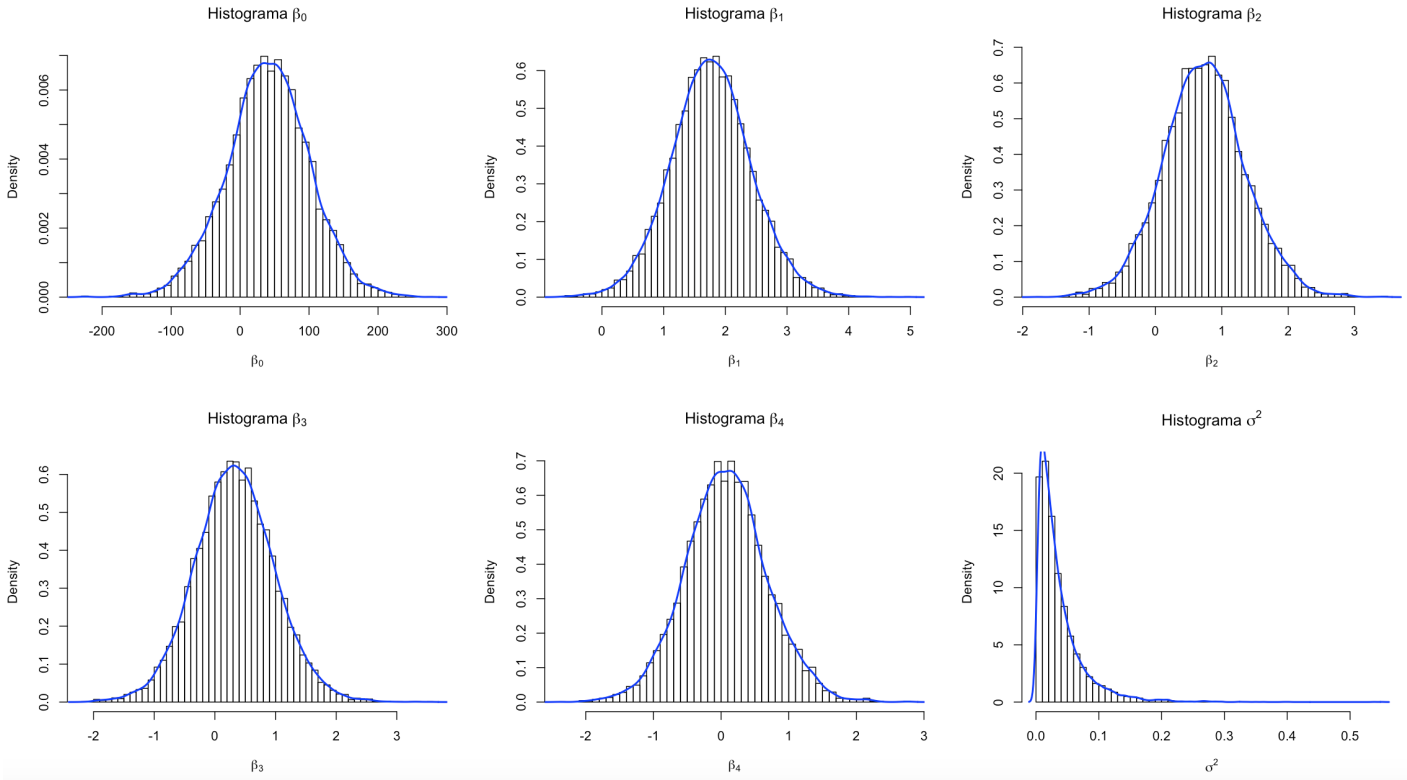
```
model.jags <- function()  {
for (i in 1:(k+1)){
   beta[i] ~ dnorm(0, 0.001)
   }
 tau ~ dgamma(0.01,0.01)
     for (i in 1:n){
       y[i] ~ dnorm(X[i,]%*%beta, tau)
   }
  }
```

Ejemplo: El siguiente conjunto de datos (Hald, 1952), también descrito en Draper and Smith (1981), consiste de 13 observaciones que relacionan el calor producido por el endurecimiento de cierto tipo de cemento con cuatro variables explicativas, cada una midiendo el contenido de un ingrediente en particular (en porcentajes).

Y	x_1	x_2	x_3	x_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Ver código: Programa17.r

Se generó una cadena durante 20,000 iteraciones con un calentamiento de 10,000. Las densidades finales de cada uno de los parámetros son:



Ver código: Programa17.r

Modelo Logístico

El modelo **logístico** asume lo siguiente:

- $(y_1, \ldots, y_n)$ siguen una distribución Bernoulli.
- $\mu_i = \mathbb{E}(y_i \mid \boldsymbol{z}_i) = p_i \in (0, 1)$
- $\eta_i = \beta_0 + \beta_1 s_1(\boldsymbol{z}_i) + \ldots + \beta_k s_k(\boldsymbol{z}_i)$ la **componente sistemática** (predictor lineal)
- $g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$, conocida como función logit(p_i)

Obs: En este caso se modela la probabilidad de éxito de y_i por medio de:

$$p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

¿Cómo describimos un modelo logístico en JAGS?
El modelo asume

$$y_i \sim Bernoulli\left(y_i|p_i\right) = Bernoulli\left(y_i \left| \frac{\exp\left(\eta_i\right)}{1+\exp\left(\eta_i\right)}\right.\right)$$

Supongamos las iniciales:

$$\beta_i \sim N(\beta_i|0, 0.001); \quad i \in \{0, 1, \ldots, k\}$$

Asumiendo que tenemos la matriz de diseño X, y el vector y, entonces el modelo es:

```
model.jags <- function()  {
      for (i in 1:(k+1)){
            beta[i] ~ dnorm(0, 0.001)
       }
      for (i in 1:n){
            logit(p[i]) <- X[i,]%*% beta
            y[i] ~ dbin(p[i], 1)
       }
  }
```

Cuando en este modelo tenemos repeticiones de las observaciones (*Bernoulli*) para cierto nivel de las covariables z, se utiliza la distribución Binomial en el componente aleatorio.

Ejemplo 2. (Gelman *et al.*, 1995) En el desarrollo de drogas y otros compuestos químicos, es normal realizar pruebas de toxicidad (experimentos de bioensayo) en grupos de animales para evaluar la conveniencia de utilizar las drogas de manera comercial. Los experimentos consisten en administrar diferentes dosis (niveles) de la droga o compuesto a diferentes grupos de animales. En general, la respuesta es dicotómica; por ejemplo, el animal vive o muere, tiene tumor o no lo tiene, o se contagia o no se contagia. Así, la información que se recolecta es de la forma

$$D = \{(x_i, n_i, y_i) : i \in J_k\},$$

en donde x_i representa el i-ésimo de k niveles de la droga que se administró a un grupo de n_i animales de los que y_i responden negativamente. La siguiente tabla muestra un conjunto de resultados de un experimento ficticio.

Dosis, x_i (log g/ml)	Número de animales, n_i	Número de muertes, y_i
-0.863	5	0
-0.296	5	1
-0.053	5	3
0.727	5	5

Para el problema anterior planteamos el modelo logístico de la siguiente forma:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i$$

de donde se puede despejar p_i de tal manera que la probabilidad de que el animal responda negativamente (muera) con una dosis igual a x_i es:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

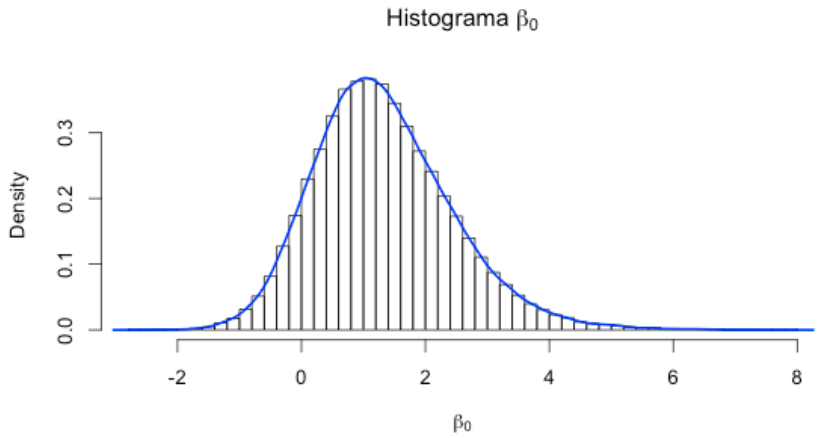
Un objetivo importante en estos estudios es determinar la dosis mediana (LD50) que es la dosis en la que la probabilidad de muerte es precisamente 0.5. En este caso, se tiene que resolver la ecuación:

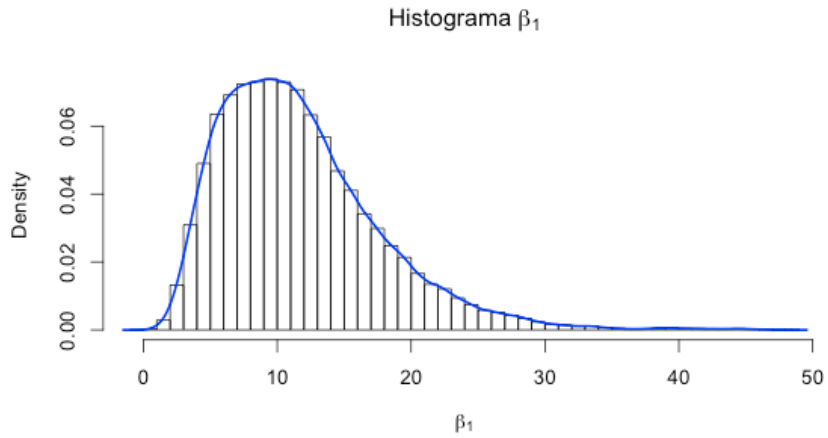
$$0.5 = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

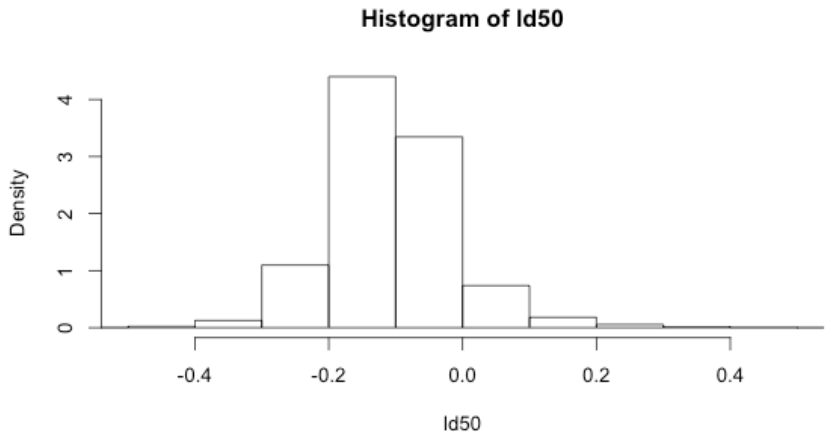
de donde se obtiene que:

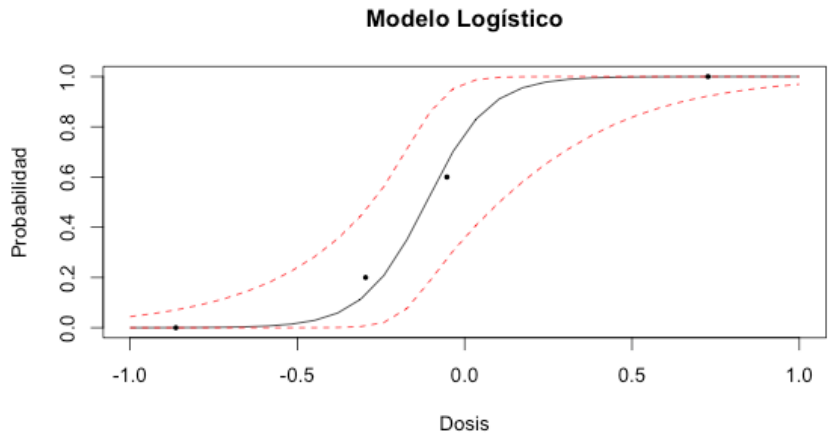
$$LD50 = -\frac{\beta_0}{\beta_1}.$$

Es entonces de interés hacer inferencias sobre esta cantidad en este tipo de problemas.
Ver código: Programa18.r









Ver código: Programa18.r

Regresión Poisson

El modelo lineal Poisson asume lo siguiente:

- $(y_1, \ldots, y_n)$ siguen una distribución Poisson.
- $\mu_i = \mathbb{E}(y_i \mid \boldsymbol{z}_i) = \lambda_i \in \mathbb{R}^+$
- $\eta_i = \beta_0 + \beta_1 s_1(\boldsymbol{z}_i) + \ldots + \beta_k s_k(\boldsymbol{z}_i)$ la **componente sistemática** (predictor lineal)
- $g(\lambda_i) = \log(\lambda_i)$.

Obs: En este caso se modela la tasa λ_i de la variable y_i como

$$\lambda_i = \exp(\eta_i)$$

¿Cómo describimos un modelo Poisson en JAGS?
El modelo asume

$$y_i \sim Poisson\left(y_i|\lambda_i\right) = Poisson\left(y_i\,|\exp\left(\eta_i\right)\right)$$

Supongamos las siguientes iniciales en los parámetros del predictor lineal

$$\beta_i \sim N(\beta_i|0, 0.001); \quad i \in \{0, 1, \ldots, k\}$$

Asumiendo que tenemos la matriz de diseño X, y el vector y, entonces el modelo es:

```
model.jags <- function()  {
      for (i in 1:(k+1)){
            beta[i] ~ dnorm(0, 0.001)
       }
      for (i in 1:n){
            log(lambda[i]) <- X[i,]%*%beta
            y[i] ~ dpois(lambda[i])
       }
}
```

Ver código: Programa19.r

*

Regresión Gamma

El modelo Gamma asume lo siguiente:

- $(y_1, \ldots, y_n)$ siguen una distribución $Gamma(\alpha, \delta_i)$
- $\mu_i = \mathbb{E}(y_i \mid \boldsymbol{z}_i) = \frac{\alpha}{\delta_i} = \mu_i \in \mathbb{R}^+, \Rightarrow \delta_i = \frac{\alpha}{\mu_i}$
- $\eta_i = \beta_0 + \beta_1 s_1(\boldsymbol{z}_i) + \ldots + \beta_k s_k(\boldsymbol{z}_i)$ la **componente sistemática** (predictor lineal)
- $g(\mu_i) = \log(\mu_i)$.

Obs: En este caso se modela la media μ_i de la variable y_i como

$$\mu_i = \exp(\eta_i)$$

¿Cómo describimos un modelo Gamma en JAGS?
El modelo asume

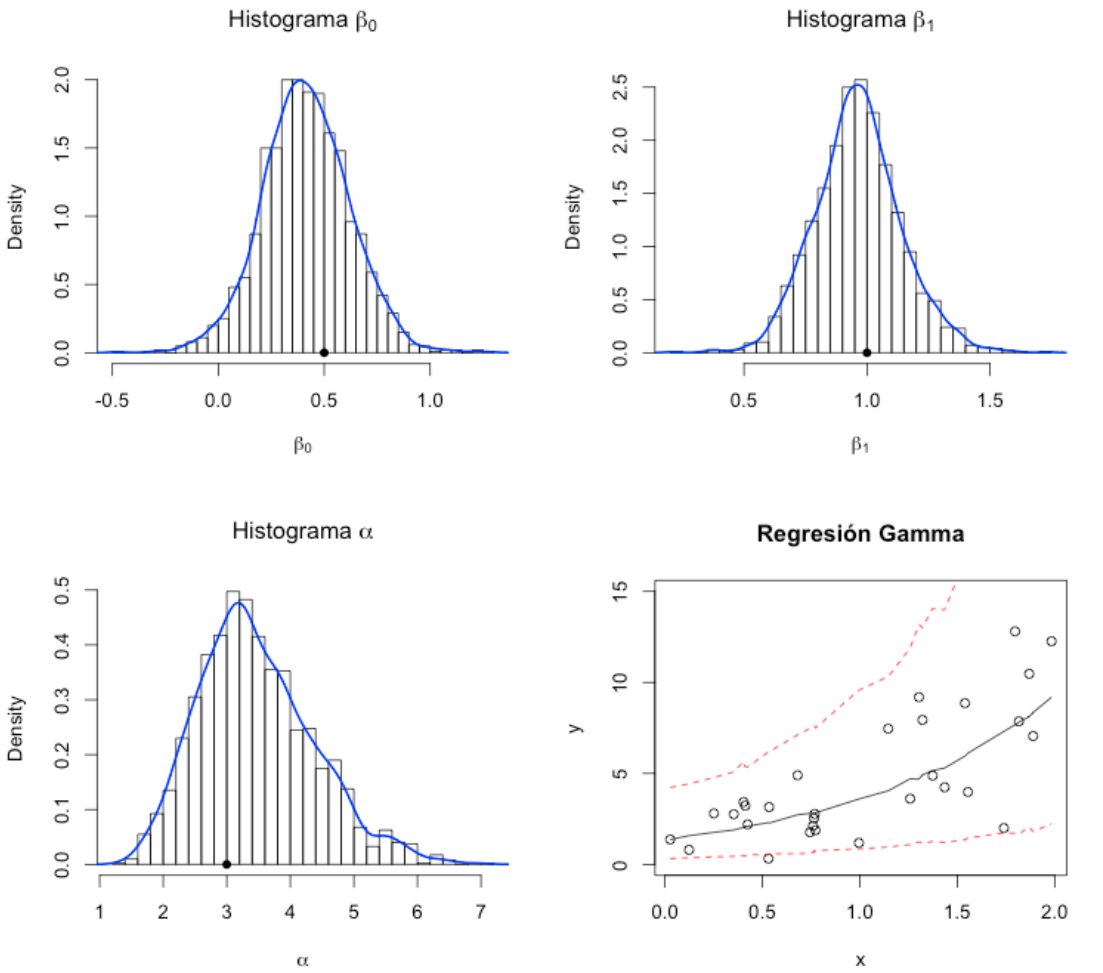
$$y_i \sim Gamma\left(y_i|\alpha,\delta_i\right) = Gamma\left(y_i\left|\alpha,\frac{\alpha}{\exp\left(\eta_i\right)}\right.\right)$$

Asumiendo las siguientes iniciales en los parámetros del predictor lineal y para el parámetro α

$$\beta_i \sim N(\beta_i|0,0.001); \quad i \in \{0,1,\ldots,k\}; \qquad \alpha \sim Ga(\alpha|0.01,0.01)$$

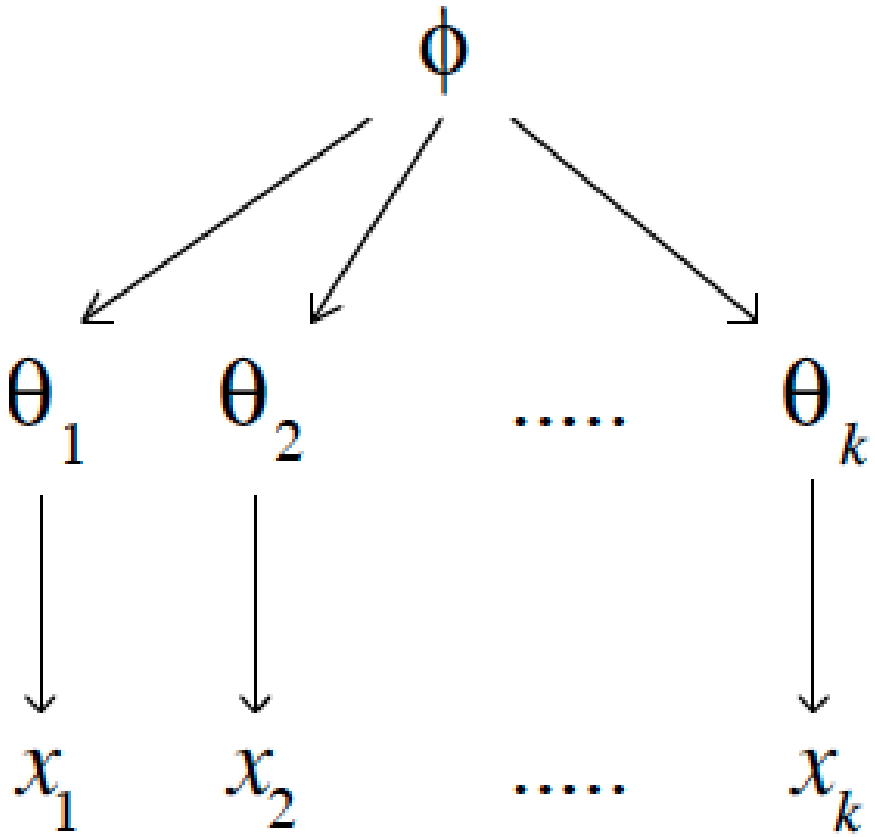
```
model.jags <- function()  {
      for (i in 1:(k+1)){
            beta[i] ~ dnorm(0, 0.001)
       }
       alpha ~ dgamma(0.01,0.01)
      for (i in 1:n){
            log(mu[i]) <-  X[i,]%*%beta
            y[i] ~ dgamma(alpha,alpha/mu[i])
       }
   }
```

Ejemplo de una regresión Gamma



Ver código: Programa20.r

Modelos Jerárquicos Lineales

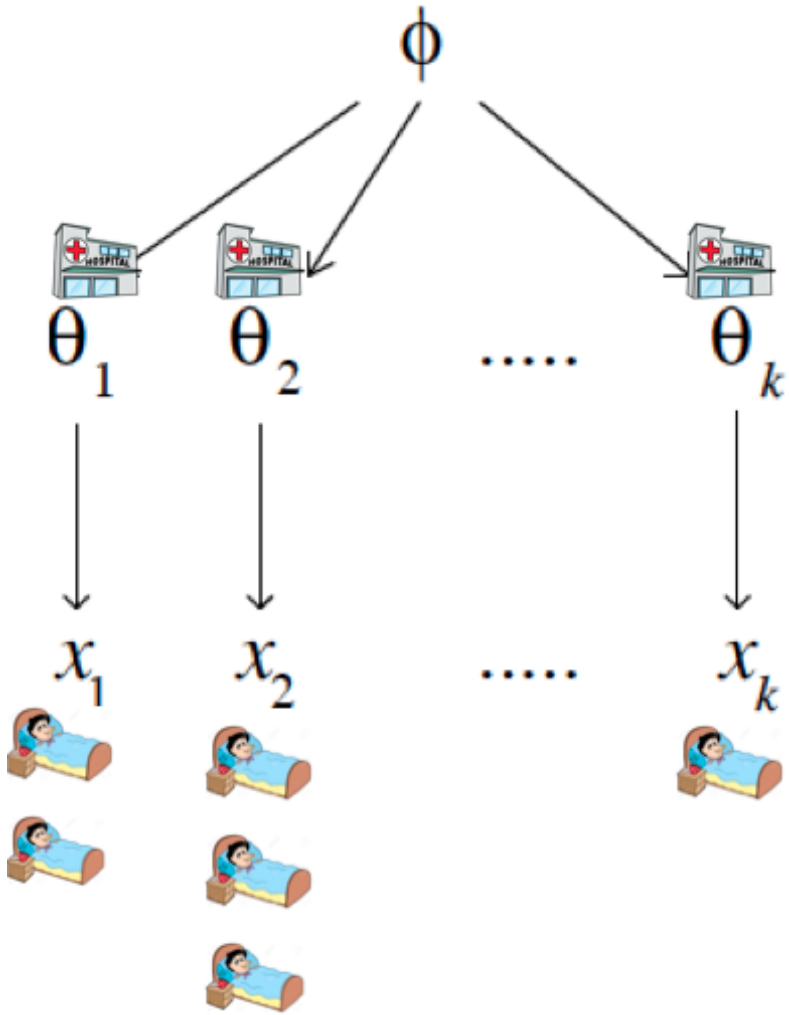


Modelos Jerárquicos Lineales: Motivación

Consideremos el siguiente problema:

- Se tiene una muestra de k hospitales
- Dentro de cada hospital, se tiene una muestra de n_i pacientes $(i = 1, \ldots, k)$.
- Para el Hospital i, los pacientes tiene una probabilidad de supervivencia (ante cierto padecimiento) de θ_i, de tal forma que condicionado a θ_i se tiene que $x_{ij} \sim Bernoulli(\theta_i)$.
- Si los hospitales fuera independientes, bastaría inferir el parámetro θ_i en cada hospital
- Si entre los hopitales suponemos **intercambiabilidad** (homogeneidad; por ejemplo hospitales de cierta Región del país), entonces se puede considerar que estos hospitales son una muestra de una distribución poblacional común que a su vez depende de **hiperparámetros** desconocidos ϕ.
- Con esta estructura **jerárquica**, suponiendo que se observan las muestras en cada hospital, resulta interesante entonces hacer inferencias sobre θ_i con $i = 1, \ldots, k$
- Si además a esta estructura se agregan covariables z, se pueden enriquecer el modelo.

Modelos Jerárquicos Lineales



Modelos Jerárquicos Lineales

Un Modelo Jerárquico tienen la siguiente estructura:

1. Nivel I, las observaciones:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\theta}) &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \\ &= \prod_{i=1}^{k} p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) = \prod_{i=1}^{k}\prod_{j=1}^{n_i} p(x_{ij}|\boldsymbol{\theta}_i) \end{aligned}$$

2. Nivel II, los parámetros:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\theta}) &= p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k|\boldsymbol{\phi}) \\ &= \prod_{i=1}^{k} p(\boldsymbol{\theta_i}|\boldsymbol{\phi}) \end{aligned}$$

3. Nivel III, los hiperparámetros:

$$p(\boldsymbol{\phi})$$

Modelos Jerárquicos Lineales

La interpretación del modelo puede ser la siguiente:

- Las observaciones $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, con $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in_i})$, provienen de experimentos distintos pero relacionados entre sí (Ej. Experimentos realizados en k centros de investigación).
- Los parámetros $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ se suponen relacionados (intercambiables, homogéneos) (Ej. θ_i puede representar la probabilidad de supervivencia en el centro de investigación i)
- Los parámetros $\boldsymbol{\phi}$ describen alguna característica relevante de la población (Ej. $g(\boldsymbol{\phi})$ con $g : \mathbb{R}^d \to \mathbb{R}$ puede representar la probabilidad de supervivencia **global** para toda la población de cierta región del país).
- En caso de existir información adicional, por ejemplo algunas característica del paciente como edad, peso, estatura, entonces los datos vienen dados por

$$\{(\boldsymbol{x}_1, \boldsymbol{z}_1), (\boldsymbol{x}_2, \boldsymbol{z}_2), \ldots, (\boldsymbol{x}_k, \boldsymbol{z}_k)\}$$

Modelos Jerárquicos Lineales

En el enfoque Bayesiano, estamos interesados en hacer inferencia sobre los parámetros,

$$(\boldsymbol{\theta}, \boldsymbol{\phi}) = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \phi_1, \ldots, \phi_d)$$

así como también para posibles futuras observaciones, digamos

$$x_{iF} \sim p\left(x|\boldsymbol{\theta}_i\right)$$

(una observación futura del i-ésimo centro de investigación).
Sin embargo, dada la estructura **jerárquica** también es plausible pensar en una observación futura x_F^* correspondiente a una futura θ_* que proviene de la misma población que generó a los parámetros θ_j existentes.

Modelos Jerárquicos Lineales

Dada la jerarquía que existe, es apropiado pensar en distribuciones iniciales de la siguiente forma:

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\theta}|\boldsymbol{\phi})\, p(\boldsymbol{\phi})$$

Por otro lado, la distribución final correspondiente es

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{x}) &= \frac{p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{x})}{p(\boldsymbol{x})} \\ &\propto p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\phi})\, p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}, \boldsymbol{\phi}) \end{aligned}$$

de donde las marginales correspondientes para hacer inferencias son:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &\propto \int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}, \boldsymbol{\phi})\, d\boldsymbol{\phi} \\ &\propto \int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{\phi})\, p(\boldsymbol{\phi})\, d\boldsymbol{\phi} \\ p(\boldsymbol{\phi}|\boldsymbol{x}) &\propto p(\boldsymbol{\phi}) \int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{\phi})\, d\boldsymbol{\theta} \end{aligned}$$

Modelos Jerárquicos Lineales

Existen casos donde se puede hacer inferencia de forma analítica, sin embargo generalmente tendremos que recurrir a aspectos computacionales para obtener aproximaciones a las densidades finales.

Ejemplo:

- Nivel I (Observaciones):

$$p\left(\boldsymbol{y}|\boldsymbol{\beta}\right) = N_n\left(\boldsymbol{y}|\boldsymbol{X\beta}, \boldsymbol{\Sigma}_y\right); \quad \boldsymbol{y} \in \mathbb{R}^n, \quad \boldsymbol{X} \in \mathbb{R}^{n\times p}; \quad \boldsymbol{\beta} \in \mathbb{R}^p; \quad \boldsymbol{\Sigma}_y \in \mathbb{R}^{n\times n}$$

- Nivel II (Parámetros):

$$p\left(\boldsymbol{\beta}|\boldsymbol{\alpha}\right) = N_p\left(\boldsymbol{\beta}|\boldsymbol{H\alpha}, \boldsymbol{\Sigma}_\beta\right); \quad \boldsymbol{\beta} \in \mathbb{R}^p; \quad \boldsymbol{H} \in \mathbb{R}^{p\times q}; \quad \boldsymbol{\alpha} \in \mathbb{R}^q; \quad \boldsymbol{\Sigma}_\beta \in \mathbb{R}^{p\times p}$$

- Nivel III (Hiperparámetros):

$$p\left(\boldsymbol{\alpha}\right) = N_q\left(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0, \boldsymbol{\Sigma}_\alpha\right); \quad \boldsymbol{\alpha} \in \mathbb{R}^p; \quad \boldsymbol{\alpha}_0 \in \mathbb{R}^q; \quad \boldsymbol{\Sigma}_\alpha \in \mathbb{R}^{q\times q}$$

En este modelo, suponiendo $\boldsymbol{\Sigma}_y$ y $\boldsymbol{\Sigma}_\beta$ conocidas, es posible demostrar que las finales toman la siguiente forma:

$$p\left(\boldsymbol{\alpha}|\boldsymbol{y}\right) = N_q\left(\boldsymbol{\alpha}\left|\boldsymbol{\mu}_{\alpha|y}, \boldsymbol{V}_{\alpha|y}\right.\right)$$

donde

$$\boldsymbol{\mu}_{\alpha|y} = \boldsymbol{\alpha}_0 + \boldsymbol{\Sigma}_\alpha \boldsymbol{H}^{'} \boldsymbol{X}^{'} \boldsymbol{V}_y^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{H}\boldsymbol{\alpha}_0\right)$$

y

$$\boldsymbol{V}_{\alpha|y} = \boldsymbol{\Sigma}_\alpha - \boldsymbol{\Sigma}_\alpha \boldsymbol{H}^{'} \boldsymbol{X}^{'} \boldsymbol{V}_y^{-1} \boldsymbol{X}\boldsymbol{H}\boldsymbol{\Sigma}_\alpha$$

con

$$\boldsymbol{V}_y = \boldsymbol{X}\left(\boldsymbol{\Sigma}_\beta + \boldsymbol{H}\boldsymbol{\Sigma}_\alpha \boldsymbol{H}^{'}\right)\boldsymbol{X}^{'} + \boldsymbol{\Sigma}_y$$

Mientras que

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = N_p\left(\boldsymbol{\beta}\left|\boldsymbol{\mu}_{\beta|y}, \boldsymbol{V}_{\beta|y}\right.\right)$$

donde:

$$\boldsymbol{\mu}_{\beta|y} = \boldsymbol{H}\boldsymbol{\mu}_{\alpha|y} + \boldsymbol{\Sigma}_\beta \boldsymbol{X}^{'} \boldsymbol{V}_y^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{H}\boldsymbol{\alpha}_0\right)$$

y

$$\boldsymbol{V}_{\beta|y} = \boldsymbol{V}_\beta - \boldsymbol{V}_\beta \boldsymbol{X}^{'} \boldsymbol{V}_y^{-1} \boldsymbol{X}\boldsymbol{V}_\beta$$

Ejemplo 5: (Simulación) Suponga el siguiente modelo jerárquico pare modelar tiempos de fallas de cierto componente de aviones de una compañía determinada.

- Nivel I (Observaciones) (9 grupos, con 20 observaciones cada uno)

$$p\left(\boldsymbol{x}_i|\alpha_i,\beta_i\right) = Ga\left(\boldsymbol{x}_i|\alpha_i,\beta_i\right); \quad i = 1,2,\ldots,9; \quad \boldsymbol{x}_i = (x_{i1},\ldots.,x_{i20})$$

- Nivel II (Parámetros):

$$p\left(\alpha_i,\beta_i\left|\alpha_1^{(0)},\beta_1^{(0)},\alpha_2^{(0)},\beta_2^{(0)}\right.\right) = Ga\left(\alpha_i\left|\alpha_1^{(0)},\beta_1^{(0)}\right.\right) Ga\left(\beta_i\left|\alpha_2^{(0)},\beta_2^{(0)}\right.\right);$$

- Nivel III (Hiperparámetros):

$$\begin{aligned} p\left(\alpha_1^{(0)},\beta_1^{(0)},\alpha_2^{(0)},\beta_2^{(0)}\middle|\boldsymbol{\theta}_0\right) &= Ga(\alpha_1^{(0)}\middle|0.01,0.01)\, Ga(\beta_1^{(0)}\middle|0.01,0.01) \\ &\quad Ga(\alpha_2^{(0)}\middle|0.01,0.01)\, Ga(\beta_2^{(0)}\middle|0.01,0.01) \end{aligned}$$

Obs: En este caso, colocar iniciales impropias no es adecuado pues genera distribuciones finales que no necesariamente son propias. Además, la estimación de los hiperparámetros no es muy precisa si se tienen pocos grupos.

El código de JAGS para este modelo es:

```
model.jags <- function()  {
  #Nivel 3; hiperparámetros (iniciales informativas)
  alpha.3[1] ~ dgamma(100,1)
  alpha.3[2] ~ dgamma(100,1)
  beta.3[1] ~ dgamma(1,1)
  beta.3[2] ~ dgamma(1,1)

  for (i in 1:k){
    #Nivel 2; Parametros
    alpha.2[i] ~ dgamma(alpha.3[1],beta.3[1])
    beta.2[i] ~ dgamma(alpha.3[2],beta.3[2])
    for(j in 1:(n[i])){
      #Nivel 1; Las Observaciones
      X[j,i] ~ dgamma(alpha.2[i],beta.2[i])
    }
  }
  #Nueva observacion del grupo 1
  x.1 ~ dgamma(alpha.2[1],beta.2[1])
  #Nueva observacion de un nuevo grupo
  alpha.2.n ~ dgamma(alpha.3[1],beta.3[1])
  beta.2.n ~ dgamma(alpha.3[2],beta.3[2])
  x.n ~ dgamma(alpha.2.n,beta.2.n)
}
```

Ver código: Programa21.r

Resultados : (Basados en 100,000 simulaciones, con un calentamiento de 10,000)

	Reales	Media	Mediana	L0.025	U0.975
alpha.2[1]	93.35	106.07	105.35	79.88	136.07
alpha.2[2]	113.21	117.56	116.83	88.71	150.02
alpha.2[3]	112.60	116.85	116.32	88.16	149.39
alpha.2[4]	103.68	108.39	107.77	82.04	137.28
alpha.2[5]	84.73	101.02	100.51	76.50	128.46
alpha.2[6]	104.42	103.60	103.22	78.62	131.20
alpha.2[7]	107.00	106.01	105.44	80.22	134.47
alpha.2[8]	105.33	101.73	101.16	77.20	129.09
alpha.2[9]	96.48	95.80	95.41	72.85	121.26
alpha.3[1]	100.00	100.16	99.88	81.85	120.16
alpha.3[2]	100.00	101.11	100.82	82.39	121.84
beta.2[1]	91.69	106.09	105.40	79.82	136.11
beta.2[2]	93.40	96.58	96.02	72.93	123.16
beta.2[3]	96.54	97.66	97.21	73.73	124.82
beta.2[4]	99.05	103.04	102.49	77.95	130.43
beta.2[5]	90.80	110.78	110.17	83.71	141.09
beta.2[6]	107.86	106.12	105.73	80.47	134.45
beta.2[7]	105.51	103.05	102.50	77.92	130.78
beta.2[8]	108.88	106.55	105.95	80.86	135.32
beta.2[9]	107.45	110.13	109.66	83.65	139.44
beta.3[1]	1.00	0.95	0.94	0.70	1.28
beta.3[2]	1.00	0.98	0.97	0.72	1.32

Ver código: Programa21.r

Ejemplo 6: Gelfand, Hills, Racine-Poon y Smith (1990) discuten el análisis bayesiano de la siguiente tabla:

Tabla 2: Crecimiento de un grupo de ratas (controles)

Rata	x_1	x_2	x_3	x_4	x_5	Rata	x_1	x_2	x_3	x_4	x_5
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

Corresponde al peso (en gramos) de 30 ratas jóvenes en un grupo de controles medido en distintos días (edad: $x_1 = 8, x_2 = 15, x_3 = 22, x_4 = 29, x_5 = 36$). Se supone un crecimiento lineal del peso respecto a la edad para cada una de las ratas.

Se tienen entonces 30 grupos, todos de tamaño 5. El modelo jerárquico que se plantea es el siguiente:

- Nivel I (Observaciones) (30 grupos, con 5 observaciones en cada uno)

$$p\left(\boldsymbol{y}|\boldsymbol{\beta_i},\tau_i\right)=N_5\left(\boldsymbol{X}\boldsymbol{\beta}_i,\tau_i\boldsymbol{I}\right);\ i=1,\ldots,30;\ \boldsymbol{X}=\begin{pmatrix}1 & 8\\ 1 & 15\\ 1 & 22\\ 1 & 29\\ 1 & 36\end{pmatrix}\in\mathbb{R}^{5\times 2};\ \boldsymbol{\beta}_i\in\mathbb{R}^{2\times 1}$$

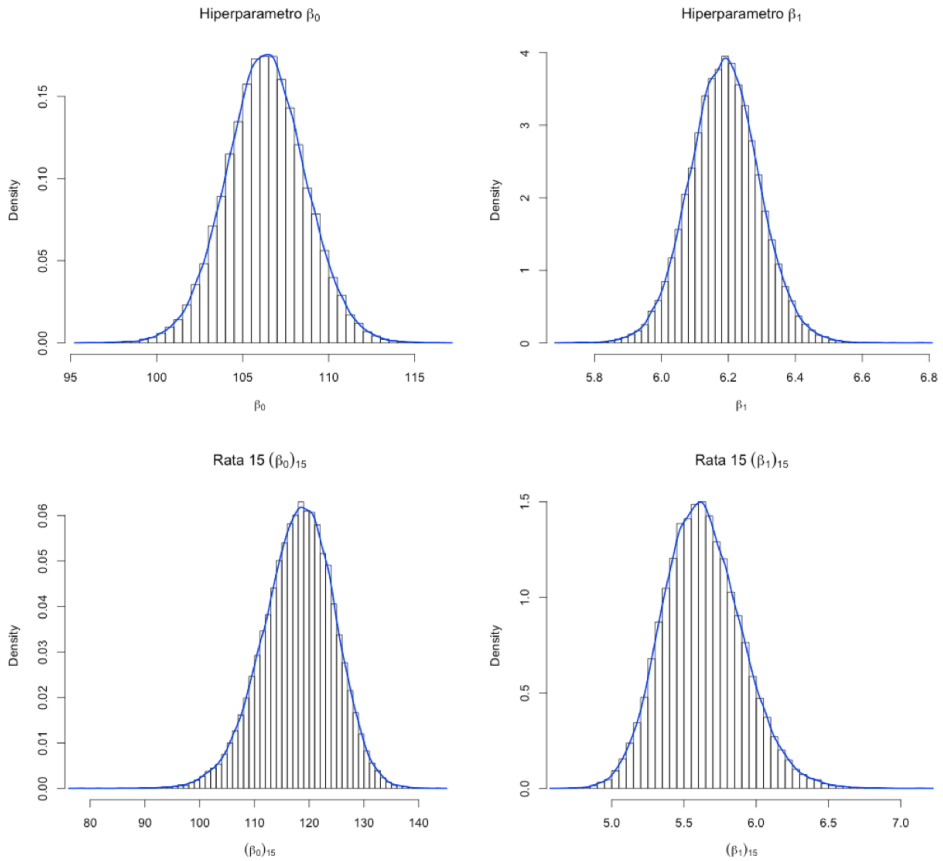
- Nivel II (Parámetros)

$$p\left(\boldsymbol{\beta_i},\tau_i\left|\boldsymbol{\alpha},\boldsymbol{\tau^{(0)}}\delta_1,\delta_2\right.\right)=N_2\left(\boldsymbol{\beta}_i\left|\boldsymbol{\alpha},\boldsymbol{\tau}^{(0)}\right.\right)Ga\left(\tau_i\left|\delta_1,\delta_2\right.\right);\ \boldsymbol{\tau^{(0)}}=\begin{pmatrix}\tau_1^{(0)} & 0\\ 0 & \tau_2^{(0)}\end{pmatrix}$$

- Nivel III (Hiperparámetros)

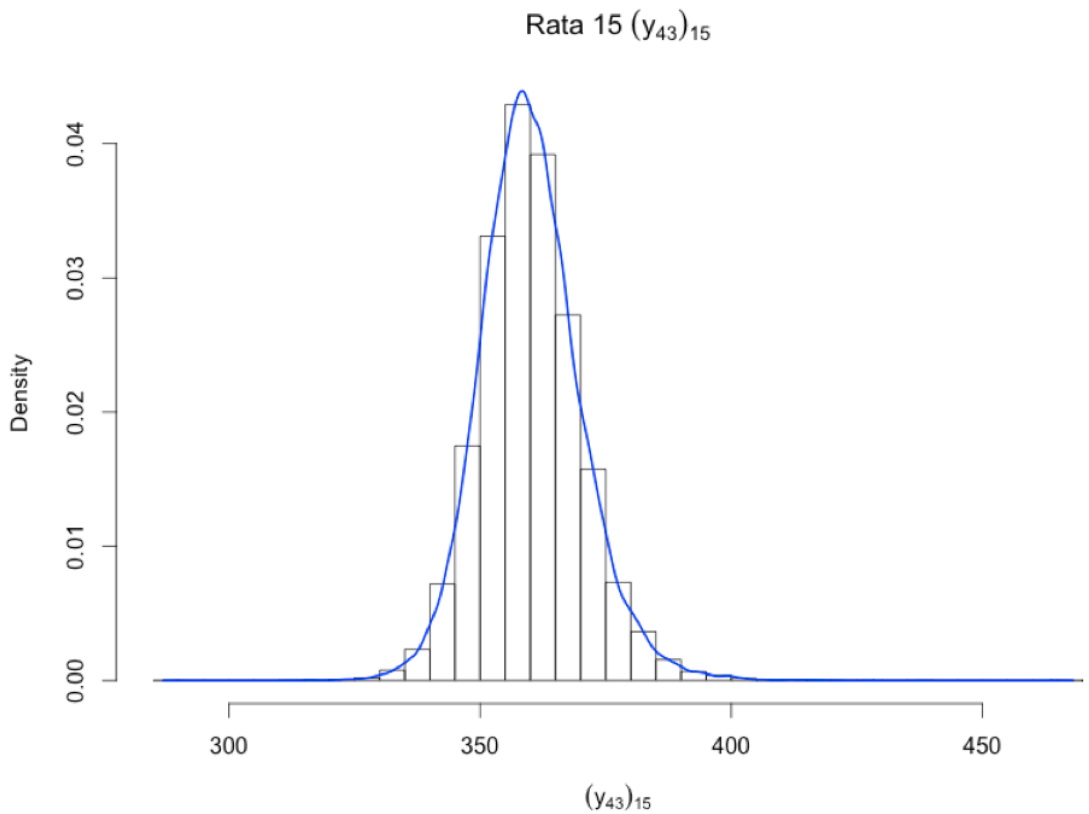
$$\begin{aligned}p\left(\boldsymbol{\alpha},\boldsymbol{\tau}^{(0)},\delta_1,\delta_2\left|\boldsymbol{\theta}_0\right.\right) &= N_2\left(\boldsymbol{\alpha}\left|\boldsymbol{0},0.001\boldsymbol{I}\right.\right)\\ &\quad Ga(\tau_1^{(0)}|0.01,0.01)Ga(\tau_2^{(0)}|0.01,0.01)\\ &\quad Ga(\delta_1|0.01,0.01)Ga(\delta_2|0.01,0.01)\end{aligned}$$

En la siguiente figura se presentan las distribuciones finales, tanto para los coeficientes poblacionales como para los coeficientes correspondientes a la Rata 15. (Los resultados son muy parecidos a los presentados por Gelfand, Hills, Racine-Poon y Smith (1990)).



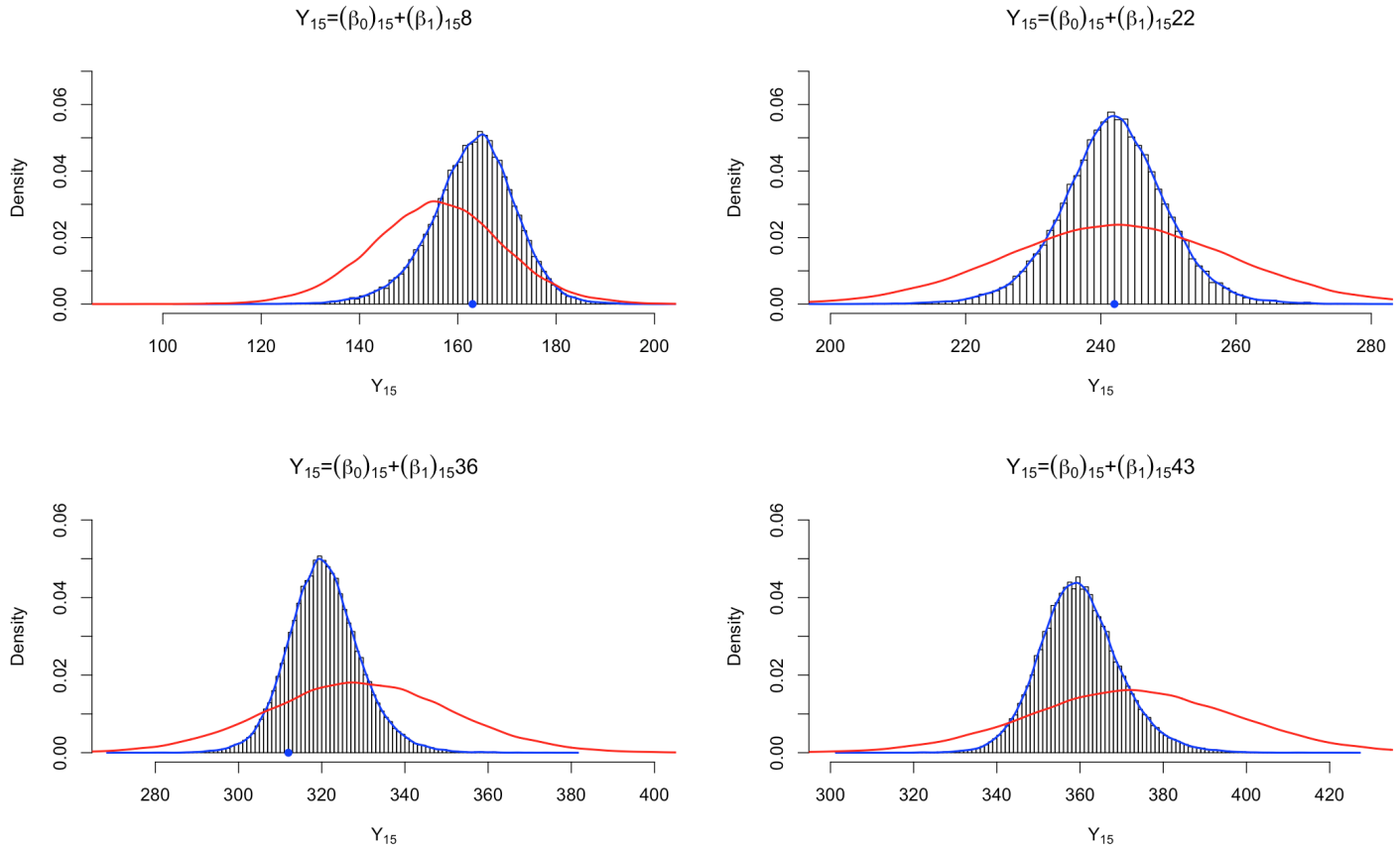
Ver código: Programa22.r

En la siguiente figura se presenta la distribución predictiva final para el peso de la Rata 15 a los 43 días de edad. El intervalo de credibilidad al 95 % obtenido vía simulación es $(340.7595, 380.4825)$.



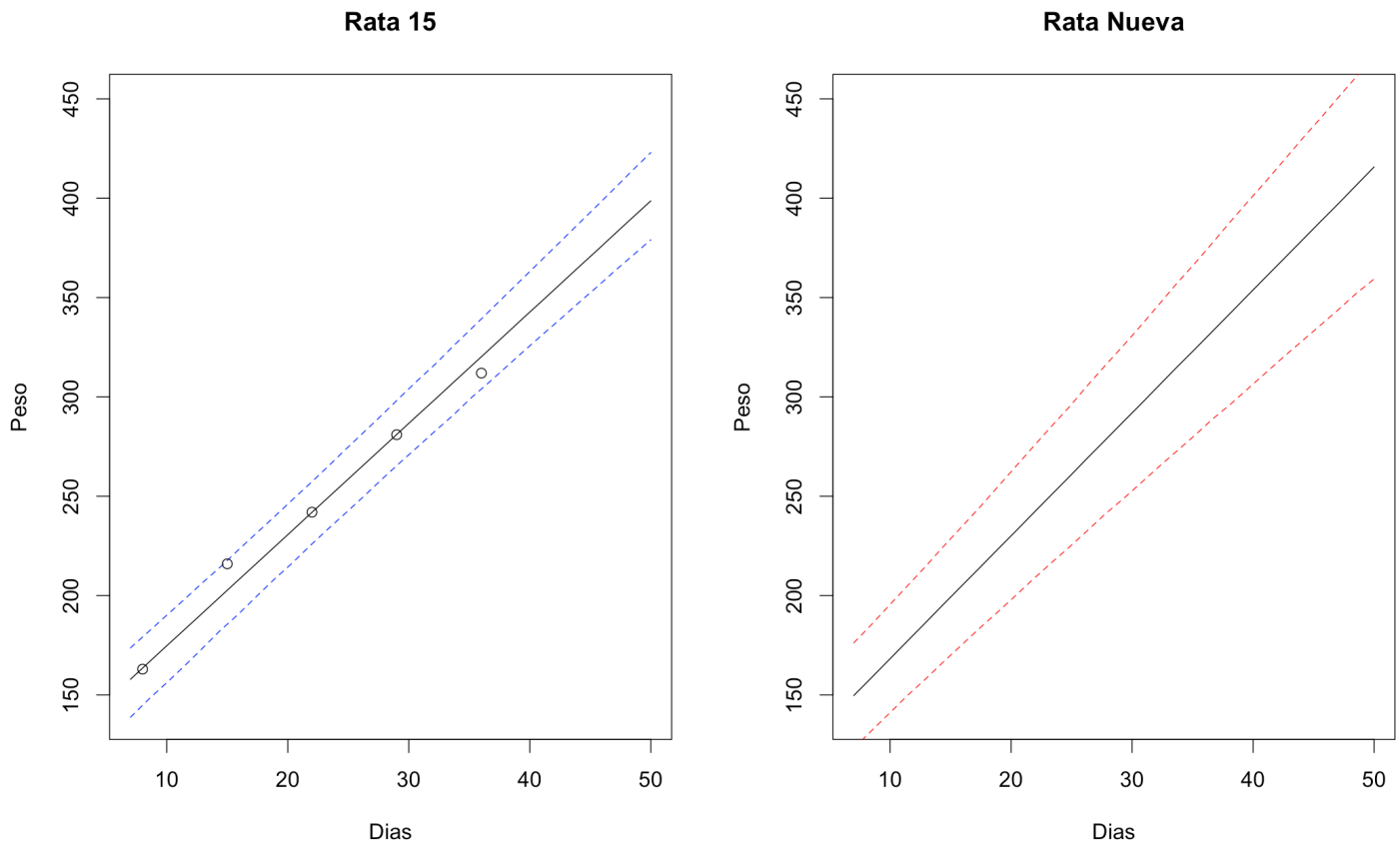
Ver código: Programa22.r

En la siguiente figura se presentan las distribuciones predictivas del peso de la Rata 15 para los días $(8, 22, 36, 43)$. Por otro lado, la línea roja presenta la predicción en esos días para el peso de una rata nueva.



Ver código: Programa22.r

Se presenta el modelo lineal ajustado con las bandas de predicción al 95 % para la Rata 15 (líneas azules). Las líneas rojas del segundo gráfico representan la banda de predicción (95 %) para el peso de una rata nueva.



Ver código: Programa22.r

Comentarios finales

Comentarios finales

¡Muchas gracias por su atención!