Estadística Bayesiana
Parte II

Irving Gómez Méndez

UNIVERSIDAD IBEROAMERICANA
CIUDAD DE MÉXICO ®

Regresión Bayesiana

Regresión

El modelamiento matemático se refiere a construir ecpresiones matemáticas que describan el comportamiento de una variable de interés Y. Frecuentemente deseamos agregar al modelo algunas variables (caractetísticas) X, que poseen información sobre la variable de interés Y.

En el análisis de regresión buscamos una función $f : \mathcal{X} \to \mathcal{Y}$ tal que $f(X)$ se una buena aproximación de Y.

Regresión Lineal Bayesiana

Considere el modelo usual de regresión

$$Y = X^T\beta + \varepsilon,$$

con $\varepsilon | X \sim \mathcal{N}(0, \sigma^2)$ y $\varepsilon \perp\!\!\!\perp \beta | X$.
Además suponga que contamos con una muestra $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ con $(X_i, Y_i) \overset{iid}{\sim} (X, Y)$. Así, el modelo puede ser escrito como

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n).$$

Regresión (Modelo Normal - Inversa Gama)

Los parámetros de este modelo son β y σ. Considere las siguientes distribuciones previas

$$\sigma^2 \sim \textsf{Inversa-Gama}(a, b), \quad a, b > 0,$$

$$\beta|\sigma^2 \sim \mathcal{N}_p(\mu_\beta, \sigma^2 V_\beta),$$

$\mu_\beta \in \mathbb{R}^p$ y V_β matriz simétrica y positiva definida.

En este caso se dice que la distribución conjunta previa para β y σ^2 es una Normal-Inversa Gama de parámetros μ_β, V_β, a y b, y se denota como

$$\beta, \sigma^2 \sim NIG(\mu_\beta, V_\beta, a, b).$$

Distribuciones Previas

Puede demostrarse que bajo este modelo

$$
\begin{aligned}
\sigma^2 &\sim \textsf{Inversa-Gama}(a, b) \\
\beta|\sigma^2 &\sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta) \\
\sigma^2|\beta &\sim \textsf{Inversa-Gama}\left(a + \frac{p}{2}, b + \frac{1}{2}(\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta)\right) \\
\beta, \sigma^2 &\sim NIG(\mu_\beta, V_\beta, a, b) \\
\beta &\sim t_{2a}\left(\mu_\beta, \frac{b}{a} V_\beta\right)
\end{aligned}
$$

Distribuciones Posteriores

$$\begin{aligned}
\sigma^2|\mathcal{D}_n &\sim \textsf{Inversa-Gama}(a^\star, b^\star)\\
\beta|\sigma^2,\mathcal{D}_n &\sim \mathcal{N}(\mu^\star, \sigma^2 V^\star)\\
\sigma^2|\beta,\mathcal{D}_n &\sim \textsf{Inversa-Gama}\left(a^\star + \frac{p}{2}, b^\star + \frac{1}{2}(\beta-\mu^\star)^T V^{\star -1}(\beta-\mu^\star)\right)\\
\beta,\sigma^2|\mathcal{D}_n &\sim NIG(\mu^\star, V^\star, a^\star, b^\star)\\
\beta|\mathcal{D}_n &\sim t_{2a^\star}\left(\mu^\star, \frac{b^\star}{a^\star}V^\star\right)
\end{aligned}$$

Distribucion Predictivas

Previa

$$\mathbf{Y}|\mathbf{X}_0, \beta, \sigma^2 \sim \mathcal{N}_n(\mathbf{X}_0\beta, \sigma^2 I)$$

$$\mathbf{Y}|\mathbf{X}_0, \sigma^2 \sim \mathcal{N}_n\left(\mathbf{X}_0\mu_\beta, \sigma^2(I + \mathbf{X}_0 V_\beta \mathbf{X}_0^T)\right)$$

$$\mathbf{Y}|\mathbf{X}_0 \sim t_{2a}\left(\mathbf{X}_0\mu_\beta, \frac{b}{a}(I + \mathbf{X}_0 V_\beta \mathbf{X}_0^T)\right)$$

Posterior

$$\mathbf{Y}|\mathbf{X}_0, \beta, \sigma^2, \mathcal{D}_n \sim \mathcal{N}_n(\mathbf{X}_0\beta, \sigma^2 I)$$

$$\mathbf{Y}|\mathbf{X}_0, \sigma^2, \mathcal{D}_n \sim \mathcal{N}_n\left(\mathbf{X}_0\mu^\star, \sigma^2(I + \mathbf{X}_0 V^\star \mathbf{X}_0^T)\right)$$

$$\mathbf{Y}|\mathbf{X}_0, \mathcal{D}_n \sim t_{2a^\star}\left(\mathbf{X}_0\mu^\star, \frac{b^\star}{a^\star}(I + \mathbf{X}_0 V^\star \mathbf{X}_0^T)\right)$$

Donde

$$
\begin{aligned}
V^{\star} &= \left(V_{\beta}^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1} \\
\mu^{\star} &= V^{\star}\left(V_{\beta}^{-1}\mu_{\beta} + \mathbf{X}^T\mathbf{Y}\right) \\
a^{\star} &= a + \frac{n}{2} \\
b^{\star} &= b + \frac{1}{2}\left[\mu_{\beta}^T V_{\beta}^{-1}\mu_{\beta} + \mathbf{Y}^T\mathbf{Y} - \mu^{\star T} V^{\star -1}\mu^{\star}\right]
\end{aligned}
$$

Tarea 3

- Mostrar ejemplo de regresión.
- Dejar tarea 3.

Análisis de Referencia

A veces se buscan distribuciones previas que tengan poco impacto en la distribución posterior. Dichas distribuciones son llamadas distribuciones previas de referencia, las cuales son descritas como vagas, constantes, difusas o no informativas.

Por ejemplo, en el caso de la distribución Normal con media θ y varianza conocida σ^2, al considerar una previa normal con media μ_0 y varianza τ_0^2, obtuvimos:

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

donde

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{Y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{y} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

Distribuciones impropias

Si hacemos $\tau_0 \to \infty$, entonces

$$\theta|\mathbf{Y} \sim \mathsf{Normal}\left(\bar{Y}, \frac{\sigma^2}{n}\right),$$

note que si $\tau_0 \to \infty$, entonces $p(\theta) \propto \mathbb{1}_{\mathbb{R}}(\theta)$. Lo que va en orden con el principio de razón insuficiente.

Note que la función $\mathbb{1}_{\mathbb{R}}(\theta)$ no posee integral finita, y por lo tanto no hay manera de normalizarla para que integre 1. Por lo tanto, no hay manera de obtener una densidad y no determina una distribución en sentido estricto. A este tipo de "densidades" que no poseen integral finita se les conoce como densidades impropias.

Retomando el caso binomial, habíamos propuesto $\theta \sim \mathsf{Uniforme}(0,1)$. Pero suponga que estamos interesados en $\phi = -\log\theta$, si $\theta \sim \mathsf{Uniforme}(0,1)$, entonces $\phi \sim \mathsf{Exponencial}(1)$. Lo que viola el principio de razón insuficiente.

Esta ambigüedad en la que no está claro qué debe ser uniforme puede conducir a importantes contradicciones.

- Mostrar paradoja de Bertrand.

Función score

Suponga que $Y \sim p(Y|\theta_0)$, definimos la función score como:

$$sc(\theta) = \frac{d}{d\theta} \log p(Y|\theta).$$

$$\begin{aligned}
\mathbb{E}_{Y|\theta_0}[sc(\theta)] &= \int_{\mathcal{Y}} \left[\frac{d}{d\theta} \log p(Y|\theta) \right] p(Y|\theta_0) dY \\
&= \int_{\mathcal{Y}} \frac{p(Y|\theta_0)}{p(Y|\theta)} \frac{d}{d\theta} p(Y|\theta) dY.
\end{aligned}$$

Entonces

$$\mathbb{E}_{Y|\theta_0}[sc(\theta_0)] = \int_{\mathcal{Y}} \frac{d}{d\theta} p(Y|\theta) \Big|_{\theta=\theta_0} dY$$

Condiciones de regularidad

Si se pueden intercambiar las operaciones de integración y derivación, entonces

$$\begin{aligned}\mathbb{E}_{Y|\theta_0}[sc(\theta_0)] &= \frac{d}{d\theta}\left[\int_{\mathcal{Y}} p(Y|\theta)dY\right]_{\theta=\theta_0} \\ &= \left.\frac{d}{d\theta}1\right|_{\theta=\theta_0} \\ &= 0.\end{aligned}$$

Información (esperada) de Fisher por unidad muestral

Por otro lado, definimos la información esperada de Fisher por unidad muestral como:

$$\begin{aligned}\mathcal{I}_{\theta_0}(\theta) &= \mathbb{E}_{Y|\theta_0}[sc^2(\theta)] \\ &= \mathbb{E}_{Y|\theta_0}\left[\frac{1}{p^2(Y|\theta)}\left(\frac{d}{d\theta}p(Y|\theta)\right)^2\right].\end{aligned}$$

Note que

$$\begin{aligned}\frac{d^2}{d\theta^2}\log p(Y|\theta) &= \frac{d}{d\theta}\left[\frac{1}{p(Y|\theta)}\frac{d}{d\theta}p(Y|\theta)\right] \\ &= -\frac{1}{p^2(Y|\theta)}\left(\frac{d}{d\theta}p(Y|\theta)\right)^2 + \frac{1}{p(Y|\theta)}\frac{d^2}{d\theta^2}p(Y|\theta).\end{aligned}$$

Entonces,

$$-\mathbb{E}_{Y|\theta_0}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\right] = \mathcal{I}_{\theta_0}(\theta) - \mathbb{E}_{Y|\theta_0}\left[\frac{1}{p(Y|\theta)}\frac{d^2}{d\theta^2}p(Y|\theta)\right].$$

Al evaluar en θ_0 y suponiendo que se pueden intercambiar las operaciones de integración y derivación, obtenemos que

$$\mathcal{I}_{\theta_0}(\theta_0) = -\mathbb{E}_{Y|\theta_0}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\bigg|_{\theta=\theta_0}\right].$$

Por lo tanto, bajo condiciones de regularidad

$$\mathbb{E}_{Y|\theta_0}[sc(\theta_0)] = 0, \quad \mathbb{V}_{Y|\theta_0}[sc(\theta_0)] = \mathcal{I}_{\theta_0}(\theta_0)$$

y

$$J(\theta_0) \equiv \mathcal{I}_{\theta_0}(\theta_0) = -\mathbb{E}_{Y|\theta_0}\left[\frac{d^2}{d\theta^2} \log p(Y|\theta)\bigg|_{\theta=\theta_0} \right].$$

- Mostrar ejemplo de caso exponencial.

Ejercicios

1. Sea $Y|\theta \sim \mathsf{Exponencial}(\theta)$, $p(Y|\theta) = \theta e^{-\theta y}\mathbb{1}_{(0,\infty)}(y)$. Demuestre que la información esperada de Fisher está dada por
$$J(\theta_0) = \frac{1}{\theta_0^2}.$$
2. Sea $Y|\theta \sim \mathsf{Poisson}(\theta)$. Demuestre que la información esperada de Fisher está dada por
$$J(\theta_0) = \frac{1}{\theta_0}.$$
3. Sea $Y|\theta \sim \mathsf{Binomial}(n, \theta)$. Demuestre que la información esperada de Fisher está dada por
$$J(\theta_0) = \frac{n}{\theta_0(1-\theta_0)}.$$

Regla de Jeffreys

La distribución no informativa de Jeffreys está dada por

$$p(\theta) \propto \sqrt{J(\theta)}.$$

La idea detrás de esta definición es que cualquier regla para determinar la densidad previa $p(\theta)$ debería de generar un resultado equivalente al ser aplicada a un parámetro transformado; es decir $p(\phi)$ calculado a partir de $p(\theta)$ y el teorema de cambio de variable debería dar el mismo resultado que calculándolo directamente a partir de la información esperada de Fisher para ϕ.

Si $p(\theta) \propto \sqrt{J(\theta)}$ y $\phi = \phi(\theta)$ es una transformación 1-1 de θ. Entonces $p(\phi) \propto \sqrt{J(\phi)}$.

Demostración

Usando regla de la cadena, se tiene que

$$\frac{d}{d\phi} \log p(Y|\phi) = \frac{d}{d\theta} \log p(Y|\phi) \frac{d\theta}{d\phi}$$

y

$$\frac{d^2}{d\phi^2} \log p(Y|\phi) = \frac{d^2}{d\theta^2} \log p(Y|\phi) \left(\frac{d\theta}{d\phi}\right)^2 + \frac{d}{d\theta} \log p(Y|\phi) \frac{d^2\theta}{d\phi^2}$$

Multiplicando por -1 y tomando el valor esperado:

$$J(\phi) = J(\theta)\left(\frac{d\theta}{d\phi}\right)^2 - \underbrace{\mathbb{E}_{Y|\theta}\left[\frac{d}{d\theta}\log p(Y|\theta)\right]}_{0}\frac{d^2\theta}{d\phi^2}.$$

Luego

$$\sqrt{J(\phi)} = \sqrt{J(\theta)}\left|\frac{d\theta}{d\phi}\right|,$$

es decir

$$p(\phi) \propto p(\theta)\left|\frac{d\theta}{d\phi}\right|.$$

■

Cota de Cramér-Rao

Sean $Y_1, \ldots, Y_n$ variables aleatorias con función de densidad conjunta $p(\mathbf{Y}|\theta)$, y sea $T(\mathbf{Y})$ cualquier función de tal que $\mathbb{E}_{\mathbf{Y}|\theta}[T(\mathbf{Y})]$ sea una función diferenciable en θ.
Además, defina la función score (de la muestra) como

$$
\begin{aligned}
sc_n(\theta) &= \frac{d}{d\theta} \log p(\mathbf{Y}|\theta) \\
&= \frac{1}{p(\mathbf{Y}|\theta)} \frac{d}{d\theta} p(\mathbf{Y}|\theta),
\end{aligned}
$$

demostramos que, bajo condiciones de regularidad

$$\mathbb{E}_{\mathbf{Y}|\theta}[sc_n(\theta)] = 0.$$

Entonces, bajo dichas condiciones, se satisface que

$$\begin{aligned}
\mathrm{Cov}_{\mathbf{Y}|\theta}(T(\mathbf{Y}), sc_n(\theta)) &= \mathbb{E}_{\mathbf{Y}|\theta}[T(\mathbf{Y})sc_n(\theta)] \\
&= \mathbb{E}_{\mathbf{Y}|\theta}\left[T(\mathbf{Y})\frac{1}{p(\mathbf{Y}|\theta)}\frac{d}{d\theta}p(\mathbf{Y}|\theta)\right] \\
&= \int_{\mathcal{Y}^n} T(\mathbf{y})\frac{d}{d\theta}p(\mathbf{y}|\theta)d\mathbf{y} \\
&= \frac{d}{d\theta}\int_{\mathcal{Y}^n} T(\mathbf{y})p(\mathbf{y}|\theta)d\mathbf{y} \\
&= \frac{d}{d\theta}\mathbb{E}_{\mathbf{Y}|\theta}(T(\mathbf{Y})).
\end{aligned}$$

Por otro lado, usando la desigualdad de Cauchy-Schwartz

$$\begin{aligned}
\mathbb{V}_{\mathbf{Y}|\theta}(T(\mathbf{Y}))\mathbb{V}_{\mathbf{Y}|\theta}(sc_n(\theta)) &\geq \left(\mathrm{Cov}_{\mathbf{Y}|\theta}(T(\mathbf{Y}), sc_n(\theta))\right)^2 \\
&= \left(\frac{d}{d\theta}\mathbb{E}_{\mathbf{Y}|\theta}(T(\mathbf{Y}))\right)^2
\end{aligned}$$

$$\Rightarrow \mathbb{V}_{\mathbf{Y}|\theta}(T(\mathbf{Y})) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_{\mathbf{Y}|\theta}(T(\mathbf{Y}))\right)^2}{\mathbb{V}_{\mathbf{Y}|\theta}(sc_n(\theta))}.$$

Relación entre la varianza de un estimador y la información de Fisher

Si, $Y_1, \ldots, Y_n$ son v.a. i.i.d. con función de densidad $p(Y|\theta)$, entonces

$$\mathbb{V}_{\mathbf{Y}|\theta}(T(\mathbf{Y})) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_{\mathbf{Y}|\theta}(T(\mathbf{Y}))\right)^2}{nJ(\theta)}.$$

Más aún, si $T(\mathbf{Y})$ es un estimador insesgado para θ, i.e. $\mathbb{E}_{\mathbf{Y}|\theta}(T(\mathbf{Y})) = \theta$, entonces

$$\mathbb{V}_{\mathbf{Y}|\theta}(T(\mathbf{Y})) \geq \frac{1}{nJ(\theta)}.$$

Note que esta desigualdad formaliza nuestro pensamiento intuitivo. Pues para cualquier valor de θ que minimice la información esperada de Fisher, tendrá como consecuencia que la mínima varianza de cualquier estimador será grande. Y viceversa, cualquier valor de θ que maximice la información esperada de Fisher, tendrá como consecuencia que la mínima varianza de cualquier estimador será pequeña.

Por otro lado, esto también indica que no todos los valores de θ poseen la misma cantidad de información. De esta manera, entendiendo la distribución previa como una manera de codificar la información que aporta cada valor de θ a priori, es que hace sentido tomarla proporcional a la información esperada de Fisher.

Por ejemplo, demostramos que para el caso Binomial $(Y|\theta \sim \mathsf{Binomial}(n, \theta))$

$$J(\theta) = \frac{n}{\theta(1-\theta)},$$

note que $\theta \to 0$ o $\theta \to 1$ corresponden a los casos más informativos y es cuando $J(\theta) \to \infty$. Mientras que $J(\theta)$ es mínima cuando $\theta = 0.5$.

Relevancia histórica de la regla de Jeffreys

El problema de la invarianza ante transformaciones monótonas del principio de razón de insuficiencia de Laplace fue una de las mayores críticas a la estadística bayesiana a inicios del siglo XX, realizada, entre otros, por Ronald A. Fisher. Sin embargo, los estudios y aportaciones de Harold Jeffreys sobre previas no informativas que fueran invariantes ante transformaciones volvieron a traer interés sobre el tema.

Relación de la previa de Jeffreys y la divergencia KL

Para el caso uniparamétrico, puede demostrarse que la regla de Jeffreys maximiza la divergencia de Kullback-Leibler. Aunque es posible que Jeffreys ignorara esta afirmación, puede ser que sí tuviera cierta idea de la existencia de una relación entre la regla que propuso y la divergencia de Kullback-Leibler. Sin embargo, cuando existen más parámetros, la regla de Jeffreys no parece ser una buena alternativa (algo que ya había notado el propio Jeffreys).

Análisis de referencia con pivotales

Si $U = Y - \theta$ es una variable aleatoria cuya distribución no depende de θ ni de Y, entonces U es una cantidad pivotal y θ es llamado un parámetro de localización, $\theta \in \mathbb{R}$.

Note que

$$\frac{dY}{dU} = 1 \quad \text{y} \quad \left|\frac{d\theta}{dU}\right| = |-1| = 1,$$

luego

$$p(U) = p(Y|\theta)\left|\frac{dY}{dU}\right| = p(Y|\theta)$$

y

$$p(U) = p(\theta|Y)\left|\frac{d\theta}{dU}\right| = p(\theta|Y).$$

Por lo tanto,

$$p(\theta|Y) = p(Y|\theta)$$

y por lo tanto $p(\theta) \propto \mathbb{1}_{\mathbb{R}}(\theta)$.

Si $U = \frac{Y}{\theta}$ es una variable aletoria cuya distribución no depende de θ ni de Y, entonces U es una cantidad pivotal y θ es llamado un parámetro de escala, $\theta > 0$.
Note que

$$\frac{dY}{dU} = \theta$$

y

$$\frac{d\theta}{dU} = -\frac{Y}{U^2} = -\frac{Y}{Y^2}\theta^2 = -\frac{\theta^2}{Y},$$

luego

$$p(U) = p(Y|\theta)\left|\frac{dY}{dU}\right| = \theta p(Y|\theta)$$

y

$$p(U) = p(\theta|Y)\left|\frac{d\theta}{dU}\right| = \frac{\theta^2}{|y|}p(\theta|Y)$$

Por lo tanto,

$$\frac{\theta^2}{|y|}p(\theta|Y) = \theta p(Y|\theta);$$

entonces

$$p(\theta|Y) = \frac{1}{\theta}|y|p(Y|\theta),$$

es decir

$$p(\theta) \propto \frac{1}{\theta}\mathbb{1}_{(0,\infty)}(\theta).$$

Ejercicio

Demuestre que si θ es un parámetro de escala y $p(\theta) \propto \frac{1}{\theta}\mathbb{1}_{(0,\infty)}(\theta)$, entonces

$$p(\theta^2) \propto \frac{1}{\theta^2}\mathbb{1}_{(0,\infty)}(\theta^2)$$

y

$$p(\log\theta) \propto \mathbb{1}_{\mathbb{R}}(\log\theta)$$

Modelo Uniforme

Sea $Y|a,b \sim \mathsf{Uniforme}(a,b)$,

$$p(Y|a,b) = \frac{1}{b-a}\mathbb{1}_{(a,b)}(Y).$$

Considere la transformación dada por

$$U = \frac{Y-a}{b-a} \Rightarrow Y = (b-a)U + a$$

y

$$\frac{dY}{dU} = b-a,$$

luego

$$p(U|a,b) = \mathbb{1}_{(0,1)}(U).$$

Es decir $U \sim \mathsf{Uniforme}(0,1)$, como la distribución de U no depende de a, b ni Y, entonces U es una cantidad pivotal, a es parámetro de localización y $b-a$ es parámetro de escala. Así, podemos proponer la previa no informativa:

$$p(a, b-a) \propto \frac{1}{b-a} \mathbb{1}_{(0,\infty)}(b-a) \mathbb{1}_{\mathbb{R}}(a).$$

Considerando la reparametrización $\phi(a, b-a) = (a, b)$, obtenemos la previa no informativa para los parámetros a y b dada por

$$p(a, b) \propto \frac{1}{b-a} \mathbb{1}_{(a,\infty)}(b) \mathbb{1}_{\mathbb{R}}(a).$$

La verosimilitud puede ser escrita como

$$p(Y|a,b) = \frac{1}{b-a}\mathbb{1}_{(-\infty,y)}(a)\mathbb{1}_{(y,\infty)}(b)$$

y la verosimilitud de una muestra observada estaría dada por

$$p(\mathbf{Y}|a,b) = \frac{1}{(b-a)^n}\mathbb{1}_{(-\infty,y_{(1)})}(a)\mathbb{1}_{(y_{(n)},\infty)}(b).$$

Luego, la distribución posterior está dada por

$$p(a,b|\mathbf{Y}) \propto \frac{1}{(b-a)^{n+1}}\mathbb{1}_{(-\infty,y_{(1)})}(a)\mathbb{1}_{(y_{(n)},\infty)}(b).$$

Al integrar se puede calcular la constante de proporcionalidad y demostrar que la densidad posterior es

$$p(a,b|\mathbf{Y}) = n(n-1)\frac{\left(y_{(n)} - y_{(1)}\right)^{n-1}}{(b-a)^{n+1}}\mathbb{1}_{(-\infty,y_{(1)})}(a)\mathbb{1}_{(y_{(n)},\infty)}(b)$$

- Mostrar modelo Uniforme con previa no informativa.

Aproximación Normal

Suponga que $Y_1, \ldots, Y_n \overset{iid}{\sim} f(Y)$, pero que nosotros modelamos como $p(Y|\theta)$. Además, $p(\theta)$ es la distribución previa de nuestro modelo. Entonces

$$p(\mathbf{Y}|\theta) = \prod_{i=1}^{n} p(Y_i|\theta)$$

será la verosimilitud de la muestra observada.
Sea θ_0 el valor que minimiza la divergencia de Kullback-Leibler entre $f(Y)$ y $p(Y|\theta)$,

$$\begin{aligned} KL(\theta) &= \mathbb{E}_{Y\sim f}\left[\log\left(\frac{f(Y)}{p(Y|\theta)}\right)\right] \\ &= \int_{\mathcal{Y}} \log\left(\frac{f(Y)}{p(Y|\theta)}\right) f(Y)dY \end{aligned}$$

Información observada de Fisher en el contexto bayesiano

Se puede demostrar que, bajo condiciones de regularidad, para n suficientemente grande

$$\theta|\mathbf{Y} \dot{\sim} \mathcal{N}\left(\theta_0, (nJ(\theta_0))^{-1}\right).$$

Más aún, si denotamos por $\hat{\theta}$ la moda de la distribución posterior, bajo las mismas condiciones de regularidad, para n suficientemente grande

$$\hat{\theta} \approx \theta_0.$$

Además, definimos la información observada de Fisher, en el contexto bayesiano y bajo condiciones de regularidad, como

$$I(\hat{\theta}) = -\left.\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\right|_{\theta=\hat{\theta}}.$$

Note que podemos reescribir esta cantidad como

$$
\begin{aligned}
I(\hat{\theta}) &= - \left.\frac{d^2}{d\theta^2} \log p(\theta|\mathbf{Y})\right|_{\theta=\hat{\theta}} \\
&= -\frac{d^2}{d\theta^2}\Big[\log p(\mathbf{Y}|\theta) + \log p(\theta) - \log p(\mathbf{Y})\Big]_{\theta=\hat{\theta}} \\
&= - \left.\frac{d^2}{d\theta^2} \log p(\mathbf{Y}|\theta)\right|_{\theta=\hat{\theta}} - \left.\frac{d^2}{d\theta^2} \log p(\theta)\right|_{\theta=\hat{\theta}}
\end{aligned}
$$

Como hemos supuesto que $Y_1, \ldots Y_n$ son iid, entonces

$$
= - \sum_{i=1}^{n} \left.\frac{d^2}{d\theta^2} \log p(Y_i|\theta)\right|_{\theta=\hat{\theta}} - \left.\frac{d^2}{d\theta^2} \log p(\theta)\right|_{\theta=\hat{\theta}}
$$

Es decir,

$$I(\hat{\theta}) = -n \left[\frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log p(Y_i|\theta) \right]_{\theta=\hat{\theta}} - \left. \frac{d^2}{d\theta^2} \log p(\theta) \right|_{\theta=\hat{\theta}}.$$

Ahora, por ley de grandes números y bajo condiciones de regularidad, se satisface que para n suficientemente grande

$$-\frac{1}{n} \sum_{i=1}^{n} \left. \frac{d^2}{d\theta^2} \log p(Y_i|\theta) \right|_{\theta=\hat{\theta}} \approx J(\theta_0),$$

además el segundo término del lado izquierdo de la primera igualdad se vuelve despreciable.

Por lo tanto, bajo condiciones de regularidad, para n suficientemente grande

$$I(\hat{\theta}) \approx nJ(\theta_0)$$

y

$$\theta|\mathbf{Y} \dot{\sim} \mathcal{N}\left(\hat{\theta}, \left(I(\hat{\theta})\right)^{-1}\right).$$

Región de $(1-\alpha)$ de probabilidad posterior

Por otro lado, podemos tomar la serie de Taylor de la log-posterior alrededor de $\hat{\theta}$, observando lo siguiente

$$\log p(\theta|\mathbf{Y}) - \log p(\hat{\theta}|\mathbf{Y}) \approx \frac{1}{2}(\theta-\hat{\theta})^T \left.\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\right|_{\theta=\hat{\theta}} (\theta-\hat{\theta})$$

$$\Rightarrow -2\log\frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} \approx (\theta-\hat{\theta})^T \left[-\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta}).$$

Por lo tanto, si θ es de dimensión k, entonces:

$$-2\log\frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} \dot{\sim} \chi^2_k$$

Sea $q_{\chi^2_k}^{1-\alpha}$ el cuantil de probabilidad $1-\alpha$ de la distribución χ^2_k, i.e.

$$\mathbb{P}\left(\chi^2_k \leq q_{\chi^2_k}^{1-\alpha}\right) = 1-\alpha.$$

Entonces

$$\mathbb{P}\left[-2\log\frac{p(\theta|\mathbf{Y})}{p(\theta|\hat{\mathbf{Y}})} \leq q_{\chi^2_k}^{1-\alpha}\right] \approx 1-\alpha$$

$$\Rightarrow \mathbb{P}\left[\frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} \geq \exp\left\{-\frac{q_{\chi^2_k}^{1-\alpha}}{2}\right\}\right] \approx 1-\alpha.$$

Es decir, aquella región de Θ,

$$R(\Theta) = \left\{ \theta : \frac{p(\theta|\mathbf{Y})}{p(\hat{\theta}|\mathbf{Y})} \geq \exp\left\{ -\frac{q_{\chi_k^2}^{1-\alpha}}{2} \right\} \right\}$$

corresponde a una región de aproximadamente $1-\alpha$ de probabilidad posterior.

Aproximación Normal del modelo Beta-Binomial

Suponga que $Y_1, \ldots, Y_n|\theta \overset{iid}{\sim} \mathsf{Bernoulli}(\theta)$, y considere la previa $\mathsf{Beta}(\alpha, \beta)$, entonces $\theta|\mathbf{Y} \sim \mathsf{Beta}(\alpha^*, \beta^*)$, donde $\alpha^* = \alpha + \sum_{i=1}^n y_i$ y $\beta^* = \beta + n - \sum_{i=1}^n y_i$, luego

$$p(\theta|\mathbf{Y}) = \text{constante} \times \theta^{\alpha^*-1}(1-\theta)^{\beta^*-1},$$

y

$$\log p(\theta|\mathbf{Y}) = \text{constante} + (\alpha^* - 1)\log\theta + (\beta^* - 1)\log(1-\theta).$$

Ahora calculamos la primera derivada:

$$\frac{d}{d\theta} \log p(\theta|\mathbf{Y}) = \frac{\alpha^* - 1}{\theta} - \frac{\beta^* - 1}{1 - \theta}$$

entonces, la moda posterior de θ, $\hat{\theta}$, satisface

$$\begin{aligned} \frac{1 - \hat{\theta}}{\hat{\theta}} &= \frac{\beta^* - 1}{\alpha^* - 1} \\ \Rightarrow \frac{1}{\hat{\theta}} &= \frac{\beta^* - 1}{\alpha^* - 1} + 1 \\ \Rightarrow \hat{\theta} &= \frac{\alpha^* - 1}{\alpha^* + \beta^* - 2}. \end{aligned}$$

Calculamos la segunda derivada y la evaluamos en el moda posterior:

$$\frac{d^2}{d\theta^2} \log p(\theta|\mathbf{Y})\bigg|_{\hat{\theta}} = -\frac{\alpha^* - 1}{\hat{\theta}^2} - \frac{\beta^* - 1}{(1-\hat{\theta})^2},$$

note que

$$1 - \hat{\theta} = \hat{\theta}\left(\frac{\beta^* - 1}{\alpha^* - 1}\right),$$

entonces

$$\begin{aligned}
\frac{d^2}{d\theta^2} \log p(\theta|\mathbf{Y})\bigg|_{\hat{\theta}} &= -\frac{\alpha^* - 1}{\hat{\theta}^2} - \frac{(\alpha^* - 1)^2}{\hat{\theta}^2(\beta^* - 1)} \\
&= -\frac{\alpha^* - 1}{\hat{\theta}^2}\left(1 + \frac{\alpha^* - 1}{\beta^* - 1}\right) \\
&= -\frac{\alpha^* - 1}{\hat{\theta}^2}\left(\frac{\alpha^* + \beta^* - 2}{\beta^* - 1}\right) \\
&= -\frac{\alpha^* + \beta^* - 2}{\hat{\theta}(1-\hat{\theta})}.
\end{aligned}$$

La última igualdad se obtiene notando que

$$\frac{\alpha^* - 1}{\beta^* - 1} = \frac{\hat{\theta}}{1 - \hat{\theta}}.$$

Luego,

$$\theta|\mathbf{Y} \;\dot{\sim}\; \mathcal{N}\left(\hat{\theta}, \frac{\hat{\theta}(1-\hat{\theta})}{\alpha^* + \beta^* - 2}\right),$$

donde

$$\hat{\theta} = \frac{\alpha^* - 1}{\alpha^* + \beta^* - 2}.$$

Note que si hacemos $\alpha = 1$ y $\beta = 1$, entonces $\hat{\theta} = \bar{y}$ y

$$\theta|\mathbf{Y} \;\dot{\sim}\; \mathcal{N}\left(\hat{\theta}, \frac{\hat{\theta}(1-\hat{\theta})}{n}\right)$$

Aproximación Normal del modelo Normal con previa conjugada

Como se mencionó con anterioridad, si $Y_1, \ldots, Y_n|\mu, \sigma^2 \overset{iid}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, entonces la previa conjugada está dada por una distribución Normal-Inversa χ^2:

$$\begin{aligned} \mu|\sigma^2 &\sim \mathsf{Normal}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \mathsf{Inversa} - \chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

Se puede demostrar que las distribuciones posteriores están dadas por:

$$\mu|\sigma^2, \mathbf{Y} \sim \mathsf{Normal}(\mu_n, \sigma^2/\kappa_n)$$
$$\sigma^2|\mathbf{Y} \sim \mathsf{Inversa} - \chi^2(\nu_n, \sigma_n^2),$$

donde

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

Además, la previa de referencia está dada por

$$p(\mu,\sigma^2) \propto \frac{1}{\sigma^2}\mathbb{1}_{\mathbb{R}}(\mu)\mathbb{1}_{(0,\infty)}(\sigma^2).$$

La cual se obtiene fijando los siguientes valores para los hiperparámetros

$$\kappa_0 = 0, \quad \mu_0 \in \mathbb{R}, \quad \nu_0 = -1, \quad \sigma_0^2 = 0,$$

por lo que

$$\kappa_n = n, \quad \mu_n = \bar{y}, \quad \nu_n = n-1, \quad \sigma_n^2 = s^2,$$

con $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$.

Entonces,

$$
\begin{aligned}
p(\mu,\sigma^2|\mathbf{Y}) =& \text{constante} \times (\sigma^2)^{-1/2} \exp\left\{-\frac{\kappa_n}{2\sigma^2}(\mu-\mu_n)^2\right\} \\
& \times (\sigma^2)^{-(\nu_n/2+1)} \exp\left\{-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right\} \mathbb{1}_{\mathbb{R}}(\mu)\mathbb{1}_{(0,\infty)}(\sigma^2),
\end{aligned}
$$

$$
\log p(\mu,\sigma^2|\mathbf{Y}) = \text{constante} - \left(\frac{\nu_n+3}{2}\right)\log(\sigma^2) - \frac{\kappa_n}{2\sigma^2}(\mu-\mu_n)^2 - \frac{\nu_n\sigma_n^2}{2\sigma^2},
$$

$$\frac{\partial}{\partial \mu} \log p(\mu, \sigma^2 | \mathbf{Y}) = -\frac{\kappa_n}{\sigma^2}(\mu - \mu_n),$$

donde reconocemos inmediatamente la moda posterior de μ, dada por $\hat{\mu} = \mu_n$, y

$$\frac{\partial}{\partial \sigma^2} \log p(\mu, \sigma^2 | \mathbf{Y}) = -\left(\frac{\nu_n + 3}{2}\right)\frac{1}{\sigma^2} + \frac{\kappa_n}{2\sigma^4}(\mu - \mu_n)^2 + \frac{\nu_n \sigma_n^2}{2\sigma^4}.$$

Entonces la moda posterior de σ^2, $\hat{\sigma}^2$, satisface la siguiente ecuación

$$-\left(\frac{\nu_n + 3}{2}\right)\frac{1}{\hat{\sigma}^2} + \frac{\nu_n \sigma_n^2}{2\hat{\sigma}^4} = 0,$$

de donde obtenemos

$$\hat{\sigma}^2 = \frac{\nu_n}{\nu_n + 3}\sigma_n^2.$$

Ahora calculamos las segundas derivadas

$$
\begin{aligned}
\frac{\partial^2}{\partial \mu^2} \log p(\mu, \sigma^2 | \mathbf{Y}) &= -\frac{\kappa_n}{\sigma^2}, \\
\frac{\partial^2}{\partial \mu \partial \sigma} \log p(\mu, \sigma^2 | \mathbf{Y}) &= \frac{\kappa_n}{\sigma^4}(\mu - \mu_n), \\
\frac{\partial^2}{\partial (\sigma^2)^2} \log p(\mu, \sigma^2 | \mathbf{Y}) &= \frac{\nu_n + 3}{2\sigma^4} - \frac{\kappa_n}{\sigma^6}(\mu - \mu_n)^2 - \frac{\nu_n \sigma_n^2}{\sigma^6}.
\end{aligned}
$$

Y evaluamos estas expresiones en las modas posteriores de los parámetros, $(\hat{\mu}, \hat{\sigma}^2)$,

$$\left.\frac{\partial^2}{\partial \mu^2} \log p(\mu, \sigma^2 | \mathbf{Y})\right|_{(\hat{\mu}, \hat{\sigma}^2)} = -\frac{\kappa_n}{\hat{\sigma}^2},$$

$$\left.\frac{\partial^2}{\partial \mu \partial \sigma^2} \log p(\mu, \sigma^2 | \mathbf{Y})\right|_{(\hat{\mu}, \hat{\sigma}^2)} = 0,$$

$$\begin{aligned}
\left.\frac{\partial^2}{\partial (\sigma^2)^2} \log p(\mu, \sigma^2 | \mathbf{Y})\right|_{(\hat{\mu}, \hat{\sigma}^2)} &= \frac{\nu_n + 3}{2\hat{\sigma}^4} - \frac{\nu_n \sigma_n^2}{\hat{\sigma}^6} \\
&= \frac{\nu_n + 3}{2\hat{\sigma}^4} - \frac{\cancel{\nu_n \sigma_n^2}}{\hat{\sigma}^4 \cancel{\nu_n \sigma_n^2}} (\nu_n + 3) \\
&= -\frac{\nu_n + 3}{2\hat{\sigma}^4}.
\end{aligned}$$

Entonces,

$$\mu, \sigma^2 | \mathbf{Y} \mathrel{\dot\sim} \mathcal{N}\left(\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix}, \begin{pmatrix} \frac{\hat{\sigma}^2}{\kappa_n} & 0 \\ 0 & \frac{2}{\nu_n+3}\hat{\sigma}^4 \end{pmatrix} \right),$$

donde

$$\hat{\mu} = \mu_n \text{ and } \hat{\sigma}^2 = \frac{\nu_n}{\nu_n + 3}\sigma_n^2.$$

Si usamos la previa de referencia, obtenemos

$$\mu, \sigma^2 | \mathbf{Y} \mathrel{\dot\sim} \mathcal{N}\left(\begin{pmatrix} \bar{y} \\ \hat{\sigma}^2 \end{pmatrix}, \begin{pmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2}{n+2}\hat{\sigma}^4 \end{pmatrix} \right),$$

donde

$$\hat{\sigma}^2 = \frac{n-1}{n+2}s^2.$$

Tarea 4

- Dejar tarea 4.

Inferencia Bayesiana

Cómo evitar a Procrustes

Un modelo bayesiano es una máquina que toma de entrada la distribución previa de los parámetros y la verosimilitud y, usando el teorema de Bayes como motor, produce la distribución posterior. Sin embargo, saber la regla matemática del funcionamiento del motor suele ser de muy poca ayuda. Restringirse únicamente a aquellos modelos que permiten la manipulación matemática es una solución *procrustea*.

Ante este problema es necesario recurrir a alguna técnica numérica que permita aproximar la manipulación matemática.

Aproximación usando una rendija

Una solución sencilla cuando se tienen pocos parámetros continuos (típicamente uno o dos) consiste en **generar una rendija de valores para los parámetros**. Sea θ_j alguno de estos valores, entonces se puede calcular la distribución posterior en θ_j (salvo por una constante de proporcionalidad) usando la fórmula:

$$p(\theta_j|\mathbf{Y}) \propto p(\mathbf{Y}|\theta_j)p(\theta_j).$$

Una importante consecuencia de este hecho es que podemos **generar una muestra de la distribución posterior** a partir de la rendija de valores propuesta, simplemente basta con **seleccionar el valor θ_j de manera proporcional a** $p(\mathbf{Y}|\theta_j)p(\theta_j)$.

Simular de la predictiva posterior

También nos puede interesar **generar una muestra de la distribución predictiva**. Una vez que se cuenta con una muestra de la distribución posterior de los parámetros, $\tilde{\theta}_1, \ldots, \tilde{\theta}_S$, se puede generar una muestra $\tilde{Y}_1, \ldots, \tilde{Y}_S$ de la distribución predictiva posterior. Simplemente hay que **simular** $\tilde{Y}_s \sim p(Y|\tilde{\theta}_s)$.

- Mostrar modelo Beta-Binomial.

Pruebas de hipótesis

Algunas veces la inferencia estadística puede ser formulada como:

1. Se cuenta con una hipótesis, la cual puede ser cierta o falsa ($H : \theta \in \Theta_1$).
2. Se obtiene evidencia estadística sobre la falsedad de la hipótesis.
3. Usamos (o deberíamos usar) el teorema de Bayes para deducir de manera lógica el impacto de la evidencia en la hipótesis

$$\mathbb{P}(H|\mathbf{Y}) = \mathbb{P}(\theta \in \Theta_1|\mathbf{Y}) = \int_{\Theta_1} p(\theta|\mathbf{Y})d\theta.$$

- Mostrar paradoja de Lindley.

Estimación por intervalo

Si se cuenta con una muestra de la distribución posterior $\tilde{\theta}_1, \ldots, \tilde{\theta}_S$, se puede estimar $\mathbb{P}(\theta \in \Theta_1 | \mathbf{Y}) = \mathbb{E}\left[\mathbb{1}_{\theta \in \Theta_1} | \mathbf{Y}\right]$ mediante

$$\frac{1}{S} \sum_{s=1}^{S} \mathbb{1}_{\tilde{\theta}_s \in \Theta_1}.$$

También se vuelve sencillo estimar intervalos (θ_1, θ_2) tales que $\mathbb{P}(\theta \in (\theta_1, \theta_2)) = 1 - \alpha$. **A estos intervalos se les llama intervalos de credibilidad**. Un intervalo de particular interés es **el de menor longitud** cuya probabilidad es $1 - \alpha$ (**highest posterior density interval, HPDI**).

Estimación puntual

Recuerde que **el estimador bayesiano consiste en toda la distribución posterior**. Sin embargo, a veces nos es requerido **reportar un único valor**. En este caso es común reportar el valor más probable a posteriori (***maximum a posteriori*, MAP**). Lamentablemente, dicho estimador puntual puede dar lugar a resultados absurdos. En vez de reportar el MAP otras opciones populares son la **media o la mediana de la distribución posterior**.

Ejemplo

Considere el ejemplo del globo terráqueo, planteado en la tarea. Suponga que en 3 lanzamientos se obtiene AAA, en este caso el MAP de θ vale 1. Lo que es un resultado absurdo.

Breve comentario sobre las pruebas de hipótesis

Retomando nuestra discusión sobre las pruebas de hipótesis. De manera más general, lo que se desea es calcular

$$\mathbb{P}(H|\text{evidencia}) = \frac{\mathbb{P}(\text{evidencia}|H)\mathbb{P}(H)}{\mathbb{P}(\text{evidencia})}.$$

Lo más importante es aumentar $\mathbb{P}(H)$, lo cual requiere un esfuerzo cognitivo y argumentativo, y no se limita a una simple prueba estadística. **La estadística no es sustituto de la argumentación y de la ciencia.**