Estadística Bayesiana

Irving Gómez Méndez

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 $\theta \sim \text{Uniforme}(0, 1)$. Pero suponga que estamos interesados en $\phi = -\log \theta$, si $\theta \sim \text{Uniforme}(0, 1)$, entonces $\phi \sim \text{Exponencial}(1)$. 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

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

Información (esperada) de Fisher

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

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

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

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

- Mostrar ejemplo de caso exponencial.

Ejercicios

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

2. Sea $Y|\theta \sim \text{Poisson}(\theta)$. Demuestre que la información esperada de Fisher está dada por
$$J(\theta_0) = \frac{1}{\theta_0}.$$

3. Sea $Y|\theta \sim \text{Binomial}(n, \theta)$. 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

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

Aproximación Normal

Convergencia al parámetro que minimiza la divergencia KL

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 $p(Y|\theta)$,

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

Caso discreto

Vamos a demostrar que, cuando n aumenta, la distribución posterior $p(\theta|\mathbf{Y})$ se concentra alrededor de θ_0. Para ello, primero consideraremos el caso en que Θ es un espacio discreto.

Teorema

Si el espacio parametral Θ es finito y $\mathbb{P}(\theta = \theta_0) > 0$, entonces $\mathbb{P}(\theta = \theta_0|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$.

Demostración

Considere el logaritmo del cociente de posteriores:

$$\log\left(\frac{p(\theta|\mathbf{Y})}{p(\theta_0|\mathbf{Y})}\right) = \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(Y_i|\theta)}{p(Y_i|\theta_0)}\right)$$

Note que

$$\log\left(\frac{p(\theta)}{p(\theta_0)}\right)$$

es una constante, al no depender de n.
Por otro lado, note que

$$\begin{aligned}
&\mathbb{E}_{Y\sim f}\left[\log\left(\frac{p(Y|\theta)}{p(Y|\theta_0)}\right)\right] \\
=&\mathbb{E}\left[\log f(Y) - \log p(Y|\theta_0) - \log f(Y) + \log p(Y|\theta)\right] \\
=&KL(\theta_0) - KL(\theta)
\end{aligned}$$

y por ley fuerte de grandes números se cumple que

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p(Y_i|\theta)}{p(Y_i|\theta_0)}\right) \xrightarrow[n\to\infty]{c.s.} KL(\theta_0) - KL(\theta) < 0$$

Por lo tanto,

$$\sum_{i=1}^{n} \log\left(\frac{p(Y_i|\theta)}{p(Y_i|\theta_0)}\right) \xrightarrow[n\to\infty]{} -\infty,$$

luego

$$\frac{p(\theta|\mathbf{Y})}{p(\theta_0|\mathbf{Y})} \xrightarrow[n\to\infty]{} 0$$

y

$$p(\theta|\mathbf{Y}) \xrightarrow[n\to\infty]{} 0, \quad \text{para todo } \theta \neq \theta_0.$$

Como la suma de las probabilidades tiene que ser 1, concluimos que

$$p(\theta_0|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$$

■

Caso continuo

Teorema

Sea Θ un espacio compacto y A un vecindario de θ_0 tal que $\mathbb{P}(\theta \in A) > 0$, entonces $\mathbb{P}(\theta \in A|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$.

Demostración

Como Θ es compacto, entonces existe una cobertura finita de Θ y se puede construir de tal manera que A es el único vecindario que incluye a θ_0. Usando el teorema anterior se puede demostrar que la probabilidad posterior para cualquier vecindario que no sea A tiende a 0 cuando $n \to \infty$ y $\mathbb{P}(\theta \in A|\mathbf{Y}) \xrightarrow[n\to\infty]{} 1$.

■

Convergencia a la distribución Normal usando la información esperada

Teorema

Bajo condiciones de regularidad (que incluyen que θ_0 no esté en la frontera de Θ), la distribución posterior de θ es aproximadamente normal con media θ_0 y varianza $(nJ(\theta_0))^{-1}$

Demostración

Sea $\hat{\theta}$ la moda de la distribución posterior. Luego,

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

Note que

$$\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{Y})\bigg|_{\theta=\hat{\theta}} = \frac{d^2}{d\theta^2}\log p(\theta)\bigg|_{\theta=\hat{\theta}} + \sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log p(Y_i|\theta)\bigg|_{\theta=\hat{\theta}}$$

Por ley fuerte de grandes números y los teoremas anteriores, tenemos que

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log p(Y_i|\theta)\bigg|_{\theta=\hat{\theta}} \xrightarrow[n\to\infty]{c.s.} \mathbb{E}_{Y\sim f}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\bigg|_{\theta=\theta_0}\right]$$

Si el modelo de la verosimilitud es correcto, entonces $f(Y) = p(Y|\theta^\star)$ para algún $\theta^\star \in \Theta$. Y, por lo tanto, la divergencia de Kullback-Leibler se puede escribir como

$$KL(\theta) = \mathbb{E}_{Y\sim f}\left[\log\left(\frac{p(Y|\theta^\star)}{p(Y|\theta)}\right)\right]$$

Recordando que $KL(\theta) \geq 0$, podemos verificar fácilmente que $KL(\theta^\star) = 0$ y, por lo tanto, $\theta_0 = \theta^\star$. Es decir, el parámetro que minimiza la divergencia de Kullback-Leibler es el verdadero parámetro. Entonces

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

Sabemos que, a medida que $n \to \infty$, la distribución se concentra en vecindarios cada vez más pequeños de θ_0, y la distancia $|\hat{\theta} - \theta_0|$ se acerca a cero.

Por lo tanto, al considerar los términos de la serie de Taylor, sólo necesitamos concentrarnos en el término cuadrático, de donde tenemos que

$$p(\theta|\mathbf{Y}) \dot{\propto} \exp\left\{-\frac{1}{2}(\theta-\theta_0)^2(nJ(\theta_0))\right\}$$
$$= \exp\left\{-\frac{(\theta-\theta_0)^2}{2(nJ(\theta_0))^{-1}}\right\}$$

Es decir, para n suficientemente grande,

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

■

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

Retomando la serie de Taylor, observamos lo siguiente

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

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

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,

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

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

Convergencia a la distribución Normal usando la información observada

En el caso de que θ sea de dimension k, entonces la serie de Taylor se escribiría como:

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

y

$$\theta|\mathbf{Y} \dot{\sim} \mathsf{Normal}(\hat{\theta}, [I(\hat{\theta})]^{-1}),$$

donde

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

- Mostrar ejemplos Beta-Binomial y Gama-Exponencial.

Modelo Normal con previa no informativa

Sean $Y_1, \ldots, Y_n | \mu, \sigma^2 \overset{iid}{\sim} \mathsf{Normal}(\mu, \sigma^2)$. Al ser μ y σ parámetros de localización y escala, respectivamente. Sabemos que una previa no informativa está dada por

$$p(\mu, \log \sigma) \propto \mathbb{1}_{\mathbb{R}}(\mu) \mathbb{1}_{\mathbb{R}}(\log \sigma)$$

y la verosimilitud de la muestra observada es

$$p(\mathbf{Y} | \mu, \log \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}$$

Entonces,

$$\log p(\mathbf{Y}|\mu, \log\sigma) = \text{constante} - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

$$= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Luego,

$$\log p(\mu, \log\sigma|\mathbf{Y}) = \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Sea $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, se sigue que

$$\begin{aligned}
&\log p(\mu, \log\sigma|\mathbf{Y}) \\
&= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^{n}(y_i - \mu)^2 \\
&= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2 \\
&= \text{constante} - n\log\sigma - \frac{1}{2}\exp\{-2\log\sigma\}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]
\end{aligned}$$

$$\frac{\partial}{\partial \mu} \log p(\mu, \log \sigma | \mathbf{Y}) = \exp\{-2 \log \sigma\} n(\bar{y} - \mu)$$

$$\frac{\partial}{\partial \log \sigma} \log p(\mu, \log \sigma | \mathbf{Y}) = -n + \exp\{-2 \log \sigma\} \left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]$$

Sea $(\hat{\mu}, \log \hat{\sigma})$ el punto en que se maximiza la densidad posterior. Podemos obtener este punto al evaluar en $(\hat{\mu}, \log \hat{\sigma})$ las expresiones anteriores e igualando a cero. De donde obtenemos que $\hat{\mu} = \bar{y}$ y

$$-n + \exp\{-2 \log \hat{\sigma}\} [(n-1)s^2] = 0$$

$$\Rightarrow \log \hat{\sigma} = \log \left(\sqrt{\frac{n-1}{n}} s \right).$$

Calculamos ahora las segundas derivadas para obtener la información observada (de Fisher)

$$\frac{\partial^2}{\partial\mu\partial\log\sigma}\log p(\mu,\log\sigma|\mathbf{Y}) = -2n\exp\{-2\log\sigma\}(\bar{y}-\mu)$$

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

$$\frac{\partial^2}{\partial(\log\sigma)^2}\log p(\mu,\log\sigma|\mathbf{Y}) = -2\exp\{-2\log\sigma\}\left[(n-1)s^2+n(\bar{y}-\mu)^2\right]$$

Entonces,

$$I(\hat{\mu}, \log \hat{\sigma}) = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & 2n \end{pmatrix}.$$

Por lo tanto,

$$\mu, \log \sigma | \mathbf{Y} \dot{\sim} \mathsf{Normal}\left(\begin{pmatrix} \hat{\mu} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{1}{2n} \end{pmatrix} \right).$$

Ejercicio

Sabemos que si, en vez de haber considerado $\mu, \log \sigma$, hubiéramos considerado μ, σ^2, entonces la previa no informativa está dada por

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

Demuestre que en este caso

$$\mu, \sigma^2 | \mathbf{Y} \dot{\sim} \text{Normal}\left(\begin{pmatrix} \hat{\mu} \\ \tilde{\sigma}^2 \end{pmatrix}, \begin{pmatrix} \frac{\tilde{\sigma}^2}{n} & 0 \\ 0 & \frac{2}{n+2}\tilde{\sigma}^4 \end{pmatrix} \right),$$

donde

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

Un modelo no regular

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

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

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 $U \sim \text{Uniforme}(0,1)$, 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

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

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 se demuestra 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.

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, $\theta_1, \ldots, \theta_m$, se puede generar una muestra $Y_1, \ldots, Y_m$ de la distribución predictiva posterior. Simplemente hay que simular $Y_j \sim p(Y|\theta_j)$.

- 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 $\theta_1, \ldots, \theta_m$, se puede estimar $\mathbb{P}(\theta \in \Theta_1 | \mathbf{Y}) = \mathbb{E}\left[\mathbb{1}_{\theta \in \Theta_1} | \mathbf{Y}\right]$ mediante

$$\frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{\theta_j \in \Theta_1}.$$

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 (MAP)

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.

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.

Estimación puntual

En vez de reportar el MAP se podría optar por la media o la mediana de la distribución posterior, pero entonces surge la pregunta de qué estimador puntual es el que deberíamos de reportar. Una manera de tomar esta decisión es a través del uso de alguna función de pérdida $L(\theta, \theta_0)$, donde θ_0 es el verdadero parámetro. Pero, ¡el verdadero parámetro es desconocido!

Para solucionar este inconveniente se minimiza la pérdida esperada, tomando el valor esperado con respecto a la distribución posterior. Es decir se calcula

$$\begin{aligned} L(\theta) &= \mathbb{E}_{\tilde{\theta}\sim p(\theta|\mathbf{Y})}[L(\theta,\tilde{\theta})] \\ &= \int_{\Theta} L(\theta,\tilde{\theta})p(\tilde{\theta}|\mathbf{Y})d\tilde{p} \end{aligned}$$

y se selecciona $\hat{\theta} \in \arg\min_{\theta\in\Theta} L(\theta)$.

Si se cuenta con una muestra $\theta_1, \ldots, \theta_m$ de la distribución posterior, entonces $L(\theta)$ puede ser estimado mediante

$$\frac{1}{m}\sum_{j=1}^{m} L(\theta,\theta_j)$$

- Mostrar modelo Beta-Binomial.

Cómo determinar qué función de pérdida usar

Considere el caso en que se requiere decidir si ordenar una evacuación o no con base en la velocidad del viento provocado por un huracán.

El riesgo a que haya fallecidos y/o personas afectadas aumenta rápidamente conforme aumenta la rapidez del viento. Sin embargo, también se induce un costo al ordenar una evacuación innecesaria, aunque este es mucho menor.

Por lo tanto, se debería usar una función de pérdida muy asimétrica, que crece rápidamente cuando la velocidad del viento excede nuestra inferencia, pero crece lentamente cuando la velocidad del viento es menor que nuestra inferencia.

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.

Validación del Modelo

Simulación de la predictiva previa

Note que también se puede simular una muestra de la predictiva previa, para ello basta con simular una muestra de los parámetros $\theta_1, \ldots, \theta_m$, a partir de la distribución previa y luego simular $Y_j \sim P(Y|\theta_j)$.

Poder simular de la predictiva previa puede ayudar a discriminar entre distintas distribuciones previas. Pues al simular de la predictiva previa podemos observar las consecuencias de las distribución previa sobre la variable de interés Y. Muchas de las técnicas convencionales para decidir la distribución previa pueden dar lugar a distribuciones más bien absurdas.

Validación del modelo

Una vez que hemos completado los dos primeros pasos del análisis bayesiano: contrucción de un modelo de probabilidad y cómputo de la distribución posterior, no deberíamos ignorar el paso (relativamente sencillo) de validar que tan bien se ajusta nuestro modelo a los datos y a nuestro conocimiento del fenómeno.

Dado que de antemano sabemos que nuestro modelo no puede incluir todos los aspectos de la realidad, podemos averiguar cuáles aspectos no son capturados por el modelo. Y, sobre todo sobre la plausibilidad de nuestro modelo para el propósito para el que fue construido. No se trata de preguntarnos si nuestro modelo es falso o verdadero, sino de cuáles son las principales deficiencias de nuestro modelo.

Si el modelo ajusta bien, entonces **datos replicados con nuestro modelo generativo deberían parecer similares a los datos observados**. Dicho de otra manera, los datos observados deben de parecer plausibles al considerar la distribución predictiva posterior. Nuestra técnica básica para validar el ajuste de un modelo es simular datos de la distribución predictiva posterior y compararlos con los datos observados.

Para que la comparación sea correcta los datos replicados, $\mathbf{Y}^{rep}$ deben de ser (cómo su nombre indica) réplicas de los datos observados. Es decir, $\mathbf{Y}^{rep}$ debe de ser de la misma dimensión que $\mathbf{Y}$, y si nuestro modelo cuenta con variables predictoras $\mathbf{X}$, entonces debemos usar exactamente los mismos valores de las variables predictoras.

Estadísticas de prueba y p-valor bayesiano

Para medir la discrepancia entre el modelo ajustado y los datos, definimos una estadística de prueba $T(\mathbf{Y})$. El p-valor bayesiano se define entonces como la probabilidad de que la estadística de prueba evaluada en los datos replicados, $T(\mathbf{Y}^{rep})$, sea más extrema que la estadística de prueba evaluada en los datos observados.

$$\begin{aligned} p_B &= \mathbb{P}(T(\mathbf{Y}^{rep}) > T(\mathbf{Y})|\mathbf{Y}) \\ &= \mathbb{E}\left[\mathbb{1}_{T(\mathbf{Y}^{rep})>T(\mathbf{Y})}|\mathbf{Y}\right] \end{aligned}$$

Como $p_B = \mathbb{E}\left[\mathbb{1}_{T(\mathbf{Y}^{rep})>T(\mathbf{Y})|\mathbf{Y}}\right]$, entonces una manera de estimar el p-valor es simulando m réplicas, $\mathbf{Y}_1^{rep}, \ldots, \mathbf{Y}_m^{rep}$ de la distribución predictiva posterior y calcular

$$\hat{p}_B = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{T(\mathbf{Y}_j^{rep})>T(\mathbf{Y})}.$$

Más aún, si contamos con una muestra $\theta_1, \ldots, \theta_m$ de la distribución posterior, podemos simular $\mathbf{Y}_j^{rep}$ a partir de la verosimilitud $p(\mathbf{Y}|\theta_j)$.

- Mostrar ejemplo de la estatura de la población !Kung.

Regresión Bayesiana

Regresión de la estatura en el peso

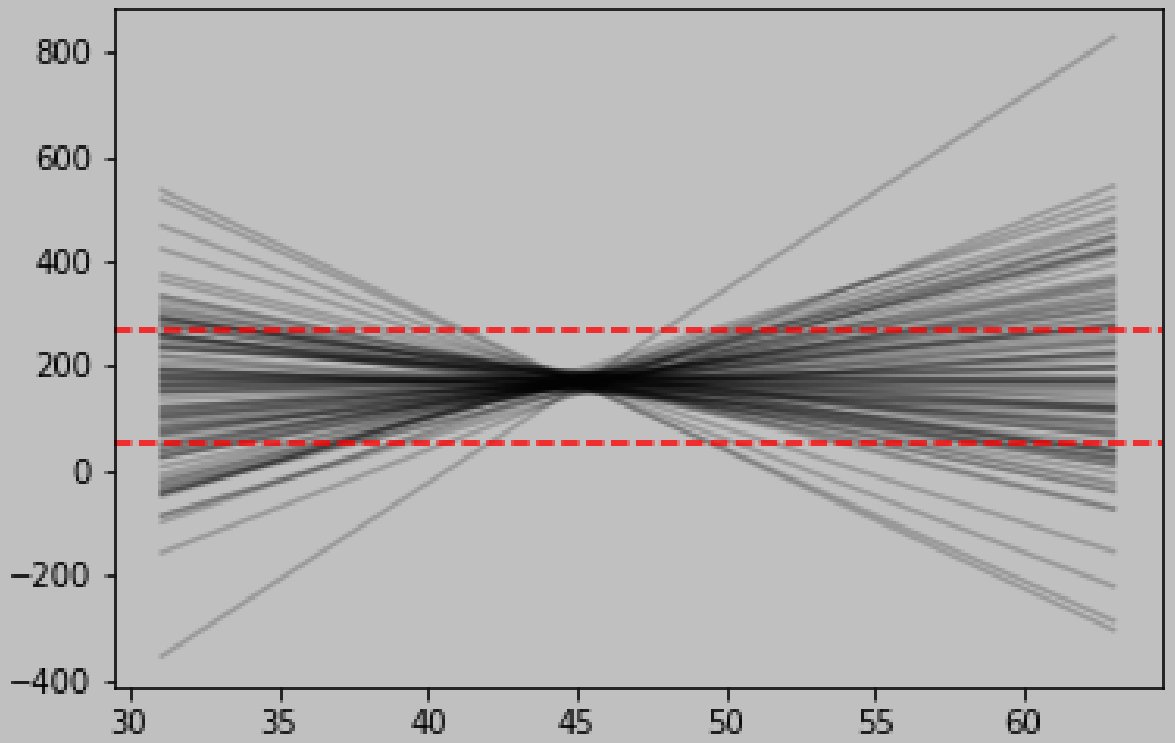
Para estos ejercicios vamos a usar los datos de la comunidad !Kung. Consideraremos sólo a los adultos, esto porque la estatura está fuertemente relacionada con la edad hasta antes de la adultez.

Sea h_i la estatura del i-ésimo individuo, w_i su peso y $\bar{w}$ el peso promedio de los adultos. Vamos a considerar el modelo:

$$
\begin{aligned}
h_i|\mu_i, \sigma &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i|\alpha, \beta &= \alpha + \beta(w_i - \bar{w}) \\
\alpha &\sim \mathcal{N}(170, 10^2) \\
\beta &\sim \mathcal{N}(0, 10^2) \\
\sigma &\sim \mathcal{U}(0, 33)
\end{aligned}
$$

Note cuando $w_i = \bar{w}$, $\mu_i = \alpha$. Es decir cuando el peso es igual al promedio de la población la estatura promedio será α, así que hace mucho sentido modelar $\alpha \sim \mathcal{N}(170, 10^2)$. Para σ simplemente usamos una distribución con poca información. Pero **¿por qué hemos puesto esa previa para β?**

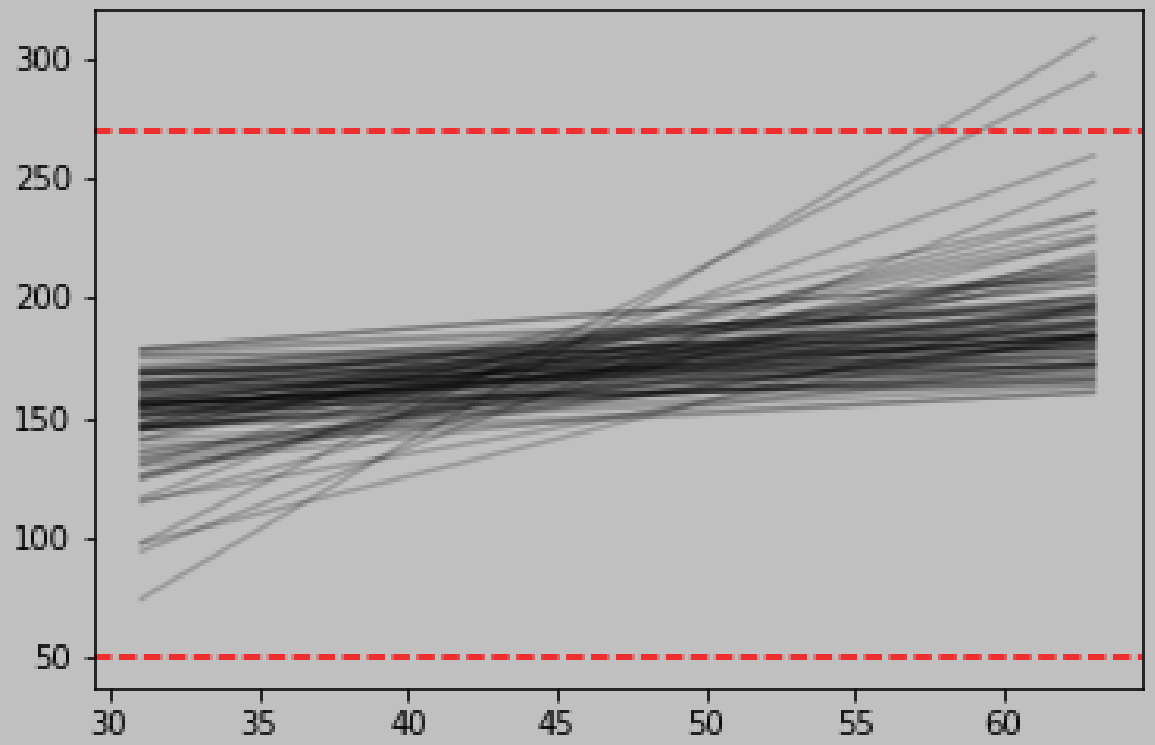
Una manera de verificar qué tan sensatas son estas previa es simulando valores de α y β, $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m$ y luego graficar $\alpha_j + \beta_j(w - \bar{w}), \quad j = 1, \ldots, m.$



¡Evidentemente esta previa para β es absurda!

Cualquiera sabe que el peso y la estatura guardan (hasta cierto punto) una correlación positiva, por lo que es sensato considerar una distribución previa estrictamente positiva para β. Por lo que consideramos el siguiente modelo:

$$
\begin{aligned}
h_i|\mu_i, \sigma &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i|\alpha, \beta &= \alpha + \beta(w_i - \bar{w}) \\
\alpha &\sim \mathcal{N}(170, 10^2) \\
\beta &\sim \text{lognormal}(0, 1) \\
\sigma &\sim \mathcal{U}(0, 33)
\end{aligned}
$$


300
250
200
150
100
50
30
35
40
45
50
55
60

Estimador MAP

Una vez establecido el modelo podemos crear una rendija de valores para α, β y σ. Y como antes, simular una muestra de la posterior. Con esta muestra podemos estimar puntualmente los parámetros y graficar la recta con mayor probabilildad a posteriori $\hat{\alpha} + \hat{\beta}(w - \bar{w})$.

Intervalos de confianza

Como μ depende de los parámetros y estos tienen una distribución posterior, entonces μ también cuenta con una distribución posterior. En particular fijando un peso w, y contando con una muestra a posteriori $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m$ podemos obtener una muestra a posteriori $\mu_1, \ldots, \mu_m$ usando la relación:

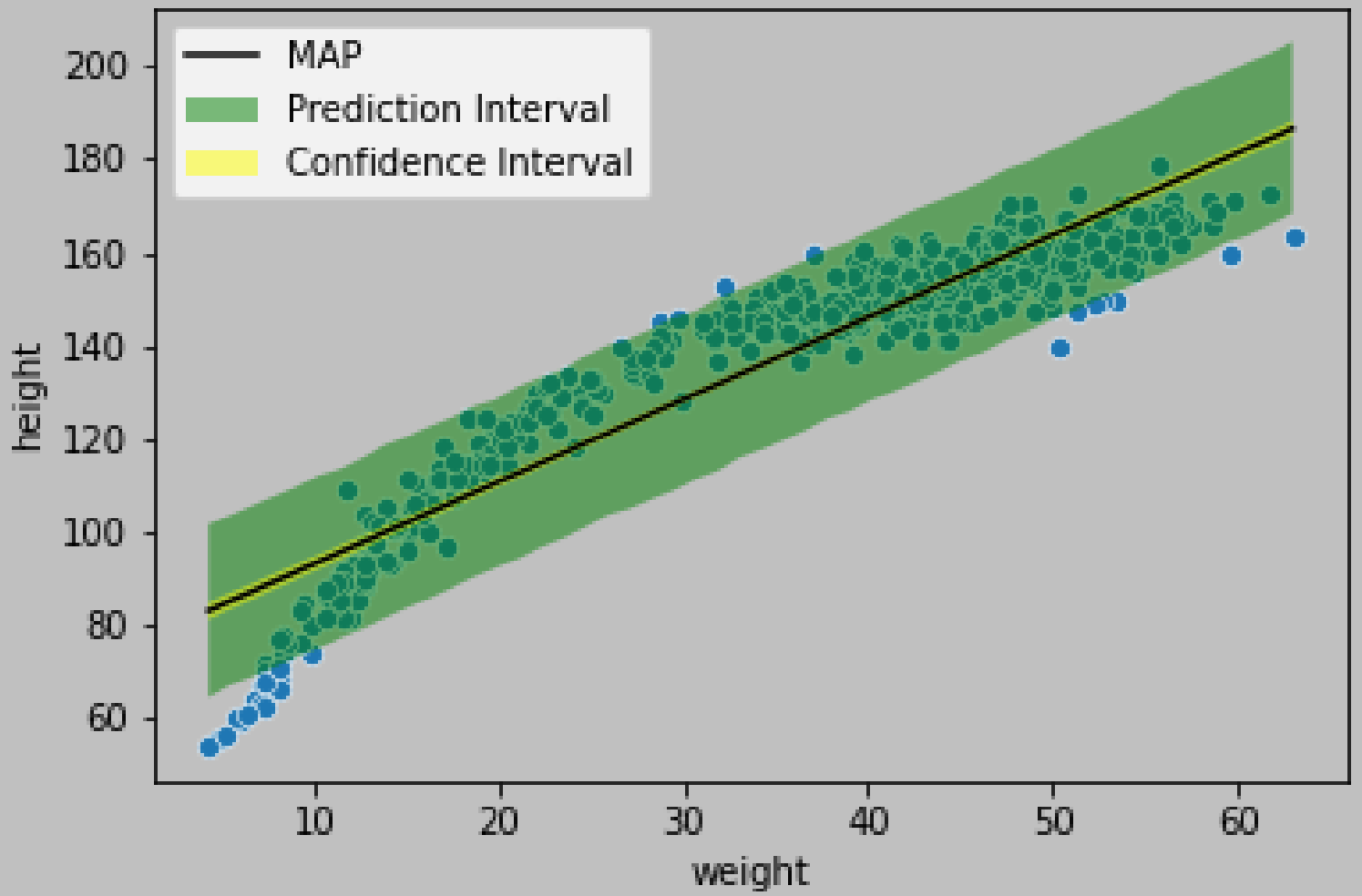
$$\mu_j = \alpha_j + \beta_j(w - \bar{w}).$$

A partir de esta muestra podemos obtener un intervalo de confianza para la estatura media dado el peso w. Variando w desde el valor más bajo hasta el valor más alto podemos obtener un intervalo de confianza para la función de regresión.

Intervalos de predicción

Finalmente, para agregar intervalos de predicción hay que recordar que $h \sim \mathcal{N}(\mu, \sigma^2)$. Así para un peso fijo w podemos obtener una muestra de la posterior de μ como se explicó antes, $\mu_1, \ldots, \mu_m$. Además podemos simular una muestra de la posterior de σ, $\sigma_1, \ldots, \sigma_m$. Por lo tanto podemos obtener una muestra de la posterior de h, $h_1, \ldots, h_m$, donde $h_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. A partir de la muestra $h_1, \ldots, h_m$ podemos obtener un intervalo de predicción para el peso w. Variando w desde el valor más bajo hasta el valor más alto podemos obtener un intervalo de predicción para la función de regresión.

- Mostrar ejemplo de regresión de la estatura de la población !Kung.


MAP
Prediction Interval
Confidence Interval
height
weight
200
180
160
140
120
100
80
60
10
20
30
40
50
60

Aproximación por cuadratura

Es claro que una de las limitantes del método de aproximación usando una rendija es que escala pobremente cuando incrementamos el número de parámetros en el modelo (maldición de la alta dimensionalidad). De los ejemplos anteriores también hemos visto que las posteriores de los parámetros tienden a tomar una forma acampanada cuando el tamaño de la muestra es grande (teorema del límite central).

Esta observación nos permite introducir el método de aproximación por cuadratura. Hay que recordar que la densidad normal tiene esta forma de campana y que el logaritmo es proporcional a $\lambda(Y - \mu)^2$, i.e. es una función cuadrática (una parábola). Usando este hecho podemos aproximar la posterior de los parámetros mediante una distribución normal.

Regresión de la estatura en el peso

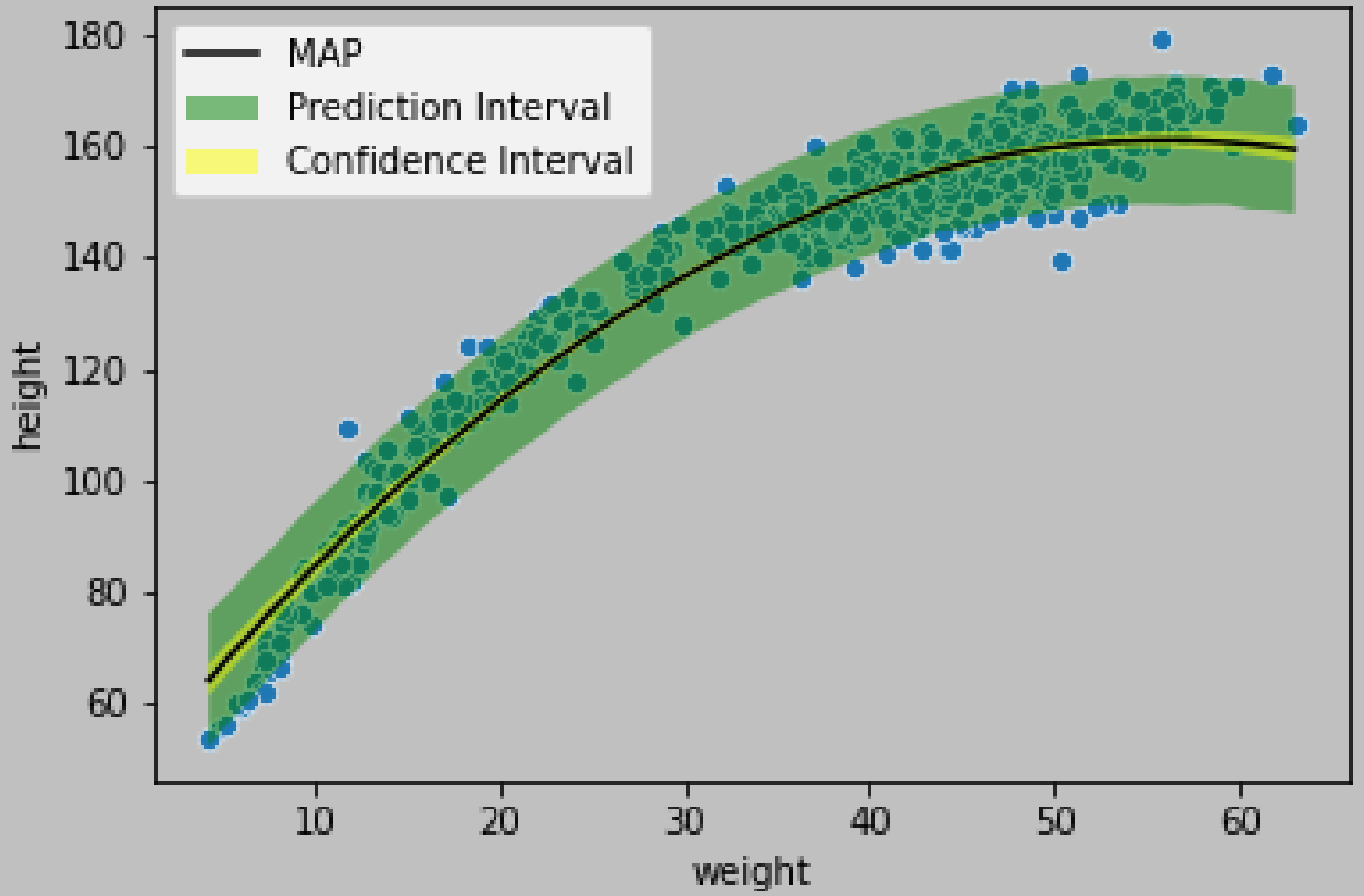
Consideremos una vez más el ejemplo de predecir la estatura de la población !Kung según su peso, pero ahora tomemos todos los individuos de la muestra. Ahora estamos dispuestos a aceptar que pueda haber mayor variabilidad en la estatura así que cambiamos la distribución previa de σ por $\sigma \sim \mathcal{U}(0, 50)$. Antes de ajustar el modelo estandarizamos la altura y el peso (restamos la media y dividimos por el desvío padrón), más adelante veremos por qué hemos hecho esto. Para estimar la posterior de los parámetros usamos la aproximación por cuadratura.

- Mostrar ejemplos de aproximación por cuadratura.

El modelo no es satisfactorio por lo que en su lugar proponemos el siguiente modelo:

$$
\begin{aligned}
h_i|\mu_i, \sigma &\sim \mathcal{N}(\mu_i, \sigma^2) \\
\mu_i|\alpha, \beta_1, \beta_2 &= \alpha + \beta_1 y_i + \beta_2 y_i^2 \\
\alpha &\sim \mathcal{N}(170, 10^2) \\
\beta_1 &\sim \text{lognormal}(0, 1) \\
\beta_2 &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{U}(0, 50),
\end{aligned}
$$

donde $y_i = (w_i - \bar{w})/s_w$.
Como estamos considerando una potencia de la variable predictora conviene estandarizarla antes para evitar posibles problemas numéricos.


MAP
Prediction Interval
Confidence Interval
height
weight
180
160
140
120
100
80
60
10
20
30
40
50
60

Tarea 3

- Dejar tarea 3.