

Estadística Bayesiana

Irving Gómez Méndez



Aproximación Normal

Aproximación Normal

Sea $Y_1, \dots, Y_n \stackrel{iid}{\sim} f(Y)$, 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 $p(Y|\theta)$,

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

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 $\mathbb{P}(\theta = \theta_0|\mathbf{Y}) \xrightarrow{n \rightarrow \infty} 1$.

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

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

y por ley fuerte de grandes números se cumple que

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

Por lo tanto,

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

luego

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

y

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

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

$$p(\theta_0|\mathbf{Y}) \xrightarrow{n \rightarrow \infty} 1$$



Teorema

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

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 $n \rightarrow \infty$ y $\mathbb{P}(\theta \in A|\mathbf{Y}) \xrightarrow[n \rightarrow \infty]{} 1$.



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,

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

Note que

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

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

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

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

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

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

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

Sabemos que, a medida que $n \rightarrow \infty$, 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

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

Es decir, para n suficientemente grande,

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



Retomando la serie de Taylor, observamos lo siguiente

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

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

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

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

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

Entonces

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

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

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 $1 - \alpha$ probabilidad posterior.

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

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

y

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

donde

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

Modelo Normal con previa no informativa

Sean $Y_1, \dots, Y_m | \mu, \sigma^2 \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. 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,

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

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

$$\log p(\mu, \log \sigma | \mathbf{Y})$$

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

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

$$= \text{constante} - n \log \sigma - \frac{1}{2} \exp \{-2 \log \sigma\} \left[(n-1)s^2 + n(\bar{y} - \mu)^2 \right]$$

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

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

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

$$\begin{aligned} -n + \exp \{-2 \log \hat{\sigma}\} [(n-1)s^2] &= 0 \\ \Rightarrow \log \hat{\sigma} &= \log \left(\sqrt{\frac{n-1}{n}} s \right). \end{aligned}$$

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,

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

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

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

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

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

Regresión Bayesiana

Regresión (Normal-Inversa Gama)

Detalles sangrientos



Tarea 2

- Dejar tarea 2.

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 (típicamente uno o dos) continuos 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, $\theta_1, \dots, \theta_m$, se puede generar una muestra y_1, \dots, y_m de la distribución predictiva posterior. Simplemente hay que simular $\mathbf{Y}_j \sim p(Y|\theta_j)$.

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

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

Estimación por intervalo

Si se cuenta con una muestra de la distribución posterior $\theta_1, \dots, \theta_m$, se puede estimar $\mathbb{P}(\theta \in \Theta_1 | \mathbf{Y}) = \mathbb{E}[\mathbb{1}_{\theta \in \Theta_1} | \mathbf{Y}]$ 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 5

Considere el ejercicio, planteado en la tarea, del globo terráqueo. Suponga que en 3 lanzamientos se obtiene AAA, en este caso el MAP 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 $\theta_1, \dots, \theta_m$ de la distribución posterior, entonces $L(\theta)$ puede ser estimado mediante

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

► Mostrar ejemplo 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.

Simular 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 $\theta_1, \dots, \theta_m$, 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.

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

Previas Poco Informativas

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:

$$h_i \sim \mathcal{N}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta(w_i - \bar{w})$$

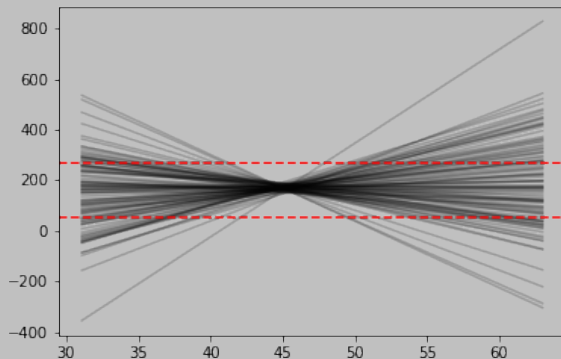
$$\alpha \sim \mathcal{N}(170, 10)$$

$$\beta \sim \mathcal{N}(0, 10)$$

$$\sigma \sim \mathcal{U}(0, 33)$$

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 $\alpha \sim \mathcal{N}(170, 10)$. 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 previas es simulando valores de α y β , $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m$ y luego graficar $\alpha_j + \beta_j(w_j - \bar{w})$, $j = 1, \dots, m$.



¡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:

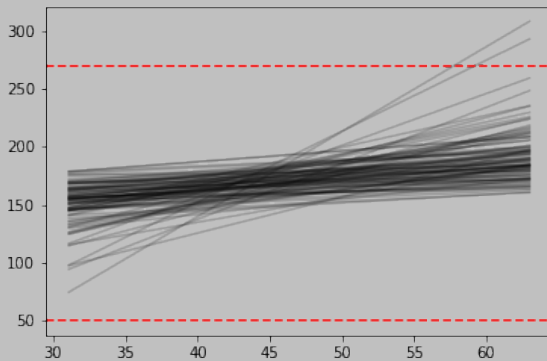
$$h_i \sim \mathcal{N}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta(w_i - \bar{w})$$

$$\alpha \sim \mathcal{N}(170, 10)$$

$$\beta \sim \text{lognormal}(0, 1)$$

$$\sigma \sim \mathcal{U}(0, 33)$$



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 probabilidad 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 $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m$ podemos obtener una muestra a posteriori μ_1, \dots, μ_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 $h \sim \mathcal{N}(\mu, \sigma)$. Así para un peso fijo w podemos obtener una muestra de la posterior de μ como se explicó antes, μ_1, \dots, μ_m . Además podemos simular una muestra de la posterior de σ , $\sigma_1, \dots, \sigma_m$. Por lo tanto podemos obtener una muestra de la posterior de h , h_1, \dots, h_m , donde $h_j \sim \mathcal{N}(\mu_j, \sigma_j)$. A partir de la muestra h_1, \dots, 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.

Aproximación por cuadratura

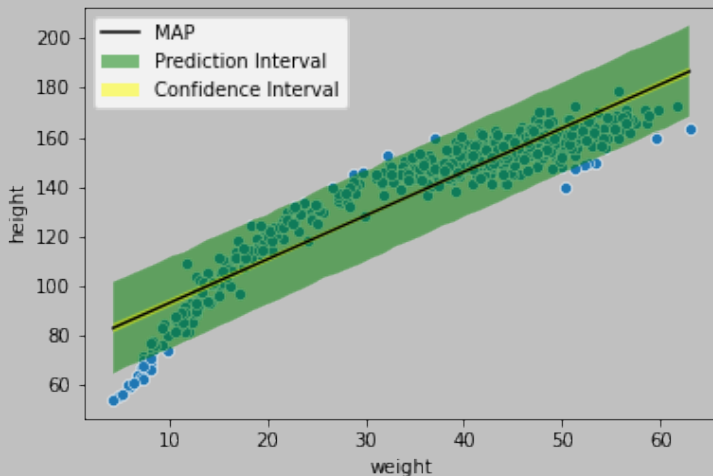
Es claro que una de las limitantes del método de aproximación usando una rendija es que escala pobremente cuando incrementamos el número de parámetros en el modelo (maldición de la alta dimensionalidad). De los ejemplos anteriores también hemos visto que las posteriores de los parámetros tienden a tomar 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 $\lambda(x - \mu)^2$, 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 (2)

- Mostrar ejemplos de aproximación por cuadratura.

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.



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

$$h_i \sim \mathcal{N}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_1 y_i + \beta_2 y_i^2$$

$$\alpha \sim \mathcal{N}(170, 10)$$

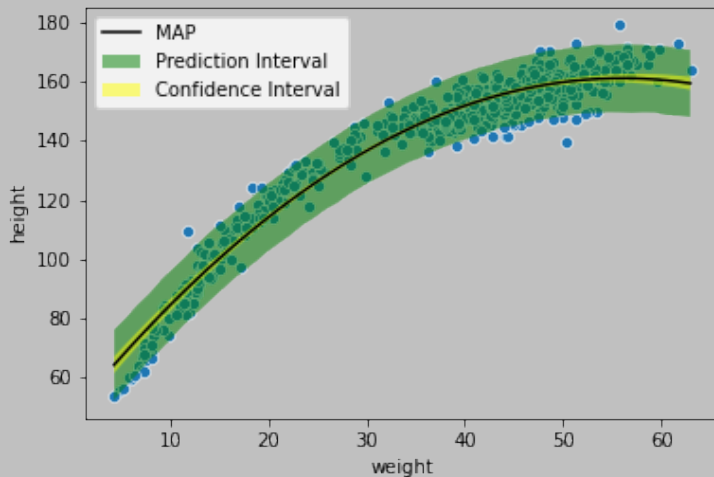
$$\beta_1 \sim \text{lognormal}(0, 1)$$

$$\beta_2 \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \mathcal{U}(0, 50),$$

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.



Tarea 3

- ▶ Dejar tarea 3.