Análisis bayesiano del modelo lineal normal

Para realizar un análisis bayesiano del modelo lineal normal requerimos especificar una distribución inicial para θ y, mediante el teorema de Bayes, actualizarla para obtener una distribución posterior dados los datos observados. Entonces, primero presento una manipulación de la función de verosimilitud para después ver algunas distribuciones iniciales frecuentemente utilizadas y, finalmente, realizar la actualización de las mismas dados los datos.

Verosimilitud

Siguiendo a **GP98Congdon06**, manipulemos la función de verosimilitud de la normal multivariada para facilitar la actualización mediante el teorema de Bayes. Observemos que:

$$\begin{aligned} f(y|\theta) &= \frac{1}{\sqrt{(2\pi)|\sigma^2\mathbb{I}_N|}} exp\left\{-\frac{1}{2}(y-X\beta)^T(\sigma^2\mathbb{I}_N)^{-1}(y-X\beta)\right\} \\ &\propto (\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\} \end{aligned} \tag{1}$$

En el análisis clásico o frecuentista, el estimador máximo verosímil para los coeficientes β es $b = (X^TX)^{-1}X^Ty$. Podemos manipular los términos dentro de la exponencial en la distribución normal con este estimador b:

$$\begin{aligned} y - X\beta &= y - Xb + Xb - X\beta = (y - Xb) + X(b-\beta) \\ \Rightarrow (y-X\beta)^T(y-X\beta) &= \left\{(y-X\beta)^T + [X(b-\beta)]^T\right\}\left\{(y-Xb)+X(b-\beta)\right\} \\ &= (y-Xb)^T(y-Xb) + (y-Xb)^TX(b-\beta) + \\ &\quad [X(b-\beta)]^T(y-Xb) + [X(b-\beta)]^TX(b-\beta) \end{aligned}$$

y, agrupando los términos cruzados en $k(y,\beta)$,

$$\Rightarrow (y-X\beta)^T(y-X\beta) = (y-Xb)^T(y-Xb) + (b-\beta)^TX^TX(b-\beta) + k(y,\beta)\,. \tag{2}$$

En realidad, $k(y,\beta) = 0$:

$$k(y,\beta) = (y-Xb)^TX(b-\beta) + [X(b-\beta)]^T(y-Xb)$$

notando que el segundo término es igual al primero pero transpuesto,

$$(y-Xb)^TX(b-\beta) = (y^T - b^TX^T)(Xb - X\beta)$$

sustituyendo el valor de b y considerando que $Xb = y$

$$\begin{aligned} (y-Xb)^TX(b-\beta) &= \left\{y^T - \left[(X^TX)^{-1}X^Ty\right]^TX^T\right\}(y-X\beta) \\ &= \left\{y^T - \left[y^TX(X^TX)^{-T}\right]X^T\right\}(y-X\beta) \\ &= \left[y^T - y^TX(X^{-1}X^{-T})X^T\right](y-X\beta) \\ &= (y^T - y^T)(y-X\beta) \end{aligned}$$

entonces,

$$(y - Xb)^T X(b - \beta) = 0 \quad \Longrightarrow \quad k(y, \beta) = 0\,.$$

Podemos entonces sustituir (2) con $k(y, X, \beta) = 0$ en (1):

$$\begin{aligned} f(y|\theta) &\propto (\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\left[(y - Xb)^T(y - Xb) + (b - \beta)^T X^T X(b - \beta)\right]\right\} \\ &\propto (\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\left[(y - Xb)^T(y - Xb) + (\beta - b)^T X^T X(\beta - b)\right]\right\} \end{aligned}$$

Igual que con el estimador b para los coeficientes, podemos utilizar el estimador máximo verosimil de la varianza, $\hat{\sigma}^2 = \frac{1}{N}(y - Xb)^T(y - Xb)$, para preparar la verosimilitud de $y|\theta$:

$$f(y|\theta) \propto (\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\left[N\hat{\sigma}^2 + (\beta - b)^T X^T X(\beta - b)\right]\right\}$$

Notemos ahora que si la varianza σ^2 fuera conocida podríamos descomponer esta distribución en dos partes, una de las cuales tiene la forma del kernel de una distribución normal para $\beta|\sigma^2$, lo que sugiere ya la familia conjugada de distribuciones iniciales:

$$f(y|\theta) \propto exp\left\{-\frac{1}{2\sigma^2}\left[(\beta - b)^T X^T X(\beta - b)\right]\right\}(\sigma^2)^{-N/2} exp\left\{-\frac{N\hat{\sigma}^2}{2\sigma^2}\right\}.$$

Finalmente, en este contexto resultará más fácil trabajar en términos de precisiones que de varianzas. Si definimos la precisión de una variable normal como $\tau = \frac{1}{\sigma^2}$, tenemos que la función de verosimilitud en el modelo normal se puede representar como sigue:

$$p(y|\theta) \propto exp\left\{-\frac{\tau}{2}\left[(\beta - b)^T X^T X(\beta - b)\right]\right\}\tau^{N/2} exp\left\{-\frac{N\hat{\sigma}^2\tau}{2}\right\}. \qquad (3)$$

Distribuciones iniciales

La primera distribución inicial que podríamos plantear sería la distribución conjugada. Recordemos que esta debe tener la misma forma funcional que la verosimilitud, por lo que (3) sugiere lo siguiente:

$$f(\theta) = f(\beta, \tau) \propto exp\left\{-\frac{\tau}{2}\left[(\beta - b_0)^T T_0(\beta - b_0)\right]\right\}\tau^{a/2} exp\left\{-\frac{r\tau}{2}\right\},$$

donde b_0, T_0, a y r sean algunos parámetros convenientes. Con esta forma, podemos determinar la familia conjugada en un proceso de dos pasos. En primer lugar, asumimos que la varianza o precisión está dada, lo que permite definir una distribución inicial para $\beta|\tau$. Posteriormente, determinaremos la distribución inicial conjugada para τ. Es decir, separaremos la distribución inicial en dos: $f(\theta) = f(\beta, \tau) = f(\beta|\tau)f(\tau)$.

La distribución condicional resulta ser una normal centrada en b_0 y con precisión τT_0, por lo que debemos completarla multiplicando por $1 = \tau^{(d-d)/2}$, donde d es el número de coeficientes, incluyendo a β_0. Así:

$$\begin{aligned} f(\theta) = f(\beta|\tau)f(\tau) &\propto exp\left\{-\frac{\tau}{2}\left[(\beta-b_0)^T T_0(\beta-b_0)\right]\right\}\tau^{a/2}exp\left\{-\frac{r\tau}{2}\right\} \\ &\propto \tau^{(d-d)/2}exp\left\{-\frac{\tau}{2}\left[(\beta-b_0)^T T_0(\beta-b_0)\right]\right\}\tau^{a/2}exp\left\{-\frac{r\tau}{2}\right\}. \end{aligned}$$

Con lo que

$$\begin{aligned} f(\beta|\tau) &\propto \tau^{d/2}exp\left\{-\frac{\tau}{2}\left[(\beta-b_0)^T T_0(\beta-b_0)\right]\right\} \text{ y} \\ f(\tau) &\propto \tau^{(a-d)/2}exp\left\{-\frac{r\tau}{2}\right\}. \end{aligned} \tag{4}$$

La distribución inicial de τ también ya tiene una forma conocida: es proporcional a una gamma. Para verlo solo basta con un poco de álgebra para verificar que el parámetro de forma debe ser $a_0 = (a-d+2)/2 = (a^\star - d)/2$ con $a^\star = a+2$ y el de tasa $r_0 = r/2$. Por lo tanto, en su conjunto, tenemos que θ tiene una distribución inicial *Normal-Gamma*:

$$\theta = (\beta, \tau) \sim NG_d\left(b_0, T_0, a_0 = \frac{a^\star - d}{2}, r_0 = \frac{r}{2}\right)$$

de forma que

$$\beta|\tau \sim N_d(b_0, \tau T_0) \text{ y } \tau \sim \Gamma\left(a_0 = \frac{a^\star - d}{2}, r_0 = \frac{r}{2}\right). \tag{5}$$

Cabe hacer notar que esta distribución inicial conjugada es propia siempre que $a^\star > d$, $r > 0$ y $B_0 = T_0^{-1}$ sea positiva definida.

Por otro lado, si se buscan distribuciones iniciales más vagas, resulta que también es posible obtener distribuciones mínimo informativas límites de esta conjugada. Por ejemplo, aunque es impropia, la inicial de Jeffreys es de esa forma con los siguientes límites: $a^\star \to d$, $r \to 0$ y $B_0 = T_0^{-1} \to \mathbf{O}$. La (4) se reduce a la siguiente expresión (**GP98**):

$$f(\theta) = f(\beta, \tau) \propto \tau^{(d-2)/2} \tag{6}$$

Distribuciones finales

Consideremos para la actualización el caso general de la distribución inicial normal gamma de (5).

$$\begin{aligned} &y|\theta \sim N_N(X\beta, \sigma^2\mathbb{I}_N) \quad \text{tal que} \quad \theta = (\beta, \sigma^2) \sim f(\beta, \sigma^2) \\ &\beta|\tau \sim N_d(b_0, \tau T_0) \\ &\tau \sim \Gamma\left(a_0 = \frac{a^\star - d}{2}, r_0 = \frac{r}{2}\right). \end{aligned} \tag{7}$$

Aplicaremos el teorema de Bayes con base en (3) y (4) buscando, al tener una inicial conjugada, mantener la forma de normal gamma. Esto es, la verosimilitud

la podemos ver también como el producto de dos distribuciones, una normal para $\beta|\tau$ centrada en el estimador máximo verosímil b y con precisión $\tau X^T X$ y una gamma para τ utilizando el estimador máximo verosímil de la varianza $\hat{\sigma}^2$.

$$\begin{aligned}
f(\theta|y) &\propto f(y|\theta)f(\theta) \\
&\propto exp\left\{-\frac{\tau}{2}\left[(\beta-b)^T X^T X(\beta-b)\right]\right\}\tau^{N/2}exp\left\{-\frac{N\hat{\sigma}^2\tau}{2}\right\} \\
&\quad \tau^{d/2}exp\left\{-\frac{\tau}{2}\left[(\beta-b_0)^T T_0(\beta-b_0)\right]\right\}\tau^{(a-d)/2}exp\left\{-\frac{r\tau}{2}\right\} \\
&\propto \tau^{d/2}exp\left\{-\frac{\tau}{2}\left[(\beta-b)^T X^T X(\beta-b)+(\beta-b_0)^T T_0(\beta-b_0)\right]\right\} \\
&\quad \tau^{(N-d+a)/2}exp\left\{-\frac{N\hat{\sigma}^2+r}{2}\tau\right\}. \qquad (8)
\end{aligned}$$

Ahora simplifiquemos el término dentro de la primera exponencial para que coincida con el kernel de una distribución normal.

$$\begin{aligned}
&(\beta-b)^T X^T X(\beta-b)+(\beta-b_0)^T T_0(\beta-b_0) \\
&= \beta^T X^T X\beta - \beta^T X^T Xb - b^T X^T X\beta + b^T X^T Xb + \\
&\quad \beta^T T_0\beta - \beta^T T_0 b_0 - b_0^T T_0\beta + b_0^T T_0 b_0
\end{aligned}$$

notando que todos estos términos son escalares de forma que sus transpuestos son ellos mismos, así como que $T_0^T = T_0$,

$$\begin{aligned}
&= \beta^T X^T X\beta - 2\beta^T X^T Xb + b^T X^T Xb + \beta^T T_0\beta - 2\beta^T T_0 b_0 + b_0^T T_0 b_0 \\
&= \beta^T(X^T X + T_0)\beta - 2\beta^T X^T Xb - 2\beta^T T_0 b_0 + b^T X^T Xb + b_0^T T_0 b_0
\end{aligned}$$

definiendo $T_1 = X^T X + T_0 \quad$ y $\quad g(X,y) = b^T X^T Xb + b_0^T T_0 b_0$,

$$\begin{aligned}
&= \beta^T T_1\beta - 2\beta^T X^T Xb - 2\beta^T T_0 b_0 + g(X,y) \\
&= \beta^T T_1\beta - 2\beta^T\left[X^T Xb + T_0 b_0\right] + g(X,y)
\end{aligned}$$

definiendo $b_1 = T_1^{-1}(X^T Xb + T_0 b_0)$ y completando el cuadrado:

$$\begin{aligned}
&= \beta^T T_1\beta - 2\beta^T T_1 b_1 + g(X,y) \\
&= (\beta-b1)^T T_1(\beta-b_1) + g(X,y) - b_1^T T_1 b_1 . \qquad (9)
\end{aligned}$$

Con esta manipulación de términos, ya podemos tener la distribución posterior de $\beta|\tau$, sustituyendo (9) en (8), como una normal d-variada con media b_1 y precisión τT_1:

$$\begin{aligned}
f(\theta|y) &\propto \tau^{d/2}exp\left\{-\frac{\tau}{2}\left[(\beta-b1)^T T_1(\beta-b_1)\right]\right\} \\
&\quad \tau^{(N-d+a)/2}exp\left\{-\frac{N\hat{\sigma}^2+g(X,y)-b_1^T T_1 b_1+r}{2}\tau\right\}.
\end{aligned}$$

La nueva media $b_1 = T_1^{-1}(X^T Xb + T_0 b_0)$ puede verse como un promedio de las medias originales— la de la inicial y el estimador máximo verosímil— ponderadas por sus precisiones (**Congdon06**). La nueva precisión es simplemente la

suma de las precisiones originales.

Ahora debemos encontrar los nuevos parámetros de forma y tasa para la distribución posterior de τ. Igualando el exponente de τ en la última expresión a $a_1 - 1$, donde a_1 es el nuevo parámetro de forma, para satisfacer la representación de una distribución gamma se llega a que $a_1 = (N - d + a^\star)/2$. El nuevo parámetro de tasa r_1 requiere ser un poco más explícitos:

$$\begin{aligned} r_1 &= \frac{N\hat{\sigma}^2 + g(X,y) - b_1^T T_1 b_1 + r}{2} \\ &= \frac{(y - Xb)^T(y - Xb) + b^T X^T X b + b_0^T T_0 b_0 - b_1^T T_1 b_1 + r}{2}\,. \end{aligned}$$

Pero resulta que $(y - Xb)^T(y - Xb) + b^T X^T X b = y^T y$:

$$\begin{aligned} (y - Xb)^T(y - Xb) + b^T X^T X b &= y^T y - 2y^T X b + b^T X^T X b + b^T X^T X b \\ &= y^T y - 2y^T X b + 2b^T X^T X b \\ &= y^T y - 2b^T X^T X b + 2b^T X^T X b \\ &= y^T y\,. \end{aligned} \tag{10}$$

Por lo que, en realidad,

$$r_1 = \frac{y^T y + b_0^T T_0 b_0 - b_1^T T_1 b_1 + r}{2}\,.$$

Con esto tenemos que la actualización de las (7) nos llevan al siguiente modelo conjugado:

$$\begin{aligned} y|\theta &\sim N_N(X\beta, \sigma^2 \mathbb{I}_N) \quad \text{tal que} \quad \theta = (\beta, \sigma^2) \sim f(\beta, \sigma^2) \\ \beta|\tau &\sim N_d(b_0, \tau T_0) \qquad \tau \sim \Gamma\left(a_0 = \frac{a^\star - p}{2}, r_0 = \frac{r}{2}\right) \\ \beta|\tau, y &\sim N_p(b_1, \tau T_1) \qquad \tau|y \sim \Gamma\left(a_1, r_1\right) \end{aligned}$$

con $a^\star > d$, $r > 0$ y $B_0 = T_0^{-1}$ positiva definida y tal que

$$\begin{aligned} T_1 &= X^T X + T_0 \qquad b_1 = T_1^{-1}(X^T X b + T_0 b_0) = T_1^{-1}(X^T y + T_0 b_0), \\ a_1 &= \frac{N - d + a^\star}{2} \qquad r_1 = \frac{y^T y + b_0^T T_0 b_0 - b_1^T T_1 b_1 + r}{2} \end{aligned} \tag{11}$$

donde $b = (X^T X)^{-1} X^T y$ es el estimador máximo verosímil de β. Más aún, si en lugar de utilizar como distribución inicial una normal gamma de esta forma se utiliza la inicial de Jeffreys de (6), podemos utilizar estas expresiones para hacer la actualización— aprovechando el carácter que la inicial de Jeffreys tiene como límite de conjugadas— considerando $a^\star \to d$, $r \to 0$ y $B_0 = T_0^{-1} \to \mathbf{O}$, por lo que se tendrían:

$$T_1 = X^T X \qquad b1 = b \qquad a_1 = \frac{N}{2} \qquad r_1 = \frac{y^T y - b^T X^T X b}{2} = \frac{N\hat{\sigma}^2}{2}$$

donde la equivalencia del estimador máximo verosímil $\hat{\sigma}^2$ puede verificarse con (10).