Modelo de regresión lineal bayesiano

Alan Ledesma Arista.[1]

[1] alan.ledesma@bcrp.gob.pe

1. Inferencia bayesiana
2. Función de verosimilitud
3. Distribuciones *prior*
   - *Prior* no informativo
   - *Prior* conjugado
   - *Prior* independiente

Inferencia bayesiana

- El interés es conocer los coeficientes $\boldsymbol{\theta}$ (de un modelo) en base a las observaciones $\boldsymbol{\mathcal{Y}}$.
- En el enfoque bayesiano los coeficientes 'verdaderos' se entienden como variables aleatorias; por lo tanto, se estiman funciones de distribución en lugar de puntos.
- El teorema de Bayes se utiliza para 'actualizar' nuestras 'creencias' sobre θ:
  1. La distribución *prior* es $p(\boldsymbol{\theta})$. Éste refleja las creencias del investigados sobre $\boldsymbol{\theta}$ que se parametrizan en una función de densidad (PDF).
  2. La función de verosimilitud es $\mathcal{L}(\boldsymbol{\theta}) \equiv p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta})$. Éste describe la probabilidad de observar $\boldsymbol{\mathcal{Y}}$ condicional a $\boldsymbol{\theta}$.
  3. La distribución *posterior* es $p(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}})$. Esta medida es el interés fundamental del análisis bayesiano. Resume lo que aprendemos sobre $\boldsymbol{\theta}$ dados los datos. También se puede entender como la actualización de nuestras creencias una vez procesados los datos.

Inferencia bayesiana

- **Teorema de Bayes**:

$$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}) = \frac{p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{\mathcal{Y}})}$$

como la inferencia se hace sobre $\boldsymbol{\theta}$ de denominador puede tratarse como un escalar. Por lo tanto

$$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}) \propto p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \equiv \mathcal{L}(\boldsymbol{\theta})p(\boldsymbol{\theta}) \qquad (1)$$

donde $\propto$ se lee como "proporcional a".

Función de verosimilitud

- El modelo de regresión lineal es

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

donde $\mathbf{y}$, $\mathbf{X}$ y $\boldsymbol{\varepsilon}$ son $n \times 1$, $n \times k$ y $n \times 1$, respectivamente.

- Si $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_{n,1}, \sigma^2 \mathbf{I}_n)$ y $\mathbf{X}$ es exógeno entonces $\hat{\mathbf{y}} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$; por lo tanto,

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2}\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2}\right].$$

- Recordar que los estimadores MCO de $\boldsymbol{\beta}$ y σ^2 son

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{and} \quad s^2 = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})}{n-k},$$

- Entonces, la función de verosimilitud se reduce a:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2}\frac{(n-k)s^2 + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})}{\sigma^2}\right] \tag{3}$$

Función de verosimilitud

Para llegar a la expresión (3) se utilizó el siguiente resultado

Álgebra recurrente:

Note que $(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)$ se puede reducir a

$$\begin{aligned}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta) =&(\mathbf{y}-\mathbf{Xb}+\mathbf{Xb}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{Xb}+\mathbf{Xb}-\mathbf{X}\beta)\\ =&[(\mathbf{y}-\mathbf{Xb})-\mathbf{X}(\beta-\mathbf{b})]'[(\mathbf{y}-\mathbf{Xb})-\mathbf{X}(\beta-\mathbf{b})]\\ =&(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})-(\mathbf{y}-\mathbf{Xb})'\mathbf{X}(\beta-\mathbf{b})-(\beta-\mathbf{b})'\mathbf{X}'(\mathbf{y}-\mathbf{Xb})\\ &+(\beta-\mathbf{b})'\mathbf{X}'\mathbf{X}(\beta-\mathbf{b})\\ =&(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})+(\beta-\mathbf{b})'\mathbf{X}'\mathbf{X}(\beta-\mathbf{b})\\ =&(n-k)s^2+(\beta-\mathbf{b})'\mathbf{X}'\mathbf{X}(\beta-\mathbf{b})\end{aligned}$$

la penúltima igualdad se deriva del siguiente resultado conocido
$\mathbf{X}'(\mathbf{y}-\mathbf{Xb})=\mathbf{X}'(\mathbf{y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})=\mathbf{X}'(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}=(\mathbf{X}'-\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}=\mathbf{0}$.

Prior no informativo

- Si no hay información *a priori* sobre $(\boldsymbol{\beta}, \sigma^2)$ se configura un *prior* no informativo.
- Un *prior* no informativo es $p(\boldsymbol{\beta}, \sigma^2) = 1$.
- El *prior* no altera la información en la función de verosimilitud. Por tanto, la distribución posterior es

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto \mathcal{L}(\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) = \mathcal{L}(\boldsymbol{\beta}, \sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2}\frac{(n-k)s^2 + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})}{\sigma^2}\right]. \end{aligned} \tag{4}$$

- Como resultado, la distribución posterior de $\boldsymbol{\beta}$ dado σ^2 es

$$p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{b})'[(\sigma^2)^{-1}(\mathbf{X}'\mathbf{X})](\boldsymbol{\beta}-\mathbf{b})\right]. \tag{5}$$

de donde $\boxed{\boldsymbol{\beta}|\sigma^2, \mathbf{y} \sim \mathcal{N}(\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})}$

- La densidad posterior marginal de σ^2 es

$$p(\sigma^2|\mathbf{y}) = \int_0^\infty p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})\mathrm{d}\boldsymbol{\beta} \propto \frac{1}{(\sigma^2)^{(n-k-2)/2+1}} \exp\left[-\frac{((n-k)/2)s^2}{\sigma^2}\right]. \tag{6}$$

entonces $\boxed{\sigma^2|\mathbf{y} \sim \Gamma^{-1}\left(\frac{n-k-2}{2}, \frac{n-k}{2}s^2\right)}$

Prior conjugado

- Un $p(\boldsymbol{\theta})$ anterior se conoce como "*prior* conjugado" de la verosimilitud $p(\mathbf{y}|\boldsymbol{\theta})$ si la distribución posterior $p(\boldsymbol{\theta}|\mathbf{y})$ está en la misma familia de distribución de probabilidad que la densidad *prior*.
- Fijemos las siguientes distribuciones *prior*: $\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(\underline{\mathbf{b}}, \sigma^2\underline{\mathbf{Q}})$ y $\sigma^2 \sim \Gamma^{-1}\left(\frac{\underline{\nu}-k-2}{2}, \frac{\underline{\eta}}{2}\right)$. I.e.,:

$$p(\boldsymbol{\beta}|\sigma^2) \propto \frac{1}{(\sigma^2)^{k/2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\underline{\mathbf{b}})'\underline{\mathbf{Q}}^{-1}(\boldsymbol{\beta}-\underline{\mathbf{b}})\right] \text{ y} \tag{7}$$

$$p(\sigma^2) \propto \frac{1}{(\sigma^2)^{(\underline{\nu}-k)/2}} \exp\left[-\frac{\underline{\eta}}{2\sigma^2}\right] \tag{8}$$

tal que

$$p(\boldsymbol{\beta},\sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \propto \frac{1}{(\sigma^2)^{\underline{\nu}/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{\eta} + (\boldsymbol{\beta}-\underline{\mathbf{b}})'\underline{\mathbf{Q}}^{-1}(\boldsymbol{\beta}-\underline{\mathbf{b}})\right)\right] \tag{9}$$

El *prior* en (9) se conoce como el *prior* normal-gamma

Prior conjugado

- La distribución posterior se calcula como sigue

$$
\begin{aligned}
p(\boldsymbol{\beta},\sigma^2|\mathbf{y}) \propto & p(\boldsymbol{\beta},\sigma^2)\mathcal{L}(\boldsymbol{\beta},\sigma^2) \\
\propto & \frac{1}{(\sigma^2)^{\underline{\nu}/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{\eta} + (\boldsymbol{\beta}-\underline{\mathbf{b}})'\underline{\mathbf{Q}}^{-1}(\boldsymbol{\beta}-\underline{\mathbf{b}})\right)\right] \\
& \times \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\frac{(n-k)s^2 + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})}{\sigma^2}\right] \\
\propto & \frac{1}{(\sigma^2)^{(\underline{\nu}+n)/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{\eta} + (n-k)s^2\right.\right. \\
& \left.\left. + (\boldsymbol{\beta}-\underline{\mathbf{b}})'\underline{\mathbf{Q}}^{-1}(\boldsymbol{\beta}-\underline{\mathbf{b}}) + (\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})\right)\right]
\end{aligned} \tag{10}
$$

Definamos

$$
\begin{aligned}
& \overline{\mathbf{Q}} = \left(\underline{\mathbf{Q}}^{-1} + \mathbf{X}'\mathbf{X}\right)^{-1}, \quad \overline{\mathbf{b}} = \overline{\mathbf{Q}}\left(\underline{\mathbf{Q}}^{-1}\underline{\mathbf{b}} + \mathbf{X}'\mathbf{X}\mathbf{b}\right), \\
& \overline{\nu} = \underline{\nu} + n \text{ and } \overline{\eta} = \underline{\eta} + (n-k)s^2 + (\mathbf{b}-\underline{\mathbf{b}})'\left(\underline{\mathbf{Q}} + (\mathbf{X}'\mathbf{X})^{-1}\right)^{-1}(\mathbf{b}-\underline{\mathbf{b}})
\end{aligned} \tag{11}
$$

Prior conjugado

- Con las definiciones (11) la expresión (10) se reduce a

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto \frac{1}{(\sigma^2)^{\overline{\nu}/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\overline{\eta} + (\boldsymbol{\beta} - \overline{\mathbf{b}})'\overline{\mathbf{Q}}^{-1}(\boldsymbol{\beta} - \overline{\mathbf{b}})\right)\right], \\
&\propto \left\{\frac{1}{(\sigma^2)^{k/2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \overline{\mathbf{b}})'\overline{\mathbf{Q}}^{-1}(\boldsymbol{\beta} - \overline{\mathbf{b}})\right]\right\} \\
&\quad \times \left\{\frac{1}{(\sigma^2)^{(\overline{\nu}-k)/2}} \exp\left[-\frac{\overline{\eta}}{2\sigma^2}\right]\right\}
\end{aligned}
\tag{12}
$$

esto es, $\boxed{\boldsymbol{\beta}|\sigma^2, \mathbf{y} \sim \mathcal{N}\left(\overline{\mathbf{b}}, \sigma^2\overline{\mathbf{Q}}\right)}$ y $\boxed{\sigma^2|\mathbf{y} \sim \Gamma^{-1}\left(\frac{\nu-k-2}{2}, \frac{\overline{\eta}}{2}\right)}$.

Álgebra recurrente:

La expresión $(\boldsymbol{\beta} - \underline{\mathbf{b}})'\underline{\mathbf{Q}}^{-1}(\boldsymbol{\beta} - \underline{\mathbf{b}}) + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})$ en (10) se puede demostrar que se reduce a

$$
\begin{aligned}
&(\boldsymbol{\beta} - \underline{\mathbf{b}})'\underline{\mathbf{Q}}^{-1}(\boldsymbol{\beta} - \underline{\mathbf{b}}) + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \\
&= (\boldsymbol{\beta} - \overline{\mathbf{b}})'\overline{\mathbf{Q}}^{-1}(\boldsymbol{\beta} - \overline{\mathbf{b}}) + (\mathbf{b} - \underline{\mathbf{b}})'\left((\mathbf{X}'\mathbf{X})^{-1} + \underline{\mathbf{Q}}\right)^{-1}(\mathbf{b} - \underline{\mathbf{b}})
\end{aligned}
$$

Prior independiente

- A veces es conveniente formular distribuciones a priori independientes entre coeficientes
- El algoritmo para simular la distribución posterior dependerá de su forma resultante:
  - Si el conjunto completo de distribuciones posteriores condicionales es fácil de simular: Gibbs sampling
  - De lo contrario: Metropolis-Hastings

Conjunto completo de distribuciones posteriores condicionales

- Si las *priors* independientes se establecen como $\boldsymbol{\beta} \sim \mathcal{N}(\underline{\mathbf{b}}, \underline{\mathbf{V}})$ y $\sigma^2 \sim \Gamma^{-1}\left(\frac{\underline{a}}{2}, \frac{\underline{b}}{2}\right)$ entonces la distribución *prior* es

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &\propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \underline{\mathbf{b}})'\underline{\mathbf{V}}^{-1}(\boldsymbol{\beta} - \underline{\mathbf{b}})\right] \frac{1}{(\sigma^2)^{\frac{\underline{a}+2}{2}}} \exp\left[-\frac{\underline{b}}{2\sigma^2}\right], \\ &\propto \frac{1}{(\sigma^2)^{\frac{\underline{a}+2}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{b} + \sigma^2(\boldsymbol{\beta} - \underline{\mathbf{b}})'\underline{\mathbf{V}}^{-1}(\boldsymbol{\beta} - \underline{\mathbf{b}})\right)\right] \end{aligned} \tag{13}$$

Bajo esta especificación, no hay expresión cerrada para la distribución posterior

Prior independiente

- La distribución posterior es

$$
\begin{aligned}
p(\boldsymbol{\beta},\sigma^2|\mathbf{y}) \propto & \mathcal{L}(\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta},\sigma^2),\\
& \propto \frac{1}{(\sigma^2)^{-n/2}} \exp\left[-\frac{1}{2\sigma^2}\left((n-k)s^2+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})\right)\right]\\
& \quad \times \frac{1}{(\sigma^2)^{\frac{\underline{a}+2}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{b}+\sigma^2(\boldsymbol{\beta}-\underline{\mathbf{b}})'\underline{\mathbf{V}}^{-1}(\boldsymbol{\beta}-\underline{\mathbf{b}})\right)\right],\\
& \propto \frac{1}{(\sigma^2)^{\frac{\underline{a}-n+2}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{b}+(n-k)s^2+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})+\sigma^2(\boldsymbol{\beta}-\underline{\mathbf{b}})'\underline{\mathbf{V}}^{-1}(\boldsymbol{\beta}\right.\right.
\end{aligned}
\tag{14}
$$

- De donde si σ^2 se toma como dado

$$p(\boldsymbol{\beta}|\sigma^2,\mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma^2}\left((\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})+\sigma^2(\boldsymbol{\beta}-\underline{\mathbf{b}})'\underline{\mathbf{V}}^{-1}(\boldsymbol{\beta}-\underline{\mathbf{b}})\right)\right] \tag{15}$$

definir $\overline{\mathbf{V}}=(\sigma^2\underline{\mathbf{V}}^{-1}+\mathbf{X}'\mathbf{X})^{-1}$ y $\overline{\mathbf{b}}=\overline{\mathbf{V}}(\sigma^2\underline{\mathbf{V}}^{-1}\underline{\mathbf{b}}+\mathbf{X}'\mathbf{y})$, entonces

$$p(\boldsymbol{\beta}|\sigma^2,\mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\overline{\mathbf{b}})'\overline{\mathbf{V}}^{-1}(\boldsymbol{\beta}-\overline{\mathbf{b}})\right] \tag{16}$$

por lo tanto, $\boxed{\boldsymbol{\beta}|\sigma^2,\mathbf{y}\sim\mathcal{N}(\overline{\mathbf{b}},\sigma^2\overline{\mathbf{V}})}$

Prior independiente

- En cambio, si $\boldsymbol{\beta}$ se toma como dado

$$p(\sigma^2|\boldsymbol{\beta},\mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{\underline{a}-n}{2}+1}} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{b}+(n-k)s^2+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})\right)\right]$$

defina $\overline{a}=\underline{a}-n$ y $\overline{b}=\underline{b}+(n-k)s^2+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})$; por lo tanto,

$$p(\sigma^2|\boldsymbol{\beta},\mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{\overline{a}}{2}+1}} \exp\left[-\frac{\overline{b}/2}{\sigma^2}\right] \tag{17}$$

Consecuentemente, $\boxed{\sigma^2|\boldsymbol{\beta},\mathbf{y} \sim \Gamma^{-1}\left(\frac{\overline{a}}{2},\frac{\overline{b}}{2}\right)}$

- The set of conditional distributions is complete: Gibbs-Sampling

Prior independiente

Un prior más general

- En un caso más general, la parte posterior no toma una forma conocida

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto & \mathcal{L}(\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2), \\
& \propto \frac{1}{(\sigma^2)^{-n/2}} \exp \left[-\frac{1}{2\sigma^2} \left((n-k)s^2 + (\boldsymbol{\beta} - \mathbf{b})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \mathbf{b}) \right) \right] p(\boldsymbol{\beta}) p(\sigma^2). \quad (18)
\end{aligned}
$$

- Metropolis-Hasting