Introducción a Ciencia de Datos

Tarea 3 - Regresión Lineal y Bayesiana

Debany Jazmín Hernández Camacho debany.hernandez@cimat.mx
Eric Ernesto Moreles Abonce eric.moreles@cimat.mx
Luis Erick Palomino Galván luis.palomino@cimat.mx

1. Regresión lineal ordinaria (OLS)

Ejercicio 1 (Derivación del estimador OLS). *Partiendo del modelo clásico*

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

demuestre que el estimador de Mínimos Cuadrados Ordinarios es

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

siempre que $X^\top X$ sea invertible.

Solución. Queremos obtener el estimador de Mínimos Cuadrados Ordinarios (OLS) para el modelo lineal

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

El objetivo es encontrar el vector $\hat{\beta}$ que minimiza la suma de cuadrados de los errores,

$$S(\beta) = (y - X\beta)^\top (y - X\beta).$$

Primero, vamos a hacer la expansión de la función de pérdida. Desarrollamos el producto,

$$S(\beta) = y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X \beta.$$

Para encontrar el punto mínimo, derivamos $S(\beta)$ con respecto a β y la igualamos a cero,

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^\top y + 2X^\top X\beta = 0.$$

De la ecuación anterior se obtiene el sistema,

$$X^\top X\, \hat{\beta} = X^\top y.$$

Estas son las *ecuaciones normales*. Si la matriz $X^\top X$ es invertible, podemos despejar directamente,

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

El estimador $\hat{\beta}$ existe y es único solo cuando $X^\top X$ es invertible, es decir, cuando no hay colinealidad perfecta entre las variables explicativas. En caso contrario, puede obtenerse una solución de mínima norma usando la pseudoinversa de Moore–Penrose,

$$\hat{\beta} = X^+ y.$$

Ejercicio 2 (Propiedades del estimador). *Calcule explícitamente:*

$$\mathbb{E}\left[\hat{\beta}\right], \qquad \mathrm{Var}\left[\hat{\beta}\right].$$

Concluya que $\hat{\beta}$ es un estimador insesgado y eficiente dentro de la clase de estimadores lineales (Teorema de Gauss–Markov).

Solución. Recordemos el modelo lineal,

$$y = X\beta + \varepsilon, \qquad \mathbb{E}\left[\varepsilon\right] = 0, \quad \mathrm{Var}\left[\varepsilon\right] = \sigma^2 I_n,$$

y el estimador OLS obtenido anteriormente,

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

Sustituimos el modelo en el estimador, entonces reemplazamos $y = X\beta + \varepsilon$:

$$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1} X^\top (X\beta + \varepsilon) \\ &= (X^\top X)^{-1} X^\top X\beta + (X^\top X)^{-1} X^\top \varepsilon.\end{aligned}$$

Ahora, simplificando,

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon.$$

Calculamos la esperanza del estimador.

$$\mathbb{E}\left[\hat{\beta}\right] = \mathbb{E}\left[\beta + (X^\top X)^{-1} X^\top \varepsilon\right] = \beta + (X^\top X)^{-1} X^\top \mathbb{E}\left[\varepsilon\right].$$

Como $\mathbb{E}\left[\varepsilon\right] = 0$, obtenemos que,

$$\mathbb{E}\left[\hat{\beta}\right] = \beta.$$

Por lo tanto, $\hat{\beta}$ es un **estimador insesgado** de β.

Luego, calculamos la varianza del estimador, entonces usando la expresión de $\hat{\beta}$,

$$\mathrm{Var}\left[\hat{\beta}\right] = \mathrm{Var}\left[(X^\top X)^{-1} X^\top \varepsilon\right].$$

Dado que $\mathrm{Var}\left[\varepsilon\right] = \sigma^2 I_n$, y recordando que $\mathrm{Var}\left[A\varepsilon\right] = A\,\mathrm{Var}\left[\varepsilon\right]\, A^\top$:

$$\mathrm{Var}\left[\hat{\beta}\right] = (X^\top X)^{-1} X^\top (\sigma^2 I_n) X (X^\top X)^{-1}.$$

Entonces, simplificando obtenemos que,

$$\mathrm{Var}\left[\hat{\beta}\right] = \sigma^2 (X^\top X)^{-1}.$$

Teorema de Gauss-Markov. El estimador $\hat{\beta}$ es:

- **Lineal** en y (porque depende de y a través de una combinación lineal);
- **Insesgado**, pues $\mathbb{E}\left[\hat{\beta}\right] = \beta$;

- **De mínima varianza** dentro de la clase de estimadores lineales insesgados.

Por lo tanto, $\hat{\beta}$ es el mejor estimador lineal insesgado (BLUE, por sus siglas en inglés *Best Linear Unbiased Estimator*).

La matriz de varianza $\sigma^2(X^\top X)^{-1}$ muestra cómo la incertidumbre de los parámetros depende tanto del nivel de ruido σ^2 como de la información contenida en X. Cuanto más grande sea $X^\top X$, menor será la varianza de los estimadores. ■

2. Regresión lineal bayesiana (prior conjugado)

Ejercicio 1 (Prior conjugado). *Suponga un prior conjugado:*

$$\beta \mid \sigma^2 \sim N(\beta_0,\, \sigma^2 V_0), \qquad \sigma^2 \sim \textit{Inv-Gamma}(a_0,\, b_0).$$

Describa brevemente la interpretación de los hiperparámetros β_0, V_0, a_0, b_0 y su relación con la información previa del modelo.

Solución. En el enfoque bayesiano de la regresión lineal, el objetivo es incorporar información previa sobre los parámetros del modelo antes de observar los datos. Para ello se definen distribuciones a priori que reflejan nuestras creencias iniciales sobre los valores de los parámetros. En este caso se considera un modelo jerárquico donde

$$y \mid \beta, \sigma^2 \sim N(X\beta,\, \sigma^2 I_n),$$

y los priors se especifican de la siguiente manera:

$$\beta \mid \sigma^2 \sim N(\beta_0,\, \sigma^2 V_0), \qquad \sigma^2 \sim \text{Inv-Gamma}(a_0,\, b_0).$$

Este tipo de prior se llama *conjugado* porque, al combinarlo con la verosimilitud normal del modelo, la distribución posterior$p(\beta, \sigma^2 \mid y)$ pertenece a la misma familia de distribuciones Normal–Inversa Gamma. Esto permite obtener resultados analíticos y simplificar la inferencia, pues la forma funcional de la posterior se conserva.

El vector β_0 representa el valor medio que se espera para los coeficientes antes de observar los datos, mientras que la matriz V_0 determina la incertidumbre previa sobre ellos. Una matriz V_0 grande indica poca información previa y, por lo tanto, un prior más difuso. Por su parte, los parámetros a_0 y b_0 controlan la forma y la escala de la distribución de la varianza σ^2; en particular, valores grandes de a_0 implican mayor confianza en una varianza cercana a su media previa $b_0/(a_0 - 1)$.

De esta manera, el prior conjugado Normal–Inversa Gamma combina la verosimilitud normal del modelo con una estructura que mantiene la coherencia matemática y facilita la obtención de la distribución posterior. Además, al ajustar los hiperparámetros (β_0, V_0, a_0, b_0), es posible controlar cuánta influencia tiene la información previa frente a los datos observados. ■

Ejercicio 2 (Distribución posterior). *Derive los parámetros posteriores (β_n, V_n, a_n, b_n) y escriba la forma explícita de la posterior conjunta*

$$p(\beta, \sigma^2 \mid y).$$

Solución. Suponiendo un prior conjugado:

$$\beta|\sigma^2 \sim \mathcal{N}(\beta_0, \sigma^2 V_0), \qquad \sigma^2 \sim \text{Inv-Gamma}(a_0, b_0)$$

De la hipótesis anterior, tenemos que:

$$\begin{aligned}
p(y|X,\beta,\sigma^2) &= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right) \\
p(\beta|\sigma^2) &= (2\pi\sigma^2)^{-P/2} |V_0|^{1/2} \exp\left(-\frac{1}{2\sigma^2}(\eta - \mu_0)^T V_0(\beta - \mu_0)\right) \\
p(\sigma^2) &= \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)} \exp\left(-b_0/\sigma^2\right)
\end{aligned}$$

De lo anterior, tenemos una expresión para la posterior:

$$\begin{aligned}
p(\beta,\sigma^2|y,X) &\propto p(y|X,\beta,\sigma^2)p(\beta|\sigma^2)p(\sigma^2) \\
&\propto (\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)\dots \\
&\quad (\sigma^2)^{-P/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_0)^T V_0(\beta - \mu_0)\right)\dots \\
&\quad (\sigma^2)^{-(a_0+1)} \exp\left(-b_0/\sigma^2\right)
\end{aligned}$$

Esto ultimo es un producto, solo se separa en renglones por cuestiones de espacio. De las primeras dos exponenciales, podemos reacomodar los términos, tal que:

$$\begin{aligned}
&(y - X\beta)^T(y - X\beta) + (\beta - \mu_0)^T V_0(\beta - \mu_0) \\
&= (\beta - \mu_n)^T(X^TX + V_0)(\beta - \mu_n) + y^Ty - \mu_n^T(X^TX + V_0)\mu_n + \mu_0^T V_0\mu_0
\end{aligned}$$

Donde definimos:

$$\mu_n = (X^TX + V_0)^{-1}(X^Ty + V_0\mu_0)$$

Sustituyendo en nuestra posterior, podemos ver que:

$$\begin{aligned}
&p(\beta,\sigma^2|y,X) \\
&\qquad \propto (\sigma^2)^{-P/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_n)^T(X^TX + V_0)(\beta - \mu_n)\right)\dots \\
&\qquad (\sigma^2)^{-\frac{N+2a_0}{2}-1} \exp\left(-\frac{2b_0 + y^Ty - \mu_n^T(X^TX + V_0)\mu_n + \mu_0^T V_0\mu_0}{2\sigma^2}\right)
\end{aligned}$$

Lo anterior es un producto de dos densidades $\mathcal{N}(\mu_n, \sigma^2 V_n^{-1})$ y una Inv-Gamma(a_n, b_n) con parámetros:

$$\begin{aligned}
V_n &= X^TX + V_0 \\
\mu_n &= (V_n)^{-1}(X^Ty + V_0\mu_0) \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^Ty + \mu_0^T V_0\mu_0 - \mu_n^T V_n\mu_n)
\end{aligned}$$

Y concluimos. ■

Ejercicio 3 (Distribuciones marginales:). *Identifique las distribuciones marginales de β y de σ^2.*

Solución. Del inciso anterior tenemos que la marginal de σ^2 es una Inv-Gamma(a_n, b_n), pues β solo esta en la $\mathcal{N}(\mu_n, \sigma^2 V_n^{-1})$, y por lo tanto al integrar sobre β nos queda la densidad previamente dicha.

Para β, tenemos que calcular:

$$p(\beta \mid y) = \int_0^\infty p(\beta|\sigma^2, y)\, p(\sigma^2|y) d\sigma^2.$$

Sustituyendo las dos densidades:

$$\begin{aligned} p(\beta \mid y) &= (2\pi)^{-p/2}|V_n|^{-1/2} \frac{b_n^{a_n}}{\Gamma(a_n)} \int_0^\infty (\sigma^2)^{-a_n-1-p/2} \dots \\ &\quad \exp\left(-\frac{1}{2\sigma^2}\left[(\beta-\beta_n)^\top V_n^{-1}(\beta-\beta_n) + 2b_n\right]\right) d\sigma^2 \\ &= (2\pi)^{-p/2}|V_n|^{-1/2} \frac{b_n^{a_n}}{\Gamma(a_n)} \int_0^\infty (\sigma^2)^{-c-1} \exp\Big(-\frac{d}{\sigma^2}\Big) d\sigma^2 \end{aligned}$$

Donde:

$$\begin{aligned} c &= a_n + \frac{p}{2} \\ d &= \frac{1}{2}\left[(\beta-\beta_n)^\top V_n^{-1}(\beta-\beta_n) + 2b_n\right] \\ \beta_n &= V_n\left(V_0^{-1}\beta_0 + X^T y\right) \end{aligned}$$

Usando la definición de la función Gamma, tenemos que:

$$\int_0^\infty (\sigma^2)^{-c-1} \exp\left(-\frac{d}{\sigma^2}\right) d\sigma^2 = \frac{\Gamma(c)}{d^c}$$

Sustituyendo:

$$\begin{aligned} p(\beta \mid y) &= (2\pi)^{-p/2}|V_n|^{-1/2} \frac{b_n^{a_n}}{\Gamma(a_n)} \cdot \frac{\Gamma\left(a_n + \frac{p}{2}\right)}{\left(\frac{1}{2}\left[(\beta-\beta_n)^\top V_n^{-1}(\beta-\beta_n) + 2b_n\right]\right)^{a_n+\frac{p}{2}}} \\ &= \frac{\Gamma\left(a_n + \frac{p}{2}\right)}{\Gamma(a_n)} \frac{b_n^{a_n}}{(\pi)^{p/2}|V_n|^{1/2}} \left((\beta-\beta_n)^T V_n^{-1}(\beta-\beta_n) + 2b_n\right)^{-\left(a_n+\frac{p}{2}\right)}. \end{aligned}$$

Factorizando $2b_n$ dentro del paréntesis:

$$(\beta-\beta_n)^T V_n^{-1}(\beta-\beta_n) + 2b_n = 2b_n\left[1 + \frac{1}{2b_n}(\beta-\beta_n)^T V_n^{-1}(\beta-\beta_n)\right]$$

Y reescribiendo constantes llegamos a la forma estándar de la densidad t de student multivariada. En particular, definiendo la matriz de escala

$$\Sigma_t = \frac{b_n}{a_n} V_n,$$

y los grados de libertad $\nu = 2a_n$, la densidad toma la forma conocida. Por tanto:

$$\beta|y \sim t_\nu(\beta_n, \Sigma_t)$$

■

3. Conexión con regularización

Ejercicio 1 (**Regresión Ridge:**). *Muestre que si se toma un prior Normal isotrópico*

$$\beta \sim \mathcal{N}(0, \tau^2 I)$$

el estimador de máxima a posteriori (MAP) es equivalente a la regresión Ridge:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|^2, \qquad \lambda = \frac{\sigma^2}{\tau^2}.$$

Solución. Consideremos ahora el prior Normal isotrópico

$$\beta \sim \mathcal{N}(0, \tau^2 I_p).$$

Queremos encontrar el estimador $\hat{\beta}_{MAP}$, es decir

$$\hat{\beta}_{MAP} = \arg\max_{\beta} p(\beta|y, \sigma^2).$$

Fijando σ^2, la log-posterior en función de β es, salvo constantes independientes de β:

$$\log p(\beta|y, \sigma^2) = -\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \frac{1}{2\tau^2}\|\beta\|^2 + \text{const}$$

Maximizar la posterior es equivalente a minimizar la expresión negativa, si al mismo tiempo multiplicamos por $2\sigma^2$, tenemos que:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\right\}$$

Con $\lambda = \sigma^2/\tau^2$.

Esta es exactamente la forma de la función objetivo de la **regresión Ridge**, donde el término $\lambda\|\beta\|^2$ actúa como penalización cuadrática.

El parámetro λ controla la magnitud de la regularización: valores grandes de λ fuerzan a los coeficientes hacia cero, reduciendo la varianza del estimador a costa de introducir un sesgo pequeño. Cuando $\lambda \to 0$, el estimador se aproxima al de mínimos cuadrados ordinarios (OLS), mientras que cuando λ crece, los coeficientes se "encogen" progresivamente hacia el origen. Esta propiedad de contracción estabiliza el ajuste cuando existe multicolinealidad o el número de predictores es grande en comparación con el tamaño de muestra.

De forma cerrada, el estimador Ridge se obtiene derivando la función de pérdida y resolviendo las ecuaciones normales modificadas:

$$(X^\top X + \lambda I_p)\hat{\beta}_{Ridge} = X^\top y, \qquad \hat{\beta}_{Ridge} = (X^\top X + \lambda I_p)^{-1} X^\top y.$$

Este estimador siempre existe y es único, aun cuando $X^\top X$ no sea invertible, gracias al término λI_p que garantiza la invertibilidad de la matriz.

El vínculo entre el prior bayesiano normal y la penalización de Ridge muestra cómo la regularización frecuentista puede interpretarse como la incorporación de información previa que favorece coeficientes pequeños. Desde esta perspectiva, Ridge surge naturalmente como el estimador MAP bajo un prior normal centrado en cero. El prior

induce una distribución posterior más concentrada, reduciendo la varianza de los parámetros estimados y proporcionando un ajuste más estable frente a colinealidad o ruido en los datos.

En síntesis, la regresión Ridge representa una versión suavizada de OLS en la que se equilibra el ajuste del modelo con la complejidad de los parámetros. El enfoque bayesiano nos permite entender este método como un proceso de inferencia posterior bajo un prior normal, mientras que la interpretación frecuentista lo ve como un método de penalización que previene el sobreajuste.

Ambas perspectivas coinciden en que el término de regularización λ desempeña un papel central al controlar la magnitud de los coeficientes y, por lo tanto, la capacidad de generalización del modelo. ■

Ejercicio 2 (Regresión Lasso). *Muestra que si en lugar de un prior Normal se utiliza un prior Laplace (doble-exponencial)*

$$p(\beta_j) \sim Exp(-\lambda \mid \beta_j),$$

el estimador MAP corresponde a la regresión Lasso:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left(\|y - X\beta\|^2 + \lambda \|\beta\|_1 \right).$$

Solución. Partimos del modelo lineal clásico

$$y \mid \beta, \sigma^2 \sim \mathcal{N}(X\beta,\, \sigma^2 I_n),$$

donde $y \in \mathbb{R}^n$ es el vector de observaciones, $X \in \mathbb{R}^{n \times p}$ la matriz de diseño y $\beta \in \mathbb{R}^p$ el vector de parámetros desconocidos. Suponemos ahora que los coeficientes β_j son a priori independientes y siguen una distribución Laplace (o doble exponencial), con densidad

$$p(\beta_j \mid \lambda) = \frac{\lambda}{2} \exp(-\lambda |\beta_j|), \qquad \lambda > 0.$$

Esta prior se caracteriza por tener una forma aguda en el origen, lo que induce una tendencia a concentrar los valores de los coeficientes cerca de cero, promoviendo así la selección automática de variables.

La verosimilitud del modelo, condicionada a β y σ^2, es proporcional a

$$p(y \mid \beta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} \|y - X\beta\|^2 \right).$$

Combinando la verosimilitud con la prior Laplace, la distribución posterior de β (ignorando constantes) se puede escribir como

$$p(\beta \mid y, \sigma^2, \lambda) \propto \exp\left(-\frac{1}{2\sigma^2} \|y - X\beta\|^2 - \lambda \|\beta\|_1 \right),$$

donde $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

El estimador de máxima a posteriori (MAP) se obtiene maximizando la posterior con respecto a β. Equivalente a esto, se puede formular el problema como la minimización del negativo del logaritmo de la posterior, de donde se obtiene

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left\{ \frac{1}{2\sigma^2} \|y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}.$$

Si multiplicamos todo por $2\sigma^2$ (sin alterar el punto que minimiza), llegamos a la siguiente forma

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda^* \|\beta\|_1 \right\},$$

donde $\lambda^* = 2\sigma^2\lambda$.

Esta expresión es exactamente la función objetivo de la regresión Lasso. Por lo tanto, un prior Laplace independiente sobre los coeficientes β_j conduce a un estimador MAP equivalente al de la regresión Lasso, con un parámetro de regularización que depende de λ y σ^2.

Este resultado muestra una relación directa entre los métodos bayesianos y las técnicas de regularización frecuentistas. En el caso del prior Laplace, la fuerte concentración de masa en torno a cero genera un efecto de "encogimiento" sobre los coeficientes pequeños, algunos de los cuales se estiman exactamente como cero. En consecuencia, el estimador MAP no solo controla la magnitud de los parámetros para evitar el sobreajuste, sino que también realiza selección automática de variables.

En el caso especial en que las columnas de X están ortonormalizadas ($X^\top X = I_p$), el problema tiene una solución analítica componente a componente mediante la función de umbralización suave (soft–thresholding):

$$\hat{\beta}_j = \text{sign}\big((X^\top y)_j\big)\ \text{máx}\left\{|(X^\top y)_j| - \lambda_{\text{std}},\, 0\right\}.$$

Este resultado ilustra de manera intuitiva el efecto del prior Laplace: los coeficientes pequeños son reducidos a cero, mientras que los de mayor magnitud solo se reducen parcialmente. En síntesis, la regresión Lasso surge naturalmente como el estimador MAP asociado a un prior Laplace sobre los parámetros, estableciendo un puente claro entre la inferencia bayesiana y las técnicas de regularización penalizada.

4. Extensiones: errores no normales

Ejercicio 1 (Modelos alternativos). *Proponga un modelo de regresión donde el error ϵ no siga una distribución Normal.*

Solución. En regresión lineal, usamos el método de mínimos cuadrados para minimizar el error cuadrático. Al considerar que el error tiene una distribución normal, implica que los errores grandes son poco probables, por lo que es similar a usar regresión lineal.
Cambiar la distribución del error es cambiar la forma de la función de verosimilitud, lo que nos permite construir modelos robustos. Por ejemplo, si consideramos la distribución **Laplace** que tiene colas más pasada, entonces los errores grandes son menos plausibles.

Recordemos que para la distribución **T-Student**, elegir los grados de libertad es poder elegir el grosor de las colas de la distribución, lo que permite un modelo más justo ante los outliers. Por ello, el modelo que vamos a proponer va a ser considerando el error con está distribución.

Sean $y_i, \ldots, y_n$ las variables dependientes, $x_i = (1, X_{i1}, \ldots, X_{im})$ y $\beta = (\beta_0, \ldots, \beta_m)$. Consideremos el modelo de regresión lineal:

$$y_i = x_i^T \beta + \varepsilon_i.$$

Donde, ahora el error sigue una distribución $\varepsilon_i \sim T - Student(\nu, 0, \sigma^2)$ donde μ es el parámetro de grados de libertad, los errores están centrados en cero y σ es el parámetro de escala. El parámetro ν, controla el grosor de

las colas de la distribución, un valor bajo ν (1-10) produce colas muy pesadas, lo que permite al modelo acomodar outliers sin que estos influyan excesivamente en la estimación lineal de regresión. Por oto lado, si $\nu \to \infty$, la distribución T-Sudent converge a una distribución normal. Para ajustar este modelo a uno bayesiano, necesitamos especificar la verosimilitud y los prior para los parámetros β, σ y ν.

Notemos que la probabilidad de observar y_i dado los parámetros se basa directamente en la función de densidad de probabilidad de la distribución T-Student, tal que

$$p\left(y_i|\beta,\sigma,\nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{1}{\nu}\left(\frac{y_i - x_i^T\beta}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}.$$

Por lo que la verosimilitud del modelo para los n datos es

$$L(\beta,\sigma^2,\nu) = \prod_{i=1}^{n} p\left(y_i|\beta,\sigma,\nu\right).$$

Para establecer distribuciones a priori para los parámetros, vamos a suponer que no tenemos conocimiento previo de ellos, por lo que buscamos una distribución que no favorezca a ningún valor en particular. Así, una elección de prioris podría ser:

1. Para los coeficientes de β: Dado que suponemos que los valores de los coeficientes son probablemente cercanos a cero, vamos a tomar un prior normal $\beta_j \sim N(0,\sigma_\beta^2)$ donde σ_β^2 es una varianza grande.
2. Para el parámetro de escala σ: queremos asegurar que la escala sea positiva y no esté demasiado restringida, por lo que tomaremos un prior cauchy $\sigma \sim Cauchy(0,\gamma)$ para $\gamma > 0$.
3. Para los grados de libertad ν: necesitamos un prior flexible para establecer un punto de partida razonable, dejando que los datos muevan la estimación de μ. Una posible distribución prior de μ es $Gamma(\alpha,\beta)$, ya los parámetros de forma α y taza β, permiten modelar la forma prior de maneras distintas. Así, si se tiene la sospecha de que podría haber outliers, podríamos elegir parámetros que pongan más peso en valores bajos de ν; pero si se cree que los datos son limpios, podríamos tomar parámetros para que la distribución Gamma tenga un pico en valores más altos.

Considerando lo anterior, el parámetro μ permite al modelo adaptarse a la naturaleza de los errores en los datos. Lo que hace un modelo más robusto ante outliers. Además, debido a que este modelo no tiene prior conjugados, la inferencia se puede hacer usando métodos computacionales como MCMC. ■

Ejercicio 2 (Consecuencias metodológicas). *Explique cuáles serían las consecuencias sobre*

1. *La forma de la verosimilitud.*

   Solución. Recordemos que la verosimilitud es la probabilidad de haber observado los datos dado una linea de regresión. Si consideramos los errores normales, por lo que la verosimilitud es

$$L(\beta,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right].$$

Es decir, la verosimilitud será una función Gaussiana, por lo que los valores atípicos tendrán una influencia mayor ya que el residuo grande se eleva al cuadrado y resulta en una verosimilitud baja para esa linea candidata. Por otro lado, si consideramos un errar con otra distribución (Laplace o T-Student), la verosimilitud tendrá una forma con colas más pesadas, por lo que la función no penalizará los residuos grandes de la misma forma que la Gaussiana, lo que permite a el modelo ajustarse a la mayoría de los datos.

Por lo que la función de verosimilitud afecta directamente la estructura de la distribución del error y las propiedades del estimador. La forma de la verosimilitud nos dice la suposición sobre la proporción de valores atípicos. ■

2. *La existencia o no de priors conjugados.*

Solución. Recordemos que un prior conjugado es una distribución de probabilidad que al ser combinada con una función de verosimilitud, obtenemos una distribución posterior que pertenece a la misma familia de distribuciones que el prior. Si tenemos un caso con conjugación como un error normal, la verosimilitud Gaussiana tiene un prior conjugado que es la distribución Normal-Gamma Inversa, que al ser combinadas, la distribución posterior resultante pertenece a la misma familia. Sin embargo, sin conjugación, al cambiar la verosimilitud (como T-Student o Laplace), el prior Normal-Gamma Inversa ya no es conjugado para estas verosimilitudes, por lo que ya no se puede derivar la distribución posterior de forma analítica. ■

3. *Los métodos de inferencias requeridos (MCMC, aproximación variacional, etc.).*

Solución. Si trabajamos con prior no conjugado, dado que no se puede calcular el posterior de manera analítica, se debe de aproximar con métodos computacionales. Dos de estos metodos son:

   a) Markov Chain Monte Carlo (MCMC). Esté método, genera miles de muestras de la distribución posterior, dándonos un mapa muy detallado de ella. El método es muy preciso, pero el coste computacional es muy alto, ya que debe dar millones de pasos y para modelos grandes o con muchos datos puede tardar días.

   b) Inferencia Variacional (VI). Este método, intenta ajustar una distribución simple como la Normal a la forma del posterior verdadero. Es una aproximación que sacrifica exactitud por velocidad, es mucho más rápida que MCMC, por lo que se suele usar cuando el número de observaciones es muy grande.

■