Alumno: Bueno Rivera Oswaldo
Alumno: Rodríguez Villagrán Juan Pablo | **Profesor:** Marco Antonio Aquino López

Regresión lineal y Bayesiana

---

1. ***Regresión lineal ordinaria (OLS).***

a) ***Derivación del estimador OLS.*** **Partiendo del modelo clásico**

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

demuestra que el estimador de mínimos cuadrados ordinarios es

$$\widehat{\beta} = (X^T X)^{-1} X^T y,$$

siempre que $X^T X$ sea invertible.

Para demostrar que dicho $\widehat{\beta}$ es el estimador de mínimos cuadrados, debemos demostrar que minimiza el error cuadrático. Consideremos la suma de los cuadrados de los errores residuales, la cual está dada por

$$\begin{aligned} SS_{Res}(\beta) &= (y - X\beta)^T (y - X\beta) \\ &= y^T y - y^T X\beta - \beta^T X^T y - \beta^T X^T X\beta \\ &= y^T y - 2y^T X\beta - \beta^T X^T X\beta. \end{aligned}$$

Derivamos lo anterior con respecto a β, entonces, teniendo en consideración que $X^T X$ es una matriz invertible, obtenemos que

$$\frac{\partial SS_{Res}(\beta)}{\partial \beta} = (-2y^T X)^T + 2X^T X\beta = -2X^T y + 2X^T X\beta.$$

Los puntos críticos respecto a β de SS_{Res} los encontramos igualando la derivada parcial a cero y resolviendo para β. Tras esto obtenemos que el punto crítico de SS_{Res} es

$$\widehat{\beta} = (X^T X)^{-1} X^T y,$$

cuyo despeje es posible gracias a la invertibilidad de $X^T X$. Verifiquemos que $\widehat{\beta}$ es en efecto un mínimo. Para esto último, utilizamos el criterio de la segunda derivada, así debemos analizar la matriz Hessiana de SS_{Res}, la cual está dada por

$$H = \frac{\partial^2 SS_{Res}(\beta)}{\partial \beta^2} = (2X^T X)^T = 2X^T X.$$

Esto último es una matriz positiva definida ya que $X^T X$ es invertible y por lo tanto X es de rango completo. Al tener que la matriz Hessiana es positiva definida concluimos que $\widehat{\beta} = (X^T X)^{-1} X^T y$ minimiza a SS_{Res}, es decir, que es el estimador de mínimos cuadrados de β.

El estimador $X\widehat{\beta}$ tiene la interpretación geométrica de que proyecta a las observaciones y sobre el subespacio generado por las columnas de X.

b) *Propiedades del estimador.* Calcula explícitamente

$$\mathbb{E}[\widehat{\beta}] \qquad \textbf{y} \qquad \textbf{Var}(\widehat{\beta}).$$

Concluya que $\widehat{\beta}$ es insesgado y eficiente dentro de la clase de estimadores lineales (teorema de Gauss-Markov).

Teniendo en cuenta que la única aleatoriedad del estimador $\widehat{\beta}$ recae en y, ya que $y = X\beta + \varepsilon$, obtenemos directamente que

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}\right] &= \mathbb{E}\left[(X^TX)^{-1}X^Ty\right] = (X^TX)^{-1}X^T\mathbb{E}\left[y\right] \\ &= (X^TX)^{-1}X^T\mathbb{E}\left[X\beta+\varepsilon\right] = (X^TX)^{-1}X^TX\beta = \beta,\end{aligned}$$

lo cual significa que $\widehat{\beta}$ es un estimador insesgado para β.
Del mismo modo, para la varianza tenemos que

$$\begin{aligned}Var\left[\widehat{\beta}\right] &= Var\left[(X^TX)^{-1}X^Ty\right] = (X^TX)^{-1}X^TVar\left[y\right]\left((X^TX)^{-1}X^T\right)^T \\ &= (X^TX)^{-1}X^T\sigma^2IX\left(X^TX\right)^{-1} = \sigma^2(X^TX)^{-1}.\end{aligned}$$

Resta ver que $\widehat{\beta}$ es eficiente dentro de la clase de estimadores lineales. Sea $c \in \mathbb{R}^p$ donde $p = rango(X)$.
Puesto que $c^T\widehat{\beta} = c^T(X^TX)^{-1}X^Ty = \sum_{i=1}^{n} a_iy_i$, se sigue que $c^T\widehat{\beta}$ es un estimador lineal de $c^T\beta$. Sea d^Ty un estimador lineal insesgado de $c^T\beta$. Se tiene entonces que

$$\begin{aligned}\mathbb{E}\left[c^T\widehat{\beta} \mid X\right] &= c^T\beta \\ c^T\beta = \mathbb{E}\left[d^Ty \mid X\right] &= d^T\mathbb{E}\left[y \mid X\right] = d^TX\beta.\end{aligned}$$

De lo anterior se sigue que $d^TX = c^T$, o bien, que $c = X^Td$. Además,

$$\begin{aligned}Var\left[c^T\widehat{\beta} \mid X\right] &= c^TVar\left[\widehat{\beta \mid X}\right]c = \sigma^2c^T\left(X^TX\right)^{-1}c \\ Var\left[d^Ty \mid X\right] &= d^TVar\left[y \mid X\right]d = \sigma^2d^Td.\end{aligned}$$

Por lo tanto,

$$\begin{aligned}Var\left[d^Ty \mid X\right] - Var\left[c^T\widehat{\beta} \mid X\right] &= \sigma^2\left(d^td - c^T(X^TX)^{-1}c\right) \\ &= \sigma^2\left(d^td - \left(d^TX\right)\left(X^TX\right)^{-1}\left(X^Td\right)\right) \\ &= \sigma^2d^T\left(I - X\left(X^TX\right)^{-1}X^T\right)d\end{aligned}$$

En la tarea 1 ya se probó que la matriz sombrero $H = X\left(X^TX\right)^{-1}X^T$ es una matriz idempotente y simétrica (y por lo tanto de proyección). Puesto que $I - X\left(X^TX\right)^{-1}X^T$ es también una matriz de proyección (al espacio ortogonal de columnas de X), se tiene que también es idempotente y simétrica. Haciendo uso de estas propiedades en lo anterior, obtenemos,

$$\begin{aligned}Var\left[d^Ty \mid X\right] - Var\left[c^T\widehat{\beta} \mid X\right] &= \sigma^2d^T\left(I - X\left(X^TX\right)^{-1}X^T\right)^T\left(I - X\left(X^TX\right)^{-1}X^T\right)d \\ &= \sigma^2\left\|\left(I - X\left(X^TX\right)^{-1}X^T\right)d\right\|^2 \geq 0.\end{aligned}$$

Esto demuestra que cualquier otro estimador lineal insesgado d^Ty de $c^T\beta$ tiene varianza mayor o igual que $c^T\widehat{\beta}$, es decir, $c^T\widehat{\beta}$ es el mejor estimador lineal insesgado de $c^T\beta$.

Alumno: Bueno Rivera Oswaldo
Alumno: Rodríguez Villagrán Juan Pablo | **Profesor:** Marco Antonio Aquino López

2. *Regresión lineal Bayesiana (priori conjugada).* Sea

$$(Y|\beta,\sigma^2) \sim N_n(X\beta,\sigma^2 I_n),$$

el modelo lineal normal, donde $Y \in \mathbb{R}^n$ es el vector de respuestas, $X \in \mathbb{R}^{n\times p}$ la matriz de diseño. Suponemos que los parámetros (β,σ^2) están relacionados por la priori conjugada

$$(\beta|\sigma^2) \sim N_p(\beta_0,\sigma^2 V_0) \qquad \textbf{y} \qquad \sigma^2 \sim \textbf{GammaInv}(a_0,b_0),$$

donde V_0 es una matriz definida positiva y $a_0, b_0 > 0$.

a) *Priori conjugada y distribución posterior.* Obtén la densidad posterior conjunta $p(\beta,\sigma^2|y)$ e identifica los parámetros (β_n, V_n, a_n, b_n) de las posteriores conjugadas

$$(\beta|\sigma^2,y) \sim N_p(\beta_n,\sigma^2 V_n) \quad \textbf{y} \quad (\sigma^2|y) \sim \textbf{GammaInv}(a_n,b_n).$$

Las densidades que da el problema por hipótesis son

$$\begin{aligned}
f_{\sigma^2}(s) &= \frac{b_0^{a_0}}{\Gamma(a_0)} s^{-(a_0+1)} \exp\left(-\frac{b_0}{s}\right) \mathbb{1}_{(0,\infty)}(s) \\
&\propto s^{-(a_0+1)} \exp\left(-\frac{b_0}{s}\right) \mathbb{1}_{(0,\infty)}(s) \\
f_{\beta|\sigma^2}(b|s) &= \frac{1}{(2\pi s)^{p/2}\sqrt{\det V_0}} \exp\left[-\frac{1}{2\sigma^2}(b-\beta_0)^T V_0^{-1}(b-\beta_0)\right] \\
&\propto s^{-p/2} \exp\left[-\frac{1}{2s}(b-\beta_0)^T V_0^{-1}(b-\beta_0)\right] \\
f_{Y|\beta,\sigma^2}(y|b,s) &= (2\pi s)^{-n/2} \exp\left[-\frac{1}{2s}\|y-Xb\|^2\right] \\
&\propto s^{-n/2} \exp\left[-\frac{1}{2s}\|y-Xb\|^2\right].
\end{aligned}$$

Recordemos que, por la fórmula del producto, la densidad posterior puede factorizar como

$$f_{\beta,\sigma^2|Y}(b,s|y) \propto f_{Y|\beta,\sigma^2}(y|b,s) f_{\beta,\sigma^2}(b,s) = f_{Y|\beta,\sigma^2}(y|b,s) f_{\beta|\sigma^2}(b|s) f_{\sigma^2}(s).$$

Como todas las densidades tienen forma exponencial, trataremos esto en escala logarítmica, así la log-posterior se puede calcular como

$$\ell_{\beta,\sigma^2|Y}(b,s|y) \propto \ell_{Y|\beta,\sigma^2}(y|b,s) + \ell_{\beta|\sigma^2}(b|s) + \ell_{\sigma^2}(s).$$

Primero calculamos la log-conjunta de β y σ

$$\begin{aligned}
\ell_{\beta,\sigma^2}(b,s) &= \ell_{\beta|\sigma^2}(b|s) + \ell_{\sigma^2}(s) \\
&= -\frac{p}{2}\ln s - \frac{1}{2s}(b-\beta_0)^T V_0^{-1}(b-\beta_0) - (a_0+1)\ln s - \frac{b_0}{s} \\
&= -\left(a_0 + 1 + \frac{p}{2}\right)\ln s - \frac{1}{s}\left[b_0 + \frac{1}{2}(b-\beta_0)^T V_0^{-1}(b-\beta_0)\right].
\end{aligned}$$

Con esta, calculamos la log-posterior conjunta de β y σ^2

$$\begin{aligned}\ell_{\beta,\sigma^2|Y}(b,s|y) &\propto \ell_{Y|\beta,\sigma^2}(y|b,s) + \ell_{\beta,\sigma^2}(b,s)\\ &= -\frac{n}{2}\ln s - \frac{1}{2s}\|y-Xb\|^2 - \left(a_0+1+\frac{p}{2}\right)\ln s - \frac{1}{s}\left[b_0 + \frac{1}{2}(b-\beta_0)^T V_0^{-1}(b-\beta_0)\right].\\ &= -\left(a_0+1+\frac{p}{2}+\frac{n}{2}\right)\ln s - \frac{1}{2s}\left[2b_0 + (b-\beta_0)^T V_0^{-1}(b-\beta_0) + \|y-Xb\|^2\right].\end{aligned}$$

Examinemos el segundo sumado, al cual podemos darle factorización explícita de forma cuadrática. Entonces, con lo anterior en cuenta, tenemos que

$$\begin{aligned}2b_0 + (b-\beta_0)^T V_0^{-1}(b-\beta_0) + \|y-Xb\|^2 &= 2b_0 + (b-\beta_0)^T V_0^{-1}(b-\beta_0) + (y-Xb)^T(y-Xb)\\ &= 2b_0 + b^T V_0^{-1} b - b^T V_0^{-1}\beta_0 - \beta_0^T V_0^{-1} b + \beta_0^T V_0^{-1}\beta_0 + y^T y - y^T(Xb) - (Xb)^T y + (Xb)^T(Xb)\\ &= 2b_0 + y^T y + b^T(V_0^{-1}+X^TX)b - b^T(V_0^{-1}\beta_0 + X^T y) - (\beta_0^T V_0^{-1} + y^T X)b + \beta_0^T V_0^{-1}\beta_0\\ &= (2b_0 + \beta_0^T V_0^{-1}\beta_0 + y^T y) + \underbrace{b^T(V_0^{-1}+X^TX)b - b^T(V_0^{-1}\beta_0 + X^T y) - (\beta_0^T V_0^{-1} + y^T X)b}_{(*)}.\end{aligned}$$

A partir de $(*)$, es fácil identificar qué falta para "completar" la forma cuadrática. El patrón que buscamos replicar es

$$(b-a)^T A(b-a) = b^T Ab - b^T Aa - a^T Ab + a^T Aa.$$

De la forma obtenida, sabemos que

$$A = V_0^{-1} + X^T X \qquad \text{y} \qquad \begin{cases} Aa = V_0^{-1}\beta_0 + X^T y \\ a^T A = \beta_0^T V_0^{-1} + y^T X. \end{cases}$$

Para obtener a, despejamos de la primera ecuación de modo que

$$\begin{aligned}Aa = V_0^{-1}\beta_0 + X^T y &\implies (V_0^{-1}+X^TX)a = V_0^{-1}\beta_0 + X^T y\\ &\iff a = (V_0^{-1}+X^TX)^{-1}(V_0^{-1}\beta_0 + X^T y),\end{aligned}$$

siempre que $V_0^{-1} + X^T X$ sea invertible. Por lo tanto,

$$\begin{aligned}2b_0 + (b-\beta_0)^T V_0^{-1}(b-\beta_0) + \|y-Xb\|^2 &= 2b_0 + (b-\beta_0)^T V_0^{-1}(b-\beta_0) + (y-Xb)^T(y-Xb)\\ &= (2b_0 + \beta_0^T V_0^{-1}\beta_0 + y^T y - \beta_n^T V_n^{-1}\beta_n) + (b-\beta_n)^T V_n^{-1}(b-\beta_n),\end{aligned}$$

donde se definieron

$$V_n = (V_0^{-1}+X^TX)^{-1} \qquad \text{y} \qquad \beta_n = V_n(V_0^{-1}\beta_0 + X^T y).$$

De esta manera, la log-verosimilitud posterior conjunta es

$$\begin{aligned}\ell_{\beta,\sigma^2|Y}(b,s|y) &\propto \ell_{Y|\beta,\sigma^2}(y|b,s) + \ell_{\beta,\sigma^2}(b,s)\\ &= -\left(a_0+1+\frac{p}{2}+\frac{n}{2}\right)\ln s - \frac{1}{2s}\left[(2b_0 + \beta_0^T V_0^{-1}\beta_0 + y^T y - \beta_n^T V_n^{-1}\beta_n) + (b-\beta_n)^T V_n^{-1}(b-\beta_n)\right]\\ &= -(a_n+1)\ln s - \frac{p}{2}\ln s - \frac{1}{2s}\left[2b_n + (b-\beta_n)^T V_n^{-1}(b-\beta_n)\right]\\ &= \left[-(a_n+1)\ln s - \frac{b_n}{s}\right] + \left[-\frac{p}{2}\ln s - \frac{1}{2s}(b-\beta_n)^T V_n^{-1}(b-\beta_n)\right],\end{aligned}$$

donde ahora se definieron

$$a_n = a_0 + \frac{n}{2} \qquad \text{y} \qquad b_n = b_0 + \frac{1}{2}\left[\beta_0^T V_0^{-1}\beta_0 + y^T y - \beta_n^T V_n^{-1}\beta_n)\right].$$

De este modo, la densidad posterior es de la forma

$$f_{\beta,\sigma^2|y}(b,s|y) \propto \left[s^{-\frac{p}{2}} \exp\left(-\frac{1}{2s}(b-\beta_n)^T V_n^{-1}(b-\beta_n)\right)\right]\left[s^{-(a_n+1)}\exp\left(-\frac{b_n}{s}\right)\right].$$

Notemos que el primer factor es el kernel de la densidad de una variable aleatoria normal de parámetros β_n y $\sigma^2 V_n$, mientras que el segundo factor es el kernel de la densidad de una variable aleatoria Gamma inversa de parámetros a_n y b_n.
El parámetro de precisión posterior V_n^{-1} es la suma de la precisión a priori V_0^{-1} con la información que aportan los datos $X^T X$, mientras que la media normal posterior β_n es la media ponderada (por la precisión) entre la media previa β_0 y la media de mínimos cuadrados $X^T y$.

b) *Distribuciones marginales.* Identifica las distribuciones marginales posteriores de β y de σ^2.
Con constantes de normalización, la densidad posterior es

$$f_{\beta,\sigma^2|Y}(b,s|y) = \left[(2\pi s)^{-p/2}\frac{1}{\sqrt{\det V_n}}\exp\left(-\frac{1}{2s}(b-\beta_n)^T V_n^{-1}(b-\beta_n)\right)\right]\left[\frac{b_n^{a_n}}{\Gamma(a_n)}s^{-(a_n+1)}\exp\left(-\frac{b_n}{s}\right)\right].$$

La densidad marginal posterior de σ^2, con esta representación, es sencilla de calcular ya que el primer factor es la densidad de una normal de parámetros β_n y $\sigma^2 V_n$, por lo que integra 1

$$\begin{aligned} f_{\sigma^2|Y}(s|y) &= \int_{\mathbb{R}^d} f_{\beta,\sigma^2|y}(b,s|y)db \\ &= \frac{b_n^{a_n}}{\Gamma(a_n)}s^{-(a_n+1)}\exp\left(-\frac{b_n}{s}\right)\int_{\mathbb{R}^d}(2\pi s)^{-p/2}\frac{1}{\sqrt{\det V_n}}\exp\left(-\frac{1}{2s}(b-\beta_n)^T V_n^{-1}(b-\beta_n)\right)db \\ &= \frac{b_n^{a_n}}{\Gamma(a_n)}s^{-(a_n+1)}\exp\left(-\frac{b_n}{s}\right), \end{aligned}$$

es decir, $(\sigma^2|y) \sim \text{GammaInv}(a_n, b_n)$. Dicho sea de paso, se demostró que $(\beta|\sigma^2, Y) \sim N_p(\beta_n, \sigma^2 V_n)$, ya que

$$f_{\beta|\sigma^2,Y}(b|s,y) = \frac{f_{\beta,\sigma^2,Y}(b,s,y)}{f_{\sigma^2,Y}(s,y)} = \frac{f_{\beta,\sigma^2|Y}(b,s|y)f_Y(y)}{f_{\sigma^2|Y}(s|y)f_Y(y)} = \frac{f_{\beta,\sigma^2|Y}(b,s|y)}{f_{\sigma^2|Y}(s|y)}.$$

Para calcular la densidad marginal posterior de β, primero definamos Q_β como la forma cuadrática de la normal, es decir, $Q_\beta = (b-\beta_n)^T V_n^{-1}(b-\beta_n)$, así la integral para la marginal es

$$\begin{aligned} f_{\beta|Y}(b|y) &= \int_{\mathbb{R}} f_{\beta,\sigma^2|Y}(b,s|y)ds \\ &= \frac{b_n^{a_n}}{\Gamma(a_n)}(2\pi)^{-p/2}\frac{1}{\sqrt{\det V_n}}\int_0^\infty s^{-(a_n+\frac{p}{2}+1)}\exp\left[-\frac{1}{s}\left(b_n+\frac{1}{2}Q_\beta\right)\right]ds \\ &= \frac{b_n^{a_n}}{\Gamma(a_n)}\frac{\Gamma(a_n+\frac{p}{2})}{\left(b_n+\frac{1}{2}Q_\beta\right)^{a_n+\frac{p}{2}}}(2\pi)^{-p/2}\frac{1}{\sqrt{\det V_n}}\int_0^\infty \frac{\left(b_n+\frac{1}{2}Q_\beta\right)^{a_n+\frac{p}{2}}}{\Gamma(a_n+\frac{p}{2})}s^{-(a_n+\frac{p}{2}+1)}\exp\left[-\frac{1}{s}\left(b_n+\frac{1}{2}Q_\beta\right)\right]ds \\ &= \frac{b_n^{a_n}}{\Gamma(a_n)}\frac{\Gamma(a_n+\frac{p}{2})}{\left(b_n+\frac{1}{2}Q_\beta\right)^{a_n+\frac{p}{2}}}(2\pi)^{-p/2}\frac{1}{\sqrt{\det V_n}}, \end{aligned}$$

donde se llegó a la última expresión utilizando que se integró la densidad de una Gamma inversa de parámetros $a_n + \frac{p}{2}$ y $b_n + \frac{1}{2}Q_\beta$. Desarrollamos esta última expresión para dar una forma que dependa de b

$$\begin{aligned} f_{\beta|Y}(b|y) &= \frac{b_n^{a_n}}{\Gamma(a_n)} \frac{\Gamma(a_n + \frac{p}{2})}{(b_n + \frac{1}{2}Q_\beta)^{a_n + \frac{p}{2}}} (2\pi)^{-p/2} \frac{1}{\sqrt{\det V_n}} \\ &= \frac{\Gamma\left(\frac{2a_n + p}{2}\right)}{\Gamma\left(\frac{2a_n}{2}\right)} (2a_n\pi)^{-p/2} \frac{1}{\sqrt{b_n^p a_n^{-p} \det V_n}} \left[1 + (b - \beta_n)^T \frac{1}{2b_n} V_n^{-1} (b - \beta_n)\right]^{-\frac{2a_n+p}{2}} \\ &= \frac{\Gamma\left(\frac{\nu + p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} (\nu\pi)^{-p/2} \frac{1}{\sqrt{\det \Sigma}} \left[1 + \frac{1}{\nu}(b - \beta_n)^T \Sigma^{-1} (b - \beta_n)\right]^{-\frac{\nu+p}{2}}, \end{aligned}$$

donde definimos $\nu = 2a_n$ y $\Sigma = \dfrac{b_n}{a_n} V_n$. Así, concluimos que la densidad posterior marginal de β es una t-Student multivariada no centada de $2a_n$ grados de libertad con centro en β_n y matriz de escala Σ, esto es

$$(\beta|y) \sim t_p(\beta_n, \Sigma, \nu).$$

De esta última densidad marginal posterior, notamos que es una distribución de colas pesadas. Particularmente, si $\nu > 1$ no tiene media finita y si tiene $\nu > 2$ tampoco tiene varianza finita. La virtud de que sea una distribución de colas pesadas está en que los intervalos de credibilidad que se dan son más anchos que si se tomara una normal, lo cual refleja la incertidumbre adicional que se tiene sobre σ^2.

Alumno: Bueno Rivera Oswaldo
Alumno: Rodríguez Villagrán Juan Pablo | **Profesor:** Marco Antonio Aquino López

3. ***Conexión por regularización.***

a) ***Regresión Ridge.*** **Muestra que si se toma una priori normal isotrópica**

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

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

$$\widehat{\beta}_{\mathbf{MAP}} = \arg\min_{\beta \in \mathbb{R}^d} \|y - X\beta\|^2 + \lambda \|\beta\|^2, \quad \textbf{con} \quad \lambda = \frac{\sigma^2}{\tau^2}.$$

Considerando que a priori $\beta \sim N(0, \tau^2 I_p)$, tenemos que

$$p(y \mid \beta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \|y - X\beta\|^2\right)$$
$$p(\beta) = (2\pi\tau^2)^{-p/2} \exp\left(-\frac{1}{2\tau^2} \|\beta\|^2\right).$$

Luego, tenemos que la distribución posterior de β satisface,

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

El estimador de máxima a posteriori maximiza $p(\beta \mid y)$, y esto último es equivalente a minimizar $-\log p(\beta \mid y)$. De nuestra ecuación anterior, se sigue que debemos minimizar

$$\frac{1}{2\sigma^2} \|y - X\beta\|^2 + \frac{1}{2\tau^2} \|\beta\|^2,$$

lo cual es equivalente a minimizar

$$\|y - X\beta\|^2 + \frac{\sigma^2}{\tau^2} \|\beta\|^2.$$

Definiendo $\lambda := \sigma^2/\tau^2$ resulta que

$$\widehat{\beta}_{MAP} = \text{argmin}_\beta \left(\|y - X\beta\|^2 + \lambda \|\beta\|^2\right),$$

que es justamente el estimador Ridge.
La priori normal centrada en 0 sugiere que se tiene la creencia de que los coeficientes de´la regresión podrían ser cercanos a 0, con varianza τ^2. Entre más pequeña sea τ^2, la creencia de que los parámetros sean 0 es mayor, esto se refleja en λ, ya que cuando $\tau^2 \to 0$, la penalización debida a λ es mucho mayor. El objetivo de esta regularización es reducir el riesgo a problemas de varianza o de multicolinealidad. Por último, es importante destacar que cuando $\tau^2 \to \infty$, o sea, cuando la priori es plana idénticamente 0, el estimador de Ridge coincide con el de mínimos cuadrados ordinarios del ejercicio 1.

b) *Regresión LASSO.* Muestra que si en lugar de una priori normal se utiliza una priori Laplace (doble-exponencial), es decir

$$p(\beta_j) \propto \exp\left(-\lambda|\beta_j|\right),$$

el estimador MAP corresponde a la regresión LASSO

$$\widehat{\beta}_{\mathbf{MAP}} = \arg\min_{\beta\in\mathbb{R}^d} \|y - X\beta\|^2 + \lambda\|\beta\|_1.$$

Supongamos ahora que $p(\beta_j) \propto \exp\left(-\lambda|\beta_j|\right)$, entonces $p(\beta) \propto \exp\left(-\lambda \|\beta\|_1\right)$. En este caso tenemos las mismas hipótesis sobre y, de manera que $p(y \mid \beta)$ se mantiene sin cambios. Luego, haciendo un análisis similar al primer inciso,

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

Maximizar la posterior es equivalente a minimizar el negativo del logaritmo de la posterior, lo cual significa minimizar

$$\frac{1}{2\sigma^2}\|y - X\beta\|^2 + \lambda\|\beta\|_1,$$

y esto equivale a minimizar

$$\|y - X\beta\|^2 + 2\sigma^2\lambda\|\beta\|_1.$$

Definiendo $\tilde{\lambda} = 2\sigma^2\lambda$ obtenemos que el estimador máxima a posteriori está dado por

$$\widehat{\beta}_{MAP} = \text{argmin}_\beta\left(\|y - X\beta\|^2 + \tilde{\lambda}\|\beta\|_1\right),$$

lo cual corresponde al estimador LASSO.
Este estimador refuerza la idea de Ridge, sólo que aquí hay incluso más masa concentrada alrededor de 0, lo que hace más probable que algún coeficiente regresor sea 0. Por esta idea, el estimador LASSO es utilizado para selección de variables.

Alumno: Bueno Rivera Oswaldo
Alumno: Rodríguez Villagrán Juan Pablo | **Profesor:** Marco Antonio Aquino López

4. *Extensiones: errores no normales.*

a) *Modelos alternativos.* Propón un modelo de regresión donde el error ϵ no siga una distribución normal: $\epsilon \sim$ Laplace$(0, b)$ (robusto a outliers) o $\epsilon \sim$ Student(ν).

b) *Consecuencias metodológicas.* Explica cuáles serían las consecuencias sobre: la forma de la verosimilitud, la existencia o no de prioris conjugadas y los métodos de inferencia requeridos (MCMC, variacional, etc.).

1. Primero consideremos un modelo de regresión $y = X\beta + \varepsilon$ donde $\varepsilon \sim Laplace(0, b)$. En este caso tendremos una verosimilitud
$$p(y \mid \beta) = \prod_{i=1}^{n} \frac{1}{2b} \exp\left\{ -\frac{|y_i - X_i^T\beta|}{b} \right\}.$$
Observemos que maximizar dicha verosimilitud sobre β es equivalente a minimizar
$$\sum_{i=1}^{n} |y_i - X_i^T\beta|,$$
dicha función es convexa (pues el valor absoluto lo es) pero no es diferenciable en los puntos donde algún $y_i - X_i^T\beta = 0$, de manera que no podemos usar las mismas técnicas de optimización que en el caso gaussiano. En esta situación se requiere aplicar otras técnicas de optimización para encontrar el estimador β que maximiza la verosimilitud, típicamente técnicas de programación lineal.

   En este caso, además no existen priors conjugados simples para β, (a diferencia del caso gaussiano). Dicho esto, es necesario utilizar métodos numéricos para muestrear la posterior. Una herramienta que facilita los métodos numéricos es la formulación de la Laplace como una mezcla de variables aleatorias normales con variable mezclante exponencial.

2. Si ahora consideramos un modelo de regresión con $\varepsilon \sim t - Student(\nu)$, tendremos una verosimilitud dada por
$$p(y \mid \beta) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\left(y_i - X_i^T\beta\right)^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$
De aquí, la log-verosimilitud está dada por
$$\ell(y \mid \beta) = n\left[\log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\log(\nu\pi)\right] - \frac{\nu+1}{2}\sum_{i=1}^{n} \log\left(1 + \frac{\left(y_i - X_i^T\beta\right)^2}{\nu}\right).$$
Observemos que al derivar lo anterior con respecto a β e igualar a cero obtenemos las ecuaciones no lineales
$$\frac{\partial \ell(y \mid \beta)}{\partial \beta}(\nu + 1)\sum_{i=1}^{n} \frac{X_{ij}(y_i - X_i^T\beta)}{\nu + (y_i - X_i^T\beta)^2} = 0,$$
de manera que la maximización analítica deja de ser una opción. Nuevamente, igual que antes se requerirán métodos numéricos para encontrar el estimador de máxima verosimilitud para β.

   Igual que en el caso Laplace no existen priors conjugados adecuados para la estimación de β bajo errores t-student, y por lo tanto es necesario utilizar métodos numéricos para muestrear la posterior, típicamente un algoritmo de mínimos cuadrados pesados iterativos. Una herramienta que facilita los métodos numéricos es la formulación de la t como una mezcla de variables aleatorias normales con variable mezclante Gamma.

3. Otra posible modificación, la cual tiene solución analítica, es el problema de mínimos cuadrados generalizados, donde se considera que el error es tal que $\varepsilon \sim N(0, \sigma^2\Omega)$, con Ω una matriz definida positiva. Aquí el estimador resulta ser

$$\widehat{\beta}_{\text{MCG}} = (X^T\Omega X)^{-1} X^T\Omega^{-1}T.$$

   Si bien esto lleva a relajar la hipótesis de isotropía, tiene el problema metodológico de conocer la estructura de covarianza de los errores.

4. Otra alternativa para relajar la hipótesis de isotropía de la varianza, ahora en el esquema Bayesiano, es considerar que la matriz de varianza covarianza es una matriz definida positiva aleatoria. Una elección para esto es proponer que la priori sea una matriz de Wishart, ya que esta es la generalización matricial de la distribución Gamma. Elegir una matriz de Wishart tiene la ventaja de que, con una verosimilitud normal, forma una familia conjugada.