



Ciencia de Datos

Tarea 3 - Parte Teórica

María Alejandra Borrego Leal
Iván García Mestiza
Rodolfo de Jesús Ramírez Lucario

13 de Octubre de 2025

Problema 1. Regresión lineal ordinaria (OLS)

1. **Derivación del estimador OLS:** Partiendo del modelo clásico:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \begin{matrix} \sigma^2 \\ \text{I} \end{matrix}),$$

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.

2. **Propiedades del estimador:** Calcule explícitamente:

$$E[\hat{\beta}], \quad \text{Var}(\hat{\beta}).$$

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

Demostración. Sean $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$ y $\boldsymbol{\beta} \in \mathbb{R}^p$, con $n \leq p$. Consideremos el modelo de regresión lineal clásico

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{con} \quad \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I),$$

y supongamos que la matriz de diseño \mathbf{X} es de rango completo.

1. Denotemos por $SS_{Res}(\boldsymbol{\beta})$ a la suma de cuadrados de los errores residuales, entonces

$$SS_{Res}(\boldsymbol{\beta}) = \sum_{i=1}^n \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

desarrollando la última expresión podemos escribir SS_{Res} como

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

esto último dado que $\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} \in \mathbb{R}$, lo que implica que es igual a su transpuesta, es decir $\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} = (\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y})^T = \mathbf{y}^T \mathbf{X} \boldsymbol{\beta}$.

Derivamos con respecto a $\boldsymbol{\beta}$ y obtenemos

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

Luego, el estimador debe satisfacer que

$$\left. \frac{\partial SS_{Res}}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \mathbf{0},$$

por consiguiente,

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y},$$

y como $p = \text{Rango}(\mathbf{X}) = \text{Rango}(\mathbf{X}^T \mathbf{X})$, entonces $\mathbf{X}^T \mathbf{X}$ es de rango completo, lo cual implica que es invertible. Entonces, tenemos que

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Finalmente, comprobamos que $\hat{\boldsymbol{\beta}}$ minimiza la suma de cuadrados residuales obteniendo su matriz Hessiana, H , y verificando que esta es positiva definida. Así, veamos que

$$H = \left. \frac{\partial^2 SS_{Res}}{\partial \boldsymbol{\beta}^2} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 2\mathbf{X}^T \mathbf{X},$$

y como $\mathbf{X}^T \mathbf{X}$ es de rango completo, entonces es positiva definida. Por lo que H es una matriz positiva definida.

Por lo tanto, el estimador de Mínimos Cuadrados Ordinarios es

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

2. Notemos que

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta.$$

Por tanto, el estimador $\hat{\beta}$ es un estimador insesgado.

Luego, veamos que

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}] \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var}(\mathbf{y}) [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T]^T \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 I_n \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2. \end{aligned}$$

Ahora bien, queremos ver que $\hat{\beta}$ es eficiente dentro de la clase de estimadores lineales, es decir que, dentro de esta clase, $\hat{\beta}$ es el estimador de mínima varianza. Entonces, sea $L \in \mathbb{R}^p$, probaremos que bajo las condiciones de Gauss-Markov, $L^T \hat{\beta}$ es el mejor estimador lineal insesgado de $L^T \beta$.

Primero, denotemos $a = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} L$, entonces $a \in \mathbb{R}^n$ y puesto que

$$L^T \hat{\beta} = L^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = [\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} L]^T \mathbf{y} = a^T \mathbf{y} = \sum_{i=1}^n a_i y_i,$$

se sigue que $L^T \hat{\beta}$ es un estimador lineal de $L^T \beta$.

Ahora, sea $C^T \mathbf{y}$ un estimador lineal insesgado de $L^T \beta$. Tenemos que

$$\mathbb{E}(L^T \hat{\beta}) = L^T \mathbb{E}(\hat{\beta}) = L^T \beta$$

y

$$\mathbb{E}(C^T \mathbf{y}) = C^T \mathbb{E}(\mathbf{y}) = C^T \mathbf{X} \beta = L^T \beta, \text{ si, y solo si, } C^T \mathbf{X} = L^T.$$

Luego,

$$\text{Var}(L^T \hat{\beta}) = L^T \text{Var}(\hat{\beta}) L = \sigma^2 L^T (\mathbf{X}^T \mathbf{X})^{-1} L$$

y

$$\text{Var}(C^T \mathbf{y}) = C^T \text{Var}(\mathbf{y}) C = \sigma^2 C^T C.$$

Por consiguiente,

$$\begin{aligned} \text{Var}(C^T \mathbf{y}) - \text{Var}(L^T \hat{\beta}) &= \sigma^2 C^T C - \sigma^2 L^T (\mathbf{X}^T \mathbf{X})^{-1} L \\ &= \sigma^2 (C^T C - L^T (\mathbf{X}^T \mathbf{X})^{-1} L), \end{aligned}$$

pero recordemos que $L^T = C^T \mathbf{X}$ de donde $L = \mathbf{X}^T C$, entonces

$$\begin{aligned}\text{Var}(C^T \mathbf{y}) - \text{Var}(L^T \hat{\boldsymbol{\beta}}) &= \sigma^2 (C^T C - C^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T C) \\ &= \sigma^2 C^T (I_n - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) C \\ &= \sigma^2 C^T (I_n - P) C,\end{aligned}$$

con $P = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. A continuación veamos que la matriz $I_n - P$ es positiva semidefinida. Para esto, recordemos (por Ejercicio 1 de la Tarea 1) que P es simétrica e idempotente, por lo que $I_n - P$ también lo es, pues

$$(I_n - P)^T = I_n^T - P^T = I_n - P,$$

y

$$(I_n - P)(I_n - P) = I_n - 2P + P^2 = I_n - 2P + P = I_n - P.$$

Luego, para cualquier $v \in \mathbb{R}^n$, tenemos que

$$\begin{aligned}v^T (I_n - P) v &= v^T (I_n - P)^T (I_n - P) v \\ &= [(I_n - P) v]^T [(I_n - P) v] \\ &= \|(I_n - P) v\|^2 \geq 0.\end{aligned}$$

Así, como $I_n - P$ es positiva semidefinida, se satisface que $C^T (I_n - P) C \geq 0$. Por consiguiente

$$\text{Var}(C^T \mathbf{y}) - \text{Var}(L^T \hat{\boldsymbol{\beta}}) \geq 0, \quad \text{si, y solo si} \quad \text{Var}(C^T \mathbf{y}) \geq \text{Var}(L^T \hat{\boldsymbol{\beta}}).$$

Por lo tanto, $L^T \hat{\boldsymbol{\beta}}$ es el mejor estimador lineal insesgado de $L^T \boldsymbol{\beta}$.

□

Problema 2. Regresión lineal Bayesiana (prior conjugado)



1. **Prior Conjugado:** Suponga un prior conjugado:

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

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

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

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

Demostración. Recordemos que, bajo el modelo lineal bayesiano, la variable de respuesta "y" tiene distribución Normal con media $X\beta$ y varianza $\sigma^2 I$ lo cuál indicaremos de la siguiente manera:

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

así, usando el Teorema de Bayes, se tiene que la distribución a posteriori de los parámetros está dada por :

$$p(\beta, \sigma^2|y) \propto p(y|\beta, \sigma^2)p(\beta, \sigma^2) \quad (1)$$

$$= p(y|\beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2) \quad (2)$$

con

$$p(y|\beta, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right\}$$

,

$$p(\beta|\sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n|V_0|^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \beta_0)^T V_0^{-1} (\beta - \beta_0) \right\}$$

y

$$p(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)} (\sigma^2)^{-(a_0+1)} \exp \left(-\frac{b_0}{\sigma^2} \right)$$

y dónde hemos supuesto que tenemos p covariables. Así, notemos que la distribución conjunta apriori de (β, σ^2) esta dada por:

$$p(\beta, \sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)(2\pi)^{n/2}|V_0|^{1/2}} \quad (3)$$

$$\times \frac{1}{\sigma^p} \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \beta_0)^T V_0^{-1} (\beta - \beta_0) \right\} (\sigma^2)^{-(a_0+1)} \exp \left(-\frac{b_0}{\sigma^2} \right)$$

$$\propto \frac{1}{\sigma^p} \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \beta_0)^T V_0^{-1} (\beta - \beta_0) \right\} (\sigma^2)^{-(a_0+1)} \exp \left(-\frac{b_0}{\sigma^2} \right)$$

$$= \frac{1}{(\sigma^2)^{\frac{p}{2}+a_0+1}} \exp \left\{ -\frac{1}{2\sigma^2} [2b_0 + (\beta - \beta_0)^T V_0^{-1} (\beta - \beta_0)] \right\}. \quad (4)$$

Por la Ecuación (2) se tiene que

$$p(\beta, \sigma^2|y) \propto \frac{1}{\sigma^p \sigma^n} (\sigma^2)^{-(a_0+1)} \exp \left\{ -\frac{b_0}{\sigma^2} - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) - \frac{1}{2\sigma^2} (\beta - \beta_0)^T V_0^{-1} (\beta - \beta_0) \right\}$$

Podemos reacomodar la expresión anterior de la siguiente manera:

$$\frac{1}{(\sigma^2)^{\frac{p}{2}+(a_0+\frac{n}{2})+1}} \exp \left\{ -\frac{1}{2\sigma^2} [2b_0 + (y - X\beta)^T (y - X\beta) + (\beta - \beta_0)^T V_0^{-1} (\beta - \beta_0)] \right\} \quad (5)$$

y notemos que desarrollando la expresión dentro de la exponencial se tiene que:

$$\begin{aligned}
 & 2b_0 + (y - \mathbf{X}\boldsymbol{\beta})^T(y - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T V_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
 &= 2b_0 + y^T y - y^T \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}^T y + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T V_0^{-1} \boldsymbol{\beta} - \boldsymbol{\beta}^T V_0^{-1} \boldsymbol{\beta}_0 \\
 & \quad - \boldsymbol{\beta}_0^T V_0^{-1} \boldsymbol{\beta} + \boldsymbol{\beta}_0^T V_0^{-1} \boldsymbol{\beta}_0 \\
 &= 2b_0 + \boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X} + V_0^{-1}) \boldsymbol{\beta} - 2(y^T \mathbf{X} + \boldsymbol{\beta}_0^T V_0^{-1}) \boldsymbol{\beta} + \boldsymbol{\beta}_0^T V_0^{-1} \boldsymbol{\beta}_0 + y^T y
 \end{aligned}$$

Definimos V_n como:

$$V_n^{-1} = \mathbf{X}^T \mathbf{X} + V_0^{-1}$$

donde nótese que si \mathbf{X} es de rango completo, entonces $\mathbf{X}^T \mathbf{X}$ es positiva definida y dado que V_0^{-1} es positiva definida, V_n^{-1} también es positiva definida. Queremos expresar el término

$$\boldsymbol{\beta}^T V_n^{-1} \boldsymbol{\beta} - 2(y^T \mathbf{X} + \boldsymbol{\beta}_0^T V_0^{-1}) \boldsymbol{\beta} + \boldsymbol{\beta}_0^T V_0^{-1} \boldsymbol{\beta}_0 + y^T y$$

de una forma cuadrática del estilo

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_n)^T V_n^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_n) + c = \boldsymbol{\beta}^T V_n^{-1} \boldsymbol{\beta} - 2\boldsymbol{\beta}_n^T V_n^{-1} \boldsymbol{\beta} + \boldsymbol{\beta}_n^T V_n^{-1} \boldsymbol{\beta}_n + c$$

por lo que requerimos encontrar $\boldsymbol{\beta}_n$, para esto, nótese que:

$$\boldsymbol{\beta}_n^T V_n^{-1} = y^T \mathbf{X} + \boldsymbol{\beta}_0^T V_0^{-1}$$

luego,

$$V_n^{-1} \boldsymbol{\beta}_n = \mathbf{X}^T y + V_0^{-1} \boldsymbol{\beta}_0$$

donde hemos usado que V_n^{-1} es simétrica. Dado que ya hemos demostrado que V_n^{-1} es invertible pues es positiva definida, $\boldsymbol{\beta}_n$ está dada por

$$\boldsymbol{\beta}_n = V_n (\mathbf{X}^T y + V_0^{-1} \boldsymbol{\beta}_0)$$

y está bien definida. Con lo anterior se tiene que

$$c = \boldsymbol{\beta}_0^T V_0^{-1} \boldsymbol{\beta}_0 + y^T y - \boldsymbol{\beta}_n^T V_n^{-1} \boldsymbol{\beta}_n$$

y así,

$$\begin{aligned}
 & 2b_0 + (y - \mathbf{X}\boldsymbol{\beta})^T(y - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T V_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
 &= 2 \left[b_0 + \frac{\boldsymbol{\beta}_0^T V_0^{-1} \boldsymbol{\beta}_0 + y^T y - \boldsymbol{\beta}_n^T V_n^{-1} \boldsymbol{\beta}_n}{2} \right] + (\boldsymbol{\beta} - \boldsymbol{\beta}_n)^T V_n^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_n)
 \end{aligned}$$

sustituyendo en la Ecuación (5),

$$\begin{aligned}
 p(\boldsymbol{\beta}, \sigma^2 | y) &\propto \frac{1}{(\sigma^2)^{\frac{p}{2} + (a_0 + \frac{n}{2}) + 1}} \\
 &\exp \left\{ -\frac{1}{2\sigma^2} \left(2 \left[b_0 + \frac{\boldsymbol{\beta}_0^T V_0^{-1} \boldsymbol{\beta}_0 + y^T y - \boldsymbol{\beta}_n^T V_n^{-1} \boldsymbol{\beta}_n}{2} \right] + (\boldsymbol{\beta} - \boldsymbol{\beta}_n)^T V_n^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_n) \right) \right\}
 \end{aligned} \tag{6}$$

Así, los parámetros posteriores (β_n, V_n, a_n, b_n) están dados por:

$$V_n = (\mathbf{X}^T \mathbf{X} + V_0^{-1})^{-1}, \quad \beta_n = V_n(\mathbf{X}^T y + V_0^{-1} \beta_0)$$

$$a_n = a_0 + \frac{n}{2} \quad \text{y} \quad b_n = b_0 + \frac{\beta_0^T V_0^{-1} \beta_0 + y^T y - \beta_n^T V_n^{-1} \beta_n}{2}$$

Dado que hemos llegado a un kernel con la misma forma que el Kernel de la Expresión (4), entonces la constante que hace que integre 1 la Expresión (6) es:

$$K = \frac{b_n^{a_n}}{\Gamma(a_n)(2\pi)^{n/2}|V_n|^{1/2}}$$

dónde hemos sustituido b_0 por b_n , a_0 por a_n , β_0 por β_n y V_0 por V_n en la Expresión (3). Así, la densidad conjunta de la distribución a posteriori de (β, σ^2) es:

$$p(\beta, \sigma^2 | y) = \frac{b_n^{a_n}}{\Gamma(a_n)(2\pi)^{n/2}|V_n|^{1/2}} \frac{1}{(\sigma^2)^{\frac{n}{2}+a_n+1}} \exp \left\{ -\frac{1}{2\sigma^2} (2b_n + (\beta - \beta_n)^T V_n^{-1} (\beta - \beta_n)) \right\} \quad (7)$$

Para obtener las distribuciones marginales de β, σ^2 obsérvese que podemos reescribir la Expresión (7) como

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

y notemos que la segunda expresión en corchetes es una Normal Multivariada con parámetros β_n como media y $\sigma^2 V_n$ como matriz de covarianzas. Así, integrando respecto a β , la segunda expresión se hace 1 y nos queda que

$$p(\sigma^2 | y) = \frac{b_n^{a_n}}{\Gamma(a_n)} (\sigma^2)^{-(a_n+1)} \exp \left(-\frac{b_n}{\sigma^2} \right).$$

por lo que $\sigma^2 | y$ tiene distribución Gamma Inversa con parámetros (a_n, b_n) .

Cómo ya se ha demostrado, podemos reescribir el kernel de $p(\beta, \sigma^2 | y)$ como

$$(\sigma^2)^{-(\frac{n}{2}+a_n+1)} \exp \left\{ -\frac{1}{2\sigma^2} (2b_n + (\beta - \beta_n)^T V_n^{-1} (\beta - \beta_n)) \right\}$$

si definimos Q como

$$Q = 2b_n + (\beta - \beta_n)^T V_n^{-1} (\beta - \beta_n)$$

y cambiamos σ^2 por r entonces podemos reescribir la expresión anterior como:

$$(r)^{-(\frac{n}{2}+a_n+1)} \exp \left\{ -\frac{1}{2r} Q \right\}$$

Si integramos sobre r podemos obtener la marginal de β . Así,

$$p(\beta|y) \propto \int_0^\infty (r)^{-(\frac{p}{2}+a_n+1)} \exp\left\{-\frac{1}{2r}Q\right\} dr$$

haciendo el cambio de variable $u = Q/2r$ se tiene que $du = -\frac{2}{4r^2}Qdr = -\frac{1}{2r^2}Qdr$ y $r = Q/2u$

$$\begin{aligned} p(\beta|y) &\propto \int_0^\infty \left(\frac{2u}{Q}\right)^{(p/2+a_n+1)} \exp\{-u\} \frac{Q}{2u^2} du \\ &= Q^{-(p/2+a_n)} \int_0^\infty u^{(p/2+a_n-1)} \exp\{-u\} du \\ &= \frac{\Gamma\left(\frac{p}{2} + a_n\right)}{Q^{p/2+a_n}} \\ &\propto Q^{-(p/2+a_n)} \\ &= [2b_n + (\beta - \beta_n)^T V_n^{-1} (\beta - \beta_n)]^{-(p/2+a_n)} \\ &\propto \left[1 + \frac{(\beta - \beta_n)^T V_n^{-1} (\beta - \beta_n)}{2b_n}\right]^{-(p/2+a_n)} \\ &= \left[1 + \frac{(\beta - \beta_n)^T V_n^{-1} (\beta - \beta_n)}{2b_n}\right]^{-\frac{p+2a_n}{2}}. \end{aligned} \quad (8)$$

Ahora, recordemos que si Y es una v.a. con distribución t -multivariada (μ, Σ, ν) entonces


$$f_Y(y; \mu, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma(\nu/2) \nu^{p/2} \pi^{p/2} |\Sigma|^{1/2}} \left[1 + \frac{1}{\nu} (\mathbf{y} - \mu)^\top \Sigma^{-1} (\mathbf{y} - \mu)\right]^{-(\nu+p)/2}$$

así, podemos notar que la Expresión (8) es el kernel de una distribución t -multivariada con parámetros

$$\mu = \beta_n, \quad \Sigma = \frac{b_n}{a_n} V_n \quad \text{y} \quad \nu = 2a_n$$

por lo que $\beta|y$ tiene una distribución t -multivariada con parámetros $(\beta_n, \frac{b_n}{a_n} V_n, 2a_n)$. □

Problema 3. Conexión con regularización

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

$$\beta \sim N_p(\mathbf{0}, \tau^2 I_p),$$

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

$$\hat{\beta}_{MAP} = \arg \min_{\beta} (\|\mathbf{Y} - X\beta\|^2 + \lambda \|\beta\|^2), \quad \lambda = \sigma^2/\tau^2.$$

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

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

el estimador MAP corresponde a la regresión Lasso:

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

Demostración. 1. Sean $\tau^2, \sigma^2 > 0$, y consideremos una familia de parámetros β con distribución (a priori) $\beta \sim N_p(\mathbf{0}, \tau^2 I_p)$, mientras que, por el modelo de regresión lineal, $\mathbf{Y} | \beta = \beta \sim N_n(X\beta, \sigma^2 I_n)$. Entonces la densidad de β está dada, para cada $\beta \in \mathbb{R}^p$, por

$$f_{\beta}(\beta) = \frac{1}{(2\pi\tau^2)^{p/2}} \exp \left[-\frac{1}{2\tau^2} \beta^T \beta \right].$$

Ahora bien, la densidad de \mathbf{Y} condicional a β es igual a, para cada $\mathbf{y} \in \mathbb{R}^n$,

$$f_{\mathbf{Y}|\beta}(\mathbf{y}|\beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right].$$

Luego, definiendo $k = (f_{\mathbf{Y}}(\mathbf{y}))^{-1}$, en donde $f_{\mathbf{Y}}$ es la marginal de \mathbf{Y} , se tiene que la distribución *a posteriori* de β , condicional a los datos es, para $\beta \in \mathbb{R}^p$,

$$\begin{aligned} f_{\beta|\mathbf{Y}}(\beta|\mathbf{y}) &= \frac{f_{\mathbf{Y}|\beta}(\mathbf{y}|\beta) f_{\beta}(\beta)}{f_{\mathbf{Y}}(\mathbf{y})} \\ &= k \frac{1}{(2\pi\tau^2)^{p/2}} \exp \left[-\frac{1}{2\tau^2} \beta^T \beta \right] \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right] \\ &\propto \exp \left[-\frac{1}{2\tau^2} \beta^T \beta - \frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right]. \end{aligned}$$

Por otro lado,

$$\begin{aligned} -\frac{1}{2\tau^2} \beta^T \beta - \frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) &= -\frac{1}{2\tau^2} \beta^T \beta - \frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} - \frac{1}{2\sigma^2} \beta^T X^T X \beta + \frac{1}{\sigma^2} \beta^T X^T \mathbf{y} \\ &= -\frac{1}{2} \beta^T \left(\frac{1}{\tau^2} I_p + \frac{1}{\sigma^2} X^T X \right) \beta + \frac{1}{\sigma^2} \beta^T X^T \mathbf{y} - \frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y}. \end{aligned}$$

Definamos

$$\Sigma := \left(\frac{1}{\tau^2} I_p + \frac{1}{\sigma^2} X^T X \right)^{-1} = \sigma^2 \left(X^T X + \frac{\sigma^2}{\tau^2} I_p \right)^{-1}, \quad \mu := \left(X^T X + \frac{\sigma^2}{\tau^2} I_p \right)^{-1} X^T \mathbf{y}.$$

Nótese que $X^T X + \frac{\sigma^2}{\tau^2} I_p$ es definida positiva, al ser suma de dos matrices definidas positivas, por lo que su inversa también lo es. Luego,

$$\begin{aligned}
-\frac{1}{2}(\beta - \mu)^T \Sigma^{-1} (\beta - \mu) &= -\frac{1}{2} (\beta^T \Sigma^{-1} \beta - 2\beta^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu) \\
&= -\frac{1}{2} \beta^T \left(\frac{1}{\tau^2} I_p + \frac{1}{\sigma^2} X^T X \right) \beta + \beta^T \left(\frac{1}{\tau^2} I_p + \frac{1}{\sigma^2} X^T X \right) \left(X^T X + \frac{\sigma^2}{\tau^2} I_p \right)^{-1} X^T \mathbf{y} \\
&\quad - \frac{1}{2} \mu^T \Sigma^{-1} \mu \\
&= -\frac{1}{2} \beta^T \left(\frac{1}{\tau^2} I_p + \frac{1}{\sigma^2} X^T X \right) \beta + \frac{1}{\sigma^2} \beta^T X^T \mathbf{y} - \frac{1}{2} \mu^T \Sigma^{-1} \mu \\
&= -\frac{1}{2} \beta^T \left(\frac{1}{\tau^2} I_p + \frac{1}{\sigma^2} X^T X \right) \beta + \frac{1}{\sigma^2} \beta^T X^T \mathbf{y} - \frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} + c \\
&= -\frac{1}{2\tau^2} \beta^T \beta - \frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) + c,
\end{aligned}$$

en donde

$$c = \frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} - \frac{1}{2} \mu^T \Sigma^{-1} \mu$$

es una constante que no depende de β . Por lo tanto,

$$f_{\beta|\mathbf{Y}}(\beta|\mathbf{y}) \propto \exp \left[-\frac{1}{2} (\beta - \mu)^T \Sigma^{-1} (\beta - \mu) \right],$$

así que la distribución *a posteriori* de β es $\beta \mid \mathbf{Y} = \mathbf{y} \sim N_p(\mu, \Sigma)$, con μ y Σ como antes. Por lo tanto, como la normal alcanza su máximo en la media, el estimador de máxima a posteriori (MAP) es:

$$\hat{\beta}_{MAP} = \mu = \left(X^T X + \frac{\sigma^2}{\tau^2} I_p \right)^{-1} X^T \mathbf{y}$$

Por otro lado, el estimador Ridge de parámetro λ es igual a

$$\hat{\beta}_{\text{Ridge}}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T \mathbf{y},$$

que es igual a $\hat{\beta}_{MAP}$ cuando $\lambda = \sigma^2/\tau^2$.

2. Supongamos ahora que para cada $j \in \{1, \dots, p\}$, β_j sigue una distribución (a priori) Laplace, de modo que su densidad $f(\beta_j)$ es proporcional a $\exp(-\lambda |\beta_j|)$. Entonces la densidad de β para cada $\beta \in \mathbb{R}^p$, es proporcional a

$$f_{\beta}(\beta) \propto \exp \left[-\lambda \sum_{j=1}^p |\beta_j| \right] = \exp [-\lambda \|\beta\|_1].$$

Ahora bien, la densidad de \mathbf{Y} condicional a β es igual a, para cada $\mathbf{y} \in \mathbb{R}^n$,

$$f_{\mathbf{Y}|\beta}(\mathbf{y}|\beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right].$$

Luego, la distribución *a posteriori* de β , condicional a los datos es, para $\beta \in \mathbb{R}^p$,

$$\begin{aligned} f_{\beta|\mathbf{Y}}(\beta|\mathbf{y}) &= \frac{f_{\mathbf{Y}|\beta}(\mathbf{y}|\beta)f_{\beta}(\beta)}{f_{\mathbf{Y}}(\mathbf{y})} \\ &\propto \exp[-\lambda\|\beta\|_1] \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right] \\ &\propto \exp \left[-\lambda\|\beta\|_1 - \frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right]. \end{aligned}$$

En este caso, el estimador de máxima a posteriori $\hat{\beta}_{MAP}$ es tal que

$$\hat{\beta}_{MAP} = \arg \max_{\beta} \left(\exp \left[-\lambda\|\beta\|_1 - \frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right] \right).$$

Ahora bien, maximizar la función dentro del paréntesis es equivalente a maximizar su logaritmo (pues este es una función uno a uno, estrictamente creciente). Luego,

$$\hat{\beta}_{MAP} = \arg \max_{\beta} \left(-\lambda\|\beta\|_1 - \frac{1}{2\sigma^2} (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \right) = \arg \max_{\beta} \left(-\lambda\|\beta\|_1 - \frac{1}{2\sigma^2} \|\mathbf{y} - X\beta\|^2 \right).$$

Ahora bien, puesto que $\min(-g) = -\max g$ para cualquier función g , se tiene que

$$\hat{\beta}_{MAP} = \arg \min_{\beta} \left(\lambda\|\beta\|_1 + \frac{1}{2\sigma^2} \|\mathbf{y} - X\beta\|^2 \right) = \arg \min_{\beta} \left[(2\sigma^2\lambda)\|\beta\|_1 + \|\mathbf{y} - X\beta\|^2 \right],$$

en donde la última igualdad se da puesto que σ^2 es una constante positiva, y multiplicar por un número positivo no cambia el argumento del mínimo. Por lo tanto, para $\lambda' = 2\sigma^2\lambda$, el estimador MAP corresponde al estimador de la regresión Lasso de parámetro λ' . \square

Problema 4. Extensiones: Errores no normales



1. **Modelos alternativos:** Proponga un modelo de regresión en donde el error ε no siga una distribución Normal. Ejemplos:

- $\varepsilon \sim \text{Laplace}(0, b)$ (robusto a outliers).
- $\varepsilon \sim \text{Student-}t(\nu)$ (colas pesadas).

2. **Consecuencias metodológicas:** Explique cuáles serían las consecuencias sobre:

- La forma de la verosimilitud.
- La existencia o no de priors conjugados.
- Los métodos de inferencia requeridos (MCMC, aproximación variacional, etc.)

Demostración. Consideremos primero el caso en el que los errores siguen una distribución Laplace($0, b$), es decir, la densidad f de ε_i para cada $i = 1, \dots, n$ es, para todo $\varepsilon \in \mathbb{R}$, igual a

$$f(\varepsilon) = \frac{1}{2b} \exp \left\{ -\frac{|\varepsilon|}{b} \right\}.$$

En este caso, notemos que la verosimilitud es proporcional a

$$L(\boldsymbol{\varepsilon}; y) \propto \exp \left\{ -\frac{1}{b} \sum_{i=1}^n |y_i - X_i^T \boldsymbol{\beta}| \right\},$$

y en comparación con el caso normal, no se tiene suma de los cuadrados de los errores, sino la suma de sus valores absolutos. Esto implica que esta distribución es más robusta a outliers, puesto que la distribución Laplace tiene colas más pesadas que la normal y la verosimilitud crece de manera lineal con respecto a los errores, a diferencia de cuando se toman normales, en donde el crecimiento es cuadrático. Como consecuencia, los outliers tienen menos influencia, puesto que los valores muy grandes de los residuales no son penalizados tan severamente.

Por otro lado, la desventaja de esta distribución es que no existe una *a priori* conjugada para $\boldsymbol{\beta}$ simple. Sin embargo, un truco usual que se puede aplicar es escribir a la distribución Laplace como una mezcla de distribuciones normales. Específicamente, si

$$\varepsilon_i | \tau_i = t \sim \mathcal{N}(0, t), \quad \tau_i \sim \text{Exponencial}(1/(2b^2)), \quad (9)$$

en donde el parámetro de la exponencial corresponde a la tasa, y los τ_i son independientes, entonces al marginalizar sobre τ_i se obtiene la distribución Laplace($0, b$). Esto permite que, condicional a $\boldsymbol{\tau}$, $\boldsymbol{\beta}$ siga una distribución normal, y por consiguiente hace los cálculos más fáciles (aunque terminan siendo densidades condicionales). Sin embargo, lo anterior permite la inferencia por métodos computacionales, como lo son:

- **Optimización convexa:** Puesto que la función de logverosimilitud es cóncava (o bien, su negativo es convexo), se pueden usar métodos de programación lineal para inferir $\boldsymbol{\beta}$ al maximizar la logverosimilitud. Esto a su vez es equivalente a la regresión cuantílica, en específico, la estimación de la mediana, puesto que es conocido que la mediana es aquella que minimiza las funciones de pérdida de sumas de valores absolutos.
- **MCMC:** Los métodos tradicionales de MCMC pueden ser aplicados gracias a (9). En efecto, se puede hacer un muestreo de $\boldsymbol{\tau}$ y a partir de ahí realizar las inferencias con la distribución normal, y ponderar de acuerdo a los pesos de $\boldsymbol{\tau}$ en su distribución exponencial. Esto también corresponde con el **muestreo de Gibbs**.

- **Mean-field VI:** Se pueden usar métodos de aproximación variacional para calcular las distribuciones posteriores de β y τ de manera aproximada, de nuevo por la relación (9). Además, debido a que la distribución Laplace tiene “picos” más definidos cerca de su media, la distribución posterior tiene un “pico” cerca del centro de los datos pero colas más pesadas, así que aquí nuevamente se puede ver que la influencia de los outliers es reducida.

Por otra parte, consideremos el caso en el que los errores siguen una distribución t de Student con ν grados de libertad y escala σ^2 . Puesto que la densidad de t decae de manera polinómica, los residuales grandes son más probables con este modelo que con el Normal, lo que hace que los parámetros sean menos afectados por los outliers. Además, mientras más chico es ν , las colas son más pesadas, lo que hace que las estimaciones sean más robustas. Ahora bien, la logverosimilitud toma la forma

$$l(\beta, \sigma^2) \propto -\frac{\nu+1}{2} \sum_{i=1}^n \log \left(1 + \frac{(y_i - X_i^T \beta)^2}{\nu \sigma^2} \right).$$

De manera análoga al caso anterior, no existe una distribución *a priori* conjugada simple, pero la distribución de los errores se puede escribir como una mezcla de una distribución normal. Específicamente, si

$$\varepsilon_i | \lambda_i = l \sim \mathcal{N}(0, \sigma^2/l), \quad \lambda_i \sim \text{Gamma} \left(\frac{\nu}{2}, \frac{\nu}{2} \right), \quad (10)$$

en donde los λ_i son independientes, entonces al marginalizar sobre λ_i se obtiene la distribución t de Student con ν grados de libertad y escala σ^2 . Esto permite que, condicional a λ , β siga una distribución normal, y de nuevo se puede hacer la inferencia por métodos computacionales, como lo son:

- **MCMC/Muestreo de Gibbs:** Dado λ , si la distribución *a priori* de β es gaussiana, la distribución *a posteriori* también lo es. Además, para obtener distribuciones conjugadas, a σ^2 se le puede asignar una Gamma inversa (como en el caso usual) y λ_i una Gamma, como antes, y se puede hacer el muestreo de Gibbs de estas distribuciones y posteriormente el cálculo usual.
- **Algoritmo EM:** También conocido como **expectation-maximization algorithm**, es un método iterativo que permite estimar el MAP (máximo *a posteriori*) cuando el modelo depende de variables latentes no observadas. En este caso se puede tratar a las λ_i como variables latentes y realizar dicha implementación, y se puede ver como un algoritmo iterativo de mínimos cuadrados ponderados, lo cual surge de la expresión de la logverosimilitud dada anteriormente.
- **Mean-field VI:** En este caso también funcionan bien los métodos de aproximación variacional, considerando los λ_i como variables latentes.

Por último, a los errores se les pueden asignar otras distribuciones, como la Cauchy, que tiene colas muy pesadas, lo cual lo hace sumamente robusto a outliers, una mezcla de normales, que permite introducir la noción de “muestras contaminadas”, entre otros. Sin embargo, esto hace que el análisis conjugado no sea factible, y se requieran métodos computacionales, como el MCMC. \square