Regresión Lineal y Bayesiana

INTRODUCCIÓN A CIENCIA DE DATOS

13 de Octubre de 2025

Jessica Rubí Lara Rosales jessica.lara@cimat.mx
Rodrigo Gonzaga Sierra rodrigo.gonzaga@cimat.mx

1. Regresión lineal ordinaria (OLS)

1.1. Derivación del estimador OLS

Partiendo del modelo clásico

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

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. Deseamos encontrar el vector β tal que la siguiente función se minimice

$$SS_{Res}(\beta) = \sum_{i=1}^{n} (y_i - [X\beta]_i)^2 = (y - X\beta)^T (y - X\beta) = ||y - X\beta||^2.$$

Para ello notemos que esta diferencia la podemos ver como:

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

Derivando la expresión obtenida respecto de β y usando que

$$\begin{aligned} \frac{\partial}{\partial \beta}(a^T \beta) &= a \\ \frac{\partial}{\partial \beta}(\beta^T A \beta) &= (A + A^T)\beta \end{aligned}$$

se tiene

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

Igualando a cero y suponiendo que $X^T X$ es invertible, es decir, X es de rango completo se tiene que el punto critico $\hat{\beta}$ es

$$-2X^T y + 2X^T X\hat{\beta} = 0 \quad \Leftrightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T y.$$

Haciendo la segunda derivada de SS_{Res} obtenemos

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

Lo cual cumple que es definida positiva si y solo si $X^T X$ es de rango completo y esto ocurre si y solo si $X^T X$ es invertible o bien X es de rango completo. Esto es cierto, pues si X es de rango completo entonces para $v \neq 0$ se cumple que $w := Xv \neq 0$ lo que implica que

$$v^T X^T X v = (Xv)^T Xv = w^T w = \sum_i w_i^2 > 0.$$

De ello podemos concluir que en efecto

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

es el estimador de Minimos Cuadrados Ordinarios.

■

1.2. Propiedades del estimador

Calcule explícitamente:

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

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

Solución. Recordando que $\mathbb{E}[y] = \mathbb{E}[X\beta + \varepsilon] = X\beta$ pues $\varepsilon \sim N(0, \sigma^2 I)$ se tiene que

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

Así que, $\hat{\beta}$ es insesgado. Por otro lado, la varianza del estimador $\hat{\beta}$ esta dada por

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

El teorema de Gauss-Markov establece que el estimador de minimos cuadrados ordinarios (OLS, ordinary least-squeares) de β es el mejor estimado lineal insesgado (BLUE, Best Linear unbiased estimator). Con mejor estimador nos referimos a que $\hat{\beta}$ tiene la mínima varianza entre la clase de todos los estimadores insesgados que son combinaciones lineales de los datos. Esto es

$$Var(\hat{\theta}) \leq Var(\tilde{\theta})$$

donde $\tilde{\theta}$ es otro estimador insesgado. Para verificar que esto se cumple para $\hat{\beta}$, tomemos $\tilde{\beta} = Cy$ otro estimador insesgado de β donde C es una matriz de $p \times n$. Como

$$\mathbb{E}[\tilde{\beta}] = C\mathbb{E}[y] = CX\beta$$

Se debe de cumplir que $CX = I_p$. Por otro lado notemos que

$$Var(\tilde{\beta}) = Var(Cy) = CVar(y)C^T = CC^T\sigma^2.$$

Sumando un cero conveniente tenemos

$$C = (X^TX)^{-1}X^T + C - (X^TX)^{-1}X^T = (X^TX)^{-1}X^T + D,$$

donde $D = C - (X^TX)^{-1}X^T$. De ello que

$$\begin{aligned} CC^T &= ((X^TX)^{-1}X^T + D)((X^TX)^{-1}X^T + D)^T \\ &= ((X^TX)^{-1}X^T + D)(X(X^TX)^{-1} + D^T) \\ &= (X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DX(X^TX)^{-1} + DD^T \\ &= (X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DX(X^TX)^{-1} + DD^T. \end{aligned}$$

Notando que

$$DX = (C - (X^TX)^{-1}X^T)X = CX - (X^TX)^{-1}X^TX = I_p - I_p = 0$$

y como $X^TD^T = (DX)^T = 0$, se sigue que

$$CC^T = (X^TX)^{-1} + DD^T.$$

Por lo tanto

$$Var(\tilde{\beta}) = \left[(X^TX)^{-1} + DD^T\right]\sigma^2$$

Así que

$$Var(\tilde{\beta}) - Var(\hat{\beta}) = DD^T\sigma^2.$$

Como para cualquier vector v se cumple que

$$v^T(DD^T)v = (D^Tv)^T(D^Tv) = ||D^Tv|| \geq 0$$

Tenemos que DD^T es semidefinida positiva, por lo tanto

$$Var(\hat{\beta}) \leq Var(\tilde{\beta})$$

Esto implica que $\hat{\beta}$ es BLUE. ■

Se puede demostrar que $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$ también máxima la verosimilitud. Además, cuando $n \to \infty$, bajo condiciones de regularidad:

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0 \quad \text{y} \quad \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, I^{-1}(\boldsymbol{\beta}_0))$$

donde $I(\boldsymbol{\beta})$ es la matriz de información de Fisher.

2. Regresión lineal bayesiana (prior conjugado)

2.1. Prior conjugado

Suponga un prior conjugado:

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

2.2. Distribución posterior

Derive los parámetros posteriores (β_n, V_n, a_n, b_n) y escriba la forma explícita de la posterior conjunta

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

Solución. Como

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

tomando una muestra $y_1, \ldots, y_n$ independiente de esta variable, definiendo $y = (y_1, \ldots, y_n)$ se tiene que la función de verosimilitud esta dada por

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

Por hipótesis las distribuciones a prioris de β y σ^2 son respectivamente

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

y

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

donde p es la dimensión del vector β, es decir, el numero de coeficientes de regresión incluyendo al intercepto. Por el teorema de Bayes

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

Desarrollando cuadrados

$$\begin{aligned}
||y - X\beta||^2 + (\beta - \beta_0)^T V_0^{-1}(\beta - \beta_0) &= y^T y - 2y^T X\beta + \beta^T X^T X\beta + \beta^T V_0^{-1}\beta - 2\beta_0^T V_0^{-1}\beta + \beta_0^T V_0^{-1}\beta_0 \\
&= \beta^T X^T X\beta + \beta^T V_0^{-1}\beta - 2y^T X\beta - 2\beta_0^T V_0^{-1}\beta + y^T y + \beta_0^T V_0^{-1}\beta_0 \\
&= \beta^T (X^T X + V_0^{-1})\beta - 2\beta^T (X^T y + V_0^{-1}\beta_0) + (y^T y + \beta_0^T V_0^{-1}\beta_0)
\end{aligned}$$

Tomando

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

Completando cuadrados

$$||y - X\beta||^2 + (\beta - \beta_0)^T V_0^{-1}(\beta - \beta_0) = (\beta - \beta_n)^T V_n^{-1}(\beta - \beta_n) + y^T y + \beta_0^T V_0^{-1}\beta_0 - \beta_n^T V_n^{-1}\beta_n$$

De ello que

$$\exp\left[-\frac{1}{2\sigma^2}((\beta - \beta_n)^T V_n^{-1}(\beta - \beta_n) + y^T t\beta_0^T V_0^{-1}\beta_0 - \beta_n^T V_n^{-1}\beta_n + 2b_0)\right]$$

Definiendo

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

Se obtiene que la distribución posterior es

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

■

2.3. Distribuciones marginales

Identifique las distribuciones marginales de β y de σ^2.

Solución. Para sacara la distribución marginal de $\sigma^2|y$ tenemos que integrar β, pero podemos notar que esto justamente va a ser integrar una $N(\beta_n, \sigma^2 V_n)$ por lo tanto

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

De ello que

$$\sigma^2|y \sim InvGamma(a_n, b_n).$$

Por otro lado, para la distribución marginal de β integramos la expresión original de $\beta, \sigma^2|y$ respecto a σ^2 y obtenemos

$$\begin{aligned} p(\beta|y) &= \int_0^\infty p(\beta, \sigma^2|y) d\sigma^2 \\ &\propto \int_0^\infty (\sigma^2)^{-(a_n+p/2+1)} \exp\left[-\frac{1}{\sigma^2}\left(b_n + \frac{1}{2}(\beta-\beta_n)^T V_n^{-1}(\beta-\beta_n)\right)\right] d\sigma^2 \end{aligned}$$

Definiendo $f(\beta) = b_n + \frac{1}{2}(\beta-\beta_n)^T V_n^{-1}(\beta-\beta_n)$ tenemos que

$$p(\beta|y) \propto \int_0^\infty (\sigma^2)^{-(a_n+p/2+1)} \exp\left[-\frac{f(\beta)}{\sigma^2}\right] d\sigma^2$$

Usando que la función Gamma es

$$\Gamma(\alpha)A^{-\alpha} = \int_0^\infty t^{-\alpha-1} e^{-A/t} dt$$

donde A es una matriz conformable (de dimensión adecuadas). Se sigue que

$$\begin{aligned} p(\beta|y) &\propto \Gamma(a_n + p/2)\left[f(\beta)\right]^{-(a_n+p/2)} \\ &\propto \left[b_n + \frac{1}{2}(\beta-\beta_n)V_n^{-1}(\beta-\beta_n)\right]^{-(a_n+p/2)} \\ &= \left[1 + \frac{(\beta-\beta_n)V_n^{-1}(\beta-\beta_n)}{2b_n}\right]^{-(a_n+p/2)} \end{aligned}$$

Lo cual coinicide con el kernel de una distribución t- student multivariada con parametros

$$\beta|y \sim t_{2_{a_n}}\left(\beta_n, \frac{b_n}{a_n}V_n\right)$$

■

3. Conexión con regularización

3.1. Regresión Ridge

Muestre que si se toma una priori Normal isotrópico

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

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

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

Solución. Suponemos que $Y|\beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I_n)$, tomando una muestra $y_1, \ldots, y_n$ independiente de esta variable y definiendo $y = (y_1, \ldots, y_n)$ se tiene que la función de verosimilitud está dada por

$$\begin{aligned} p(y|\beta) &= \frac{1}{(2\pi)^{n/2}|\sigma^2 I_n|^{1/2}} \exp\left(-\frac{1}{2}(y - X\beta)^T(\sigma^2 I_n)^{-1}(y - X\beta)\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right) \end{aligned}$$

Además, como por hipótesis

$$\beta \sim N_p(0, \tau^2 I_p).$$

Se tiene que la distribución a priori es

$$p(\beta) = \frac{1}{(2\pi)^{p/2}|\tau^2 I_p|^{1(2}} \exp\left(-\frac{1}{2}\beta^T(\tau^2 I_p)^{-1}\beta\right) \propto \exp\left(-\frac{1}{2\tau^2}\|\beta\|^2\right)$$

Entonces por el teorema de Bayes la distribución posterior es

$$\begin{aligned} p(\beta|y) &\propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right)\exp\left(-\frac{1}{2\tau^2}\|\beta\|^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \frac{1}{2\tau^2}\|\beta\|^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\|y - X\beta\|^2 + \lambda\|\beta\|^2\right)\right), \end{aligned}$$

donde $\lambda = \frac{\sigma^2}{\tau^2}$. Notemos que encontrar el β que maximiza la posterior es equivalente a minimizar la parte dentro de los parentesis del exponente de la exponencial, esto es

$$\hat{\beta}_{MAP} = \arg\max_{\beta} p(y|\beta) = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda||\beta||^2$$

Obteniendo así que el estimador de máxima posterior (MAP) es equivalente a la regresión Ridge. Esto nos dice que el nivel de regularización en Ridge está determinado por la relación (o razón) entre la varianza de los datos (verosimilitud) y la varianza de la distribución a priori. ■

3.2. 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:

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

Solución. Usando nuevamente que $Y|\beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I_n)$ y tomando una muestra $y_1, \ldots, y_n$ independiente de esta variable y definiendo $y = (y_1, \ldots, y_n)$ por independencia su función de verosimilitud es

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

Como estamos suponiendo que las β_j son independientes tenemos que la distribución a priori es

$$p(\beta) \propto \prod_{j=1}^{n} \exp(-\lambda|\beta_j|) = \exp\left(-\lambda \sum_j |\beta_j|\right) = \exp\left(-\lambda \|\beta\|_1\right)$$

Así que del teorema de Bayes

$$\begin{aligned} p(\beta|y) \propto p(y|\beta, \sigma^2)p(\sigma^2) &= \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right)\exp\left(-\lambda\|\beta\|_1\right) \\ &= \exp\left(-\left(\|y - X\beta\|^2 + c\|\beta\|_1\right)\right), \end{aligned}$$

donde $c = 2\sigma^2\lambda$. Definiendo

$$\hat{\beta}_{MAP} = \arg\max_{\beta} p(\beta|y)$$

Es equivalente a minimizar el negativo del exponente, esto es

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \|y - X\beta\|^2 + c\|\beta\|_1$$

De ello que el estimador MAP corresponde a la regresión Lasso.

■

4. Extensiones: errores no normales

De aucerdo con [Lange et al., 1989] y [Gelman et al., 1995], la inferencia estadística basada en la distribución normal (univariada o multivariada) es vulnerable a los valores atípicos. A pesar de este hecho y del considerable interés en los procedimientos robustos en la literatura estadística matemática, la mayoría de los análisis estadísticos aplicados continúan basándose en el modelo normal. Incluso en la regresión lineal, donde la robustez ha sido un tema de mucho interés, incursionando en el software estadístico disponible para los profesionales, los procedimientos se dirigen principalmente a detectar valores atípicos.

Por lo visto en [Lopez, 2024], se ha dependido principalmente de las distribuciones normal, binomial y de Poisson, y combinaciones de estas, para modelar datos y parámetros. Sin embargo, el uso de una clase limitada de distribuciones da como resultado una clase limitada y potencialmente inapropiada de inferencias. Muchos problemas caen fuera del rango de los modelos convenientes, y los modelos deben elegirse para ajustarse a la ciencia subyacente y a los datos, no simplemente por su conveniencia analítica o computacional.

4.1. Modelo alternativo

En el marco bayesiano, se puede reducir la influencia de una observación aberrante reemplazando el modelo de población normal para los θ_j por una familia de distribuciones de colas pesadas, que permite la posibilidad de observaciones extremas. Por colas largas, nos referimos a una distribución con un contenido de probabilidad relativamente alto lejos del centro, donde la escala de "lejos" se determina, por ejemplo, en relación con el diámetro de una región que contiene el 50 % de la probabilidad en la distribución. Ejemplos de distribuciones de colas largas incluyen la familia de distribuciones t, de la cual el caso más extremo es la Cauchy o t_1, y también los modelos de mezcla (finitos), que generalmente usan una distribución simple como la normal para la mayor parte de los valores pero permiten una probabilidad discreta de observaciones o valores de parámetros de una distribución alternativa que puede tener un centro diferente y generalmente tiene una dispersión mucho mayor.

Observe si se aplica un método para realizar inferencia robusta en una variedad de conjuntos de datos reales. Por ejemplo, enfóquese en reemplazar la distribución normal por la distribución **t-student** en los modelos estadísticos. Específicamente, suponga que los datos muestrales y_i, para $(1 \leq i \leq n)$ se registran para n unidades. Típicamente, se asume que las y_i son vectores aleatorios normales independientes. Si $N_k(\mu, \Sigma)$ denota la distribución normal k-variante con media μ y matriz de covarianza Σ, entonces

$$y_i \overset{\text{ind}}{\sim} N_{v_i}(\mu_i(\theta), \Sigma_i(\varphi)), \tag{1}$$

donde v_i es el número de componentes de y_i, que puede variar de una unidad a otra en algunas aplicaciones, μ_i es un vector de media $(v_i \times 1)$ de forma conocida indexado por un conjunto de parámetros desconocidos θ, y Σ_i es una matriz de covarianza $(v_i \times v_i)$ de forma conocida indexada por un conjunto de parámetros desconocidos φ. Las funciones μ_i y Σ_i pueden involucrar covariables fijas conocidas x_i, registradas para cada unidad i. Proponemos reemplazar (1) con el modelo

$$y_i \overset{\text{ind}}{\sim} t_{v_i}(\mu_i(\theta), \Psi_i(\varphi), \nu), \tag{2}$$

donde $t_v(\mu, \Psi, \nu)$ denota la distribución t k-variante [Dunnett and Sobel, 1954] con vector de ubicación μ, matriz de escala Ψ, ν grados de libertad (gl), y densidad

$$p(y \mid \mu, \Psi, \nu) = \frac{|\Psi|^{-1/2}\Gamma((\nu+k)/2)}{[\Gamma(1/2)]^k\Gamma(\nu/2)\nu^{k/2}} \times \left(1 + \frac{(y-\mu)^T\Psi^{-1}(y-\mu)}{\nu}\right)^{-(\nu+k)/2}.$$

Las inferencias sobre θ y φ en el contexto multivariante t pueden proceder mediante métodos de verosimilitud análogos a aquellos para el modelo normal (1).

Los siguientes hechos conocidos sobre la t multivariante son instructivos y se utilizan más adelante. Supongamos que $y \mid u \sim N_k(\mu, \Psi/u)$ para un escalar $u \sim \chi^2_\nu/\nu$, donde ν es positivo y puede ser un no entero. Entonces tenemos las siguientes propiedades.

Propiedad 1: $y \sim t_k(\mu, \Psi, \nu)$.

Propiedad 2: $E(y) = \mu$ $(\nu > 1)$ y $\text{cov}(y) \equiv \Sigma = \nu\Psi/(\nu - 2)$ $(\nu > 2)$.

Propiedad 3: $u \mid y \sim \chi^2_{\nu+k}/(\nu + \delta^2)$, donde $\delta^2 = (y - \mu)^T\Psi^{-1}(y - \mu)$.

Propiedad 4: $\delta^2/k \sim F_{k,\nu}$.

Nótese que la distribución t multivariante se aproxima a la distribución normal con matriz de covarianza Ψ cuando $\nu \to \infty$. Cuando $\nu < \infty$, la estimación de máxima verosimilitud (MV) de θ y ciertas funciones de φ son robustas en el sentido de que los casos atípicos con grandes distancias de Mahalanobis $\delta_i^2 = (y_i - \mu_i)^T\Psi_i^{-1}(y_i - \mu_i)$ son ponderados hacia abajo. En particular, las estimaciones de MV de θ (con q componentes, por ejemplo) para el modelo normal (1) satisfacen la ecuación de verosimilitud $\partial l/\partial\theta = \sum_{i=1}^n A_i\Sigma_i^{-1}(y_i - \mu_i) = 0$, donde l denota la log-verosimilitud y A_i es la matriz $(q \times v_i)$ de derivadas parciales de μ_i con respecto a θ. Las estimaciones de MV de θ bajo el modelo t (2) satisfacen $\sum_{i=1}^n w_i A_i\Psi_i^{-1}(y_i - \mu_i) = 0$, donde

$$w_i = (\nu + v_i)/(\nu + \delta_i^2)$$

es el peso asignado al caso i; w_i claramente disminuye al aumentar δ_i^2.

4.2. Familias conjugadas

Si bien la elección de una distribución de probabilidad para modelar nuestra incertidumbre sobre θ no resulta crucial en tanto sea factible elicitar con cualquiera de ellas una distribución a priori, resulta conveniente tanto para el análisis como desde un punto de vista computacional el que $p(\theta)$ y $p(\theta|x)$ pertenezcan a la misma familia [Erdely, 2023].

La definición y la construcción de una familia conjugada depende de la existencia e identificación de estadísticos suficientes de dimensión finita para una función de verosimilitud dada. Si existe este estadístico suficiente entonces la dimensionalidad puede ser reducida. Cuando existe el **estadístico suficiente**, entonces existe una familia conjugada [Correa Morales, 2018]. Esto se resume en el siguiente teorema:

Teorema [Mendoza, 2011]: Sea X una v.a. con f.d.p. $f(x \mid \theta)$, $\theta \in \Theta$. Suponga que para toda m.a. $x^{(n)}$ de X existe una estadística $T_n(x^{(n)})$ de dimensión fija r que es suficiente para θ. Si como consecuencia del teorema de factorización se tiene que $p(x^{(n)} \mid \theta) = h(x^{(n)})g(T_n(x^{(n)}), \theta)$ y además se cumple que $\int_\Theta g(T_n(x^{(n)}), \theta)d\theta < \infty$, entonces existe una familia paramétrica conjugada (básica) para θ.

A continuación, se mostrará que las distribuciones t-Student y Laplace, no tienen estadísticas suficientes finitas.

4.2.1. Distribución t-student

La función de densidad de una distribución t-Student con ν grados de libertad es:

$$f(x \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R}.$$

Para una muestra $X_1, \ldots, X_n$ i.i.d., la verosimilitud es:

$$L(\nu) = \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\right]^n \prod_{i=1}^{n} \left(1 + \frac{x_i^2}{\nu}\right)^{-(\nu+1)/2}.$$

Reescribiendo:

$$L(\nu) = \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\right]^n \cdot \nu^{-n(\nu+1)/2} \cdot \prod_{i=1}^{n} \left(\nu + x_i^2\right)^{-(\nu+1)/2}.$$

El producto $\prod_{i=1}^{n}(\nu + x_i^2)$ depende de los datos de manera no trivial y **no** puede reducirse a una función de un estadístico de dimensión fija independiente de n y ν, pues ν aparece tanto fuera como dentro del producto de forma no exponencial. Por lo tanto, la distribución t-Student **no pertenece a la familia exponencial** en el parámetro ν. No existe un estadístico suficiente de **dimensión fija** para ν.

4.2.2. Distribución de Laplace

La función de densidad es:

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right), \quad b > 0.$$

La verosimilitud es:

$$L(\mu) = (2b)^{-n} \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |x_i - \mu|\right).$$

Aquí pueden ocurrir tres casos (pero el general es el tercero):

- **Caso 1**: b conocido, μ desconocido. Con esto, el término $\sum_{i=1}^{n} |X_i - \mu|$ **depende del parámetro** μ, por lo que no es un estadístico. No existe un estadístico suficiente de dimensión fija para μ, por lo tanto no se puede garantizar la existencia de su familiar conjugada.

- **Caso 2**: μ conocido, b desconocido.

  $$L(b) = (2b)^{-n} \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |x_i - \mu|\right).$$

  Reescribiendo:

  $$L(b) = (2)^{-n} \cdot b^{-n} \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |x_i - \mu|\right).$$

  Esto es de la familia exponencial con:

$$T(X) = \sum_{i=1}^{n} |X_i - \mu|$$
$$\eta(b) = -\frac{1}{b}$$
$$A(b) = n \ln b$$

Entonces $T(X)$ es suficiente para b. Por lo que, la verosimilitud es proporcional a:

$$L(b) \propto b^{-n} \exp\left(-\frac{T}{b}\right).$$

Esta forma corresponde a la distribución **Gamma Inversa**. Si una distribución a priori para b es:

$$\pi(b) \propto b^{-(\alpha+1)} \exp\left(-\frac{\beta}{b}\right),$$

es decir, $b \sim$ Gamma-Inversa(α, β), entonces:

$$\pi(b \mid \mathbf{x}) \propto b^{-(n+\alpha+1)} \exp\left(-\frac{T+\beta}{b}\right).$$

Por lo tanto:

$$b \mid \mathbf{x} \sim \text{Gamma-Inversa}(\alpha + n, \beta + T).$$

La familia conjugada para b (con μ conocido) es la **Gamma Inversa**. Pero pedir que un hiperparámetro sea conocido es pedirle mucho al modelo, por lo que se sigue el siguiente caso.

- **Caso 3**: ambos parámetros desconocidos

$$L(\mu, b) = (2b)^{-n} \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |x_i - \mu|\right).$$

No hay estadístico suficiente. Por lo que Para b, la conjugada condicional es Gamma Inversa, para μ no hay conjugada estándar, por lo tanto no existe una familia conjugada conjunta para (μ, b)

4.3. Inferencia vía simulación

Siguiendo a [Correa Morales, 2018], muchos de los problemas que ocurren a través del enfoque bayesiano requieren de solución a través de métodos numéricos. Esto debido a que los modelos *a posteriori* resultantes en los análisis, usualmente son de alta complejidad y no tienen solución **cerrada**. Es por esto, que toma mucha importancia la implementación de métodos numéricos para llegar a la solución de nuestros problemas. Suponga que se tienen los parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ de la distribución *a posteriori*:

$$f(\boldsymbol{\theta} \mid \mathbf{x}^n) \propto L_n(\boldsymbol{\theta}) f(\boldsymbol{\theta}). \tag{11.8}$$

Surge entonces la pregunta de cómo extraer inferencias sobre un parámetro individual. La clave está en encontrar la densidad posterior marginal para el parámetro de interés.

$$f(\theta_1 \mid \mathbf{x}^n) = \int \cdots \int f(\theta_1, \cdots, \theta_p \mid \mathbf{x}^n)\, d\theta_2 \ldots d\theta_p. \tag{11.9}$$

En la práctica, puede no ser factible calcular esta integral. La simulación puede ayudar.

4.3.1. MCMC: Monte Carlo por Cadenas de Markov

Conforme a [Correa Morales, 2018], los métodos MCMC son algoritmos iterativos que se utilizan cuando el muestreo directo de una distribución de interés ξ no es factible. La aproximación a la teoría de cadenas de Markov, es iniciar con alguna distribución de transición (una matriz de transición en el caso discreto) que modela algún proceso de interés, para determinar las condiciones bajo la cual existe una distribución estacionaria o invariante y entonces identificar la forma de la distribución límite. Los métodos MCMC involucran la solución inversa de este problema ya que la distribución estacionaria es conocida, y es la distribución de transición la que necesita ser determinada, a pesar que el la práctica existen un número infinito de distribuciones de las cuales escoger.

Una cadena de Markov es generada muestreando

$$\theta^{(t+1)} \sim p\left(\theta|\theta^{(t)}\right)$$

Este p es llamado el kernel de transición de la cadena de Markov. Así $\theta^{(t+1)}$ depende solo de $\theta^{(t)}$, y no de $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(t-1)}$.

Existen dos problemas mayores que rodean la implementación e inferencias de los métodos MCMC. El primero tiene que ver con la convergencia y el segundo con la dependencia entre las muestras de la distribución posterior.

Hay una extensa lista de MCMC:

- Metropolis-Hastings
- Gibbs Sampling
- Hamiltonian Monte Carlo (HMC)
- No-U-Turn Sampler (NUTS)
- Leapfrog
- T-walk

Existe todo un estudio acerca de los métodos MCMC, que ya no se mencionará en este trabajo.

Referencias

[Correa Morales, 2018] Correa Morales, Juan Carlos y Barrera Causil, C. J. (2018). *Introducción a la estadística bayesiana: notas de clase*. Instituto Tecnológico Metropolitano.

[Dunnett and Sobel, 1954] Dunnett, C. W. and Sobel, M. (1954). A bivariate generalization of student's t-distribution, with tables for certain special cases. *Biometrika*, 41(1-2):153–169.

[Erdely, 2023] Erdely, A. y Gutiérrez-Peña, E. (2023). Monograf\'ia de estad\'istica bayesiana. *arXiv preprint arXiv:2309.06601*.

[Gelman et al., 1995] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

[Lange et al., 1989] Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896.

[Lopez, 2024] Lopez, M. (2024). Presentation_10 - diapositivas de ciencia de datos. Recuperado de GitHub. Diapositivas de clase del curso de Ciencia de Datos.

[Mendoza, 2011] Mendoza, Manuel y Regueiro, P. (2011). Estadística bayesiana. *Instituto Teconológico Autónomo de México*.