Tarea 3: Regresión Lineal y Bayesiana

Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Alumnos: Canché, Elías; Sánchez, Hazel; Aguilar, Guillermo.
Profesor: Dr. Marco Antonio Aquino López.

Instrucciones Generales.

Resuelva cuidadosamente cada apartado. Todas las demostraciones deben presentarse paso a paso, con claridad en las justificaciones matemáticas y con una breve interpretación estadística de los resultados.

1 Regresión lineal ordinaria (OLS)

Problem 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

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

siempre que $\mathrm{X}^T X seainvertible.$

Solución.

Partamos del modelo clásico donde $y \in \mathbb{R}^n$ es el vector de observaciones de la variable dependiente; $X \in \mathbb{R}^{n \times p}$, con $p \leq n$ y rango$(X) = p$; $\beta \in \mathbb{R}^p$ es el vector de parámetros desconocidos a estimar. $\varepsilon \in \mathbb{R}^n$ es el vector de perturbaciones aleatorias.

El principio de Mínimos Cuadrados Ordinarios consiste en encontrar el estimador $\hat{\beta}$ que minimiza la suma de los residuos al cuadrado. Definimos la función objetivo:

$$S(\beta) = \|y - X\beta\|^2 = (y - X\beta)^T (y - X\beta)$$

Expandiendo el producto, obtenemos

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

Dado que $S(\beta)$ es una función estrictamente convexa en β (debido a que $X^T X$ es definida positiva bajo el supuesto de rango$(X) = p$), el mínimo global se alcanza cuando el gradiente de $S(\beta)$ con respecto a β es igual a cero. Entonces,

$$\nabla_\beta S(\beta) = -2X^T y + 2X^T X \beta = 0,$$

dividiendo entre 2 y reordenando:

$$X^T X\beta = X^T y.$$

Dado que rango$(X) = p$, la matriz $X^T X \in \mathbb{R}^{p\times p}$ es simétrica, definida positiva y, por lo tanto, invertible. Multiplicando ambos lados de la ecuación normal por $(X^T X)^{-1}$

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

simplificando, llegamos a que

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

Por lo tanto, el estimador de Mínimos Cuadrados Ordinarios es

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

siempre que $(X^T X)$ sea invertible.

Problem 2: Propiedades del estimador

Calcule explícitamente:

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

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

Solución.

Partimos del estimador de MCO

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

Por linealidad,

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

Dado que $\varepsilon \sim N(0, \sigma^2 I)$, entonces $E[\varepsilon] = 0$, así

$$E[\hat{\beta}] = \beta + (X^\top X)^{-1} X^\top E[\varepsilon] = \beta,$$

por tanto $\hat{\beta}$ es insesgado.

Para calcular la varianza nótese que $\mathrm{Var}(\varepsilon) = \sigma^2 I$, por lo que

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

Tomemos otro estimador $\tilde{\beta} = Ay$, con A de tamaño $p\times n$, y supongamos que es insesgado: $E[\tilde{\beta}] = \beta$. La condición de insesgadez implica $AX = I_p$ (pues $E[\tilde{\beta}] = AX\beta = \beta$ para todo β). Escribimos

$$A = (X^\top X)^{-1} X^\top + M,$$

de modo que $MX = 0$. Entonces

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2 AA^\top = \sigma^2 \big((X^\top X)^{-1} X^\top + M\big)\big(X(X^\top X)^{-1} + M^\top\big).$$

Al expandir y usar $MX = 0$ se obtiene

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1} + \sigma^2 M M^\top.$$

Como $\sigma^2 M M^\top$ es semidefinida positiva, se cumple

$$\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta}) = \sigma^2 M M^\top \succeq 0.$$

Por tanto $\hat{\beta}$ minimiza la matriz de varianza entre los estimadores lineales insesgados, es decir, es la *mejor* (en varianza) dentro de esa clase.

2 Regresión lineal bayesiana (prior conjugado)

Problem 3: Prior conjugado

Suponga un prior conjugado:

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

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.

Considere el modelo lineal usual $y \mid \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n)$, con X de tamaño $n \times p$. La verosimilitud es

$$p(y \mid \beta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta) \Big\}.$$

El prior conjunto (condicional) es

$$p(\beta, \sigma^2) = p(\beta \mid \sigma^2) p(\sigma^2) \propto (\sigma^2)^{-p/2} \exp\Big\{ -\frac{1}{2\sigma^2}(\beta - \beta_0)^\top V_0^{-1} (\beta - \beta_0) \Big\} \, (\sigma^2)^{-(a_0+1)} \exp\Big\{ -\frac{b_0}{\sigma^2} \Big\}.$$

Multiplicando verosimilitud por prior y agrupando potencias de σ^2 y exponentes obtenemos el kernel de la posterior

$$p(\beta, \sigma^2 \mid y) \propto (\sigma^2)^{-\left(a_0 + \frac{n}{2} + \frac{p}{2} + 1\right)} \exp\Big\{ -\frac{1}{2\sigma^2} \Big[(y - X\beta)^\top (y - X\beta) + (\beta - \beta_0)^\top V_0^{-1} (\beta - \beta_0) + 2b_0 \Big] \Big\}.$$

Ahora completamos el cuadrado en β dentro del corchete. Podemos derivar los parámetros

$$V_n = (V_0^{-1} + X^\top X)^{-1}, \qquad \beta_n = V_n \big(V_0^{-1} \beta_0 + X^\top y \big).$$

Al completar el cuadrado se obtiene la siguiente descomposición

$$(y - X\beta)^\top (y - X\beta) + (\beta - \beta_0)^\top V_0^{-1} (\beta - \beta_0) = (\beta - \beta_n)^\top V_n^{-1} (\beta - \beta_n) + S_n,$$

donde

$$S_n = y^\top y + \beta_0^\top V_0^{-1} \beta_0 - \beta_n^\top V_n^{-1} \beta_n.$$

Por tanto la posterior conjunta toma la forma

$$p(\beta, \sigma^2 \mid y) \propto (\sigma^2)^{-(a_n+1)} \exp\Big\{ -\frac{1}{2\sigma^2}(\beta-\beta_n)^\top V_n^{-1}(\beta-\beta_n)\Big\} \exp\Big\{ -\frac{b_n}{\sigma^2}\Big\},$$

con

$$a_n = a_0 + \frac{n}{2}, \qquad b_n = b_0 + \frac{1}{2}S_n = b_0 + \frac{1}{2}\big(y^\top y + \beta_0^\top V_0^{-1}\beta_0 - \beta_n^\top V_n^{-1}\beta_n\big).$$

Problem 4: Distribuciones marginales

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

Solución.

De la forma de la posterior conjunta, podemos ver la factorización

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

donde

$$\beta \mid \sigma^2, y \sim N(\beta_n,\ \sigma^2 V_n), \qquad \sigma^2 \mid y \sim \text{Inv-Gamma}(a_n, b_n).$$

3 Conexión con regularización

Problem 5: Regresión Ridge

Muestre que si se toma un prior Normal isotrópico:

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

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

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

Solución.

Dado β, la distribución de y es $y \mid \beta \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Luego, la función de verosimilitud es

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

Tomamos un prior Normal isotrópico dado por

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

Por el teorema de Bayes $p(\beta \mid y) \propto p(y \mid \beta)\, p(\beta)$ y sustituyendo obtenemos

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

multiplicando por $2\sigma^2$ y utilizando la función arg min sobre β llegamos a que

$$\hat{\beta}_{\text{MAP}} = \arg\min_{\beta} \left[\|y - X\beta\|^2 + \frac{\sigma^2}{\tau^2} \|\beta\|^2 \right].$$

Comparando con la formulación de Ridge

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|^2$$

obtenemos $\lambda = \frac{\sigma^2}{\tau^2}$. Dado que $\sigma \neq 0$ y $\tau \neq 0$, entonces λ está bien definido y así, el estimador MAP es equivalente al estimador Ridge con parámetro de regularización $\lambda = \frac{\sigma^2}{\tau^2}$.

Problem 6: 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:

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

Solución.

Análogamente al problema anterior, dado β, la distribución de y es $y \mid \beta \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Luego, la función de verosimilitud es

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

Tomamos un prior Laplace para β, dado por

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

Por el teorema de Bayes $p(\beta \mid y) \propto p(y \mid \beta)\, p(\beta)$ y sustituyendo se obtiene

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

multiplicando por $2\sigma^2$ y tomando la función arg min sobre β llegamos a que

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

Nótese que definiendo $\tilde{\lambda} = 2\sigma^2\lambda$, obtenemos

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

lo cual, es la formulación de la Regresión Lasso. Renombrando $\tilde{\lambda}$ como λ, tenemos:

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

Por lo tanto, bajo el prior de Laplace el estimador MAP es equivalente con el estimador Lasso.

4 Extensiones: errores no normales

Problem 7: Modelos alternativos

Proponga un modelo de regresión donde el error ε no siga una distribución Normal. Ejemplos:

$$\varepsilon \sim \text{Laplace}(0, b) \quad \text{(robusto a outliers)},$$

$$\varepsilon \sim \text{Student-t}(\nu) \quad \text{(colas pesadas)}.$$

Solución.

Regresión con errores Laplace

A continuación propondremos un modelo de regresión lineal donde el error ε se distribuye con una distribución Laplace$(0, b)$, donde la función de densidad de Laplace es

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

entonces, la función de verosimilitud para una observación es

$$p(y_i \mid \beta, b) = \frac{1}{2b} \exp\left(-\frac{|y_i - x_i^\top \beta|}{b}\right),$$

luego, de manera general se obtiene que

$$p(y \mid \beta, b) = \left(\frac{1}{2b}\right)^n \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |y_i - x_i^\top \beta|\right).$$

Luego, el estimador de máxima verosimilitud es

$$\hat{\beta}_{\text{MLE}} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i^\top \beta|.$$

Lo anterior, en la literatura se lo conoce como **regresión cuantil** para el percentil 0.5 (mediana). Además, nótese que este modelo minimiza la suma de valores absolutos en lugar de cuadrados, lo cual lo hace sensible a datos extremos y por lo tanto robusto para detectar outliers.

Regresión con errores Student-t

Supongamos que los errores ε tienen una distribución Student, $\varepsilon \sim \text{Student-t}(\nu)$, donde $\nu > 0$ son los grados de libertad. Así, la función de densidad es

$$p(\varepsilon) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\varepsilon^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Luego, la función de verosimilitud para todas las observaciones:

$$p(y \mid \beta, \nu) = \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\right]^n \prod_{i=1}^{n} \left(1 + \frac{(y_i - x_i^\top \beta)^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

entonces, el estimador de máxima verosimilitud

$$\hat{\beta}_{\text{MLE}} = \arg\min_{\beta} \sum_{i=1}^{n} \log\left(1 + \frac{(y_i - x_i^\top \beta)^2}{\nu}\right)$$

Este modelo permite trabajar con datos que siguen una distribución con cola pesada lo cual permite una mayor probabilidad de valores extremos.

Problem 8: 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.).

Solución.

Al utilizar modelos de regresión con errores no normales como los propuestos anteriormente (Laplace o Student-t), es inevitable afectar a la metodología inherente del modelo. A continuación, se discutirán las consecuencias metodológicas de considerar errores no normal.

Consecuencias sobre la forma de la verosimilitud

Regresión con errores Laplace. La verosimilitud toma la forma

$$\mathcal{L}(\beta \mid y, X) = \left(\frac{1}{2b}\right)^n \exp\left(-\frac{1}{b}\sum_{i=1}^{n} |y_i - x_i^\top \beta|\right)$$

- La función de verosimilitud no es diferenciable en todos los puntos debido al valor absoluto, esto puede afectar en la aplicabilidad en algún problema, pues muchas veces se supone que la verosimilitud es suave.
- El problema de optimización se convierte en un programa lineal, el cual se puede solucionar con métodos de optimización no suaves, esto puede provocar un costo computacional muy alto, lo cual afecta la implementación, sobre todo en altas dimensiones.

Regresión con errores Student-t. La verosimilitud es

$$\mathcal{L}(\beta, \nu \mid y, X) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(y_i - x_i^\top \beta)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Consecuencias:

- La verosimilitud es diferenciable pero más compleja que la normal.
- Aparece un parámetro adicional ν (grados de libertad) que debe estimarse, esto afecta en la aplicabilidad pues en caso de no tener suficiente información, entonces no será posible estimarlo adecuadamente.

Consecuencias sobre la existencia de priors conjugados

Regresión con errores Laplace. En este caso, no existe prior conjugado para β con verosimilitud Laplace esto provoca que la distribución posterior no tenga forma analítica cerrada. Además, la posterior debe aproximarse numéricamente.

Regresión con errores Student-t. En este caso, no existe familia conjugada cuando tanto β como ν son desconocidos. Si ν es fijo, se pueden encontrar priors conjugados aproximados bajo algunas condiciones condiciones, esto le va una ventaja en aplicabilidad. Para ν desconocido, la posterior conjunta $p(\beta, \nu \mid y)$ no es analítica.

Consecuencias sobre los métodos de inferencia

Regresión con errores Laplace. Dado la discusión de los puntos anteriores, para este método se recomienda lo siguiente.

- **Optimización**: Programación lineal, métodos de punto interior, o algoritmos para regresión cuantil.
- **Inferencia Bayesiana**: MCMC (Gibbs sampling con variables auxiliares, Metropolis-Hastings) o aproximación variacional.
- Los intervalos de confianza frecuentistas, sin embargo, requieren bootstrap o métodos asintóticos especiales.

Regresión con errores Student-t. Dado la discusión de los puntos anteriores, para este método se recomienda lo siguiente.

- **Optimización**: Algoritmos EM tratando la representación de mezcla de normales, o métodos de gradiente.
- **Inferencia Bayesiana**: MCMC (Gibbs sampling usando la representación escala-mezcla de normales) $\varepsilon_i \mid \omega_i \sim N(0, \omega_i), \quad \omega_i \sim$ Gamma-Inversa.
- **Aproximación variacional**: Familias variacionales que capturen la dependencia entre β y los parámetros de escala.

En general, en ambos casos, aumenta significativamente respecto al caso normal; los métodos MCMC y variacionales requieren más iteraciones y verificación de convergencia, se necesitan librerías especializadas (Stan, PyMC3, JAGS) en lugar de fórmulas cerradas; por lo que se gana robustez y flexibilidad, pero se pierde cierta capacidad analítica y eficiencia computacional.