Tarea 3
Ciencia de Datos.

Alfredo Bistrain
Jesús Salazar

October 14, 2025

Ejercicio

1 Regresión lineal ordinaria (OLS)

1. Derivación del estimador OLS: Partiendo del modelo clásico

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$$

demuestre que el estimador de Mínimos Cuadrados Ordinarios es

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

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

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:
Considere el modelo de regresión lineal,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2 I_n\right)$. Se busca al valor de $\hat{\boldsymbol{\beta}}$ que minimice la suma de los errores al cuadrado, es decir,

$$\begin{aligned} SS_{\text{Res}}\left(\boldsymbol{\beta}\right) &= \sum_{i=1}^{n} \left(y_i - [X\boldsymbol{\beta}]_i\right)^2 \\ &= (y - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top X^\top\boldsymbol{y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta} \\ &= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top X\boldsymbol{\beta} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}. \end{aligned}$$

Derivando la expresión anterior con respecto a β y notando que $\mathrm{t}^\top X$ es una matriz simétrica, se tiene,

$$\begin{aligned}\frac{\partial SS_{\text{Res}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \left(-2\boldsymbol{y}^\top X\right)^\top + 2X^\top X\boldsymbol{\beta} \\ &= -2X^\top \boldsymbol{y} + 2X^\top X\boldsymbol{\beta}.\end{aligned}$$

Igualando a cero la expresión anterior, se puede observar que el estimador que se busca $\overline{\boldsymbol{\beta}}$ debe satisfacer

$$X^\top X\overline{\boldsymbol{\beta}} = X^\top \boldsymbol{y}.$$

Este sistema de ecuaciones son la ecuaciones normales para le modelo de regresión linea. Para poder despejar $\overline{\boldsymbol{\beta}}$ es necesario que $X^\top X$ sea invertible. Esto es que tenga rango completo. En este caso,

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

Sabemos que $\hat{\boldsymbol{\beta}}$ minimiza $SS_{\text{Res}}\left(\boldsymbol{\beta}\right)$ si y solo si la matriz Hessiana de SS_{Res} es positiva definida.

$$\begin{aligned}H &= \frac{\partial^2 SS_{\text{Res}}}{\partial \boldsymbol{\beta}^2} \\ &= \left(2X^\top X\right)^\top \\ &= 2X^\top X.\end{aligned}$$

que es positiva definida si $\text{rango}(X) = p$.

(2)

Para solucionar este inciso, bastan observar que Como $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 I_n\right)$, entonces $\boldsymbol{Y} \sim N\left(X\boldsymbol{\beta}, \sigma^2 I_n\right)$, esto por propiedades de la distribución normal multivariada, por otro lado, observe que,

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\left(\left(X^\top X\right)^{-1} X^\top \boldsymbol{Y}\right) = \left(X^\top X\right)^{-1} X^\top \mathbb{E}(\boldsymbol{Y}) = \left(X^\top X\right)^{-1} X^\top X\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Del mismo modo, la varianza de $\widehat{\boldsymbol{\beta}}$ se puede obtener como,

$$\begin{aligned}\text{Var}(\widehat{\boldsymbol{\beta}}) &= \text{Var}\left[\left(X^\top X\right)^{-1} X^\top \boldsymbol{y}\right] \\ &= \left(X^\top X\right)^{-1} X^\top \,\text{Var}(\boldsymbol{y}) \left[\left(X^\top X\right)^{-1} X^\top\right]^\top \\ &= \left(X^\top X\right)^{-1} X^\top X \left(X^\top X\right)^{-1} \sigma^2 = \left(X^\top X\right)^{-1} \sigma^2.\end{aligned}$$

De este modo,

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left(X^\top X\right)^{-1} \sigma^2\right),$$

y por el teorema de Gauss-Markov, se tiene que $\hat{\beta}$ es un estimador insesgado y eficiente.

...

Ejercicio

2 Regresión lineal bayesiana (prior conjugado)

1. Prior conjugado: Suponga un prior conjugado:

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

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

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

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

Solución:
(1)

Considere el modelo de regresión lineal,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2 I_n\right)$. Además, suponga que los parámetros son variables aletorias con distribuciones a priori,

$$\begin{aligned} \beta \mid \sigma^2 &\sim \mathcal{N}(\beta_0, \sigma^2 V_0) \\ \sigma^2 &\sim \text{Inv-Gamma}(a_0, b_0), \end{aligned}$$

donde V_0, β_0, a_0, b_0 son hiperparametros de las distribuciones. De este modo, la función de verosimilitud está dada por,

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

Observe que la distribución a priori conjunta de los parámetros se puede obtener como,

$$\begin{aligned} p(\beta, \sigma^2) &= p(\beta \mid \sigma^2) p(\sigma^2) \\ &= (2\pi\sigma^2)^{-k/2} |V_0|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \beta_0)^\top V_0^{-1}(\beta - \beta_0)\right\} \\ &\quad \times \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right). \end{aligned}$$

Por otro lado, se sabe que la distribución posterior es proporcional a la expresión,

$$\begin{aligned}
p(\beta, \sigma^2 \mid y) &\propto p(y \mid \beta, \sigma^2) p(\beta \mid \sigma^2) p(\sigma^2) \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \right\} \\
&\times (\sigma^2)^{-k/2} \exp\left\{ -\frac{1}{2\sigma^2} (\beta - \beta_0)^\top V_0^{-1} (\beta - \beta_0) \right\} \\
&\times (\sigma^2)^{-a_0 - 1} \exp\left(-\frac{b_0}{\sigma^2} \right).
\end{aligned}$$

Note que si se juntan las dos primeras exponenciales y se factoriza el termino $-\frac{1}{2\sigma^2}$, los términos dentro de la exponencial se pueden reescribir como,

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

Para simplificar la notación, defina,

$$\begin{aligned}
A &= X^\top X + V_0^{-1} \\
b &= X^\top y + V_0^{-1} \beta_0 \\
c &= y^\top y + \beta_0^\top V_0^{-1} \beta_0
\end{aligned}$$

Observe que la última expresión se puede reescribir como,

$$\begin{aligned}
&\beta^\top A \beta - 2\beta^\top b + c \\
&= \beta^\top A \beta - 2\beta^\top b + b^\top A^{-1} b - b^\top A^{-1} b + c \\
&= (\beta - A^{-1} b)^\top A (\beta - A^{-1} b) - b^\top A^{-1} b + c \\
&= (\beta - A^{-1} b)^\top A (\beta - A^{-1} b) + (c - b^\top A^{-1} b).
\end{aligned}$$

Regresando esta expresión a ecuación original se tiene,

$$\begin{aligned}
p(\beta, \sigma^2 \mid y) &\propto (\sigma^2)^{-(n+k)/2 - a_0 - 1} \exp\left\{ -\frac{1}{2\sigma^2} \left[(\beta - A^{-1} b)^\top A (\beta - A^{-1} b) + (c - b^\top A^{-1} b) \right] - \frac{b_0}{\sigma^2} \right\} \\
&= (\sigma^2)^{-k/2} \exp\left\{ -\frac{1}{2\sigma^2} (\beta - A^{-1} b)^\top A (\beta - A^{-1} b) \right\} \\
&\times (\sigma^2)^{-n/2 - a_0 - 1} \exp\left\{ -\frac{1}{\sigma^2} \left[b_0 + \frac{1}{2} (c - b^\top A^{-1} b) \right] \right\}.
\end{aligned}$$

Así, multiplicando con las constantes adecuadas para completar las distribuciones correspondientes se tiene,

$$\begin{aligned}
p(\beta, \sigma^2 \mid y) &\propto \frac{1}{(2\pi\sigma^2)^{k/2} |A^{-1}|^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\beta - A^{-1} b)^\top A (\beta - A^{-1} b) \right\} \\
&\times \frac{\left[b_0 + \frac{1}{2} (c - b^\top A^{-1} b) \right]^{a_0 + n/2}}{\Gamma(a_0 + n/2)} (\sigma^2)^{-(a_0 + n/2) - 1} \exp\left\{ -\frac{1}{\sigma^2} \left[b_0 + \frac{1}{2} (c - b^\top A^{-1} b) \right] \right\}.
\end{aligned}$$

...

De esta manera, si se define a

$$
\begin{aligned}
V_n &= A^{-1} = (X^\top X + V_0^{-1})^{-1} \\
\beta_n &= A^{-1}b = V_n(X^\top y + V_0^{-1}\beta_0) \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}\left[c - \beta_n^\top V_n^{-1}\beta_n\right] \\
&= b_0 + \frac{1}{2}\left[y^\top y + \beta_0^\top V_0^{-1}\beta_0 - \beta_n^\top V_n^{-1}\beta_n\right],
\end{aligned}
$$

se tiene que la distribución a posterior es una distribución normal-gamma inversa con parámetros, $(\beta_n, \sigma^2 V_n, a_n, b_n)$

$$
\begin{aligned}
p(\beta, \sigma^2 \mid y) = &\frac{1}{(2\pi\sigma^2)^{k/2}|V_n|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \beta_n)^\top V_n^{-1}(\beta - \beta_n)\right\} \\
&\times \frac{b_n^{a_n}}{\Gamma(a_n)}(\sigma^2)^{-a_n-1}\exp\left(-\frac{b_n}{\sigma^2}\right).
\end{aligned}
$$

Por último, observe que la distribución a posterior marginal de parámetro σ^2 se puede encontrar a partir de la definición,

$$
\begin{aligned}
p(\sigma^2 \mid y) &= \int p(\beta, \sigma^2 \mid y)\, d\beta \\
&= \int \frac{1}{(2\pi\sigma^2)^{k/2}|V_n|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \beta_n)^\top V_n^{-1}(\beta - \beta_n)\right\} \\
&\quad \times \frac{b_n^{a_n}}{\Gamma(a_n)}(\sigma^2)^{-a_n-1}\exp\left(-\frac{b_n}{\sigma^2}\right) d\beta \\
&= \frac{b_n^{a_n}}{\Gamma(a_n)}(\sigma^2)^{-a_n-1}\exp\left(-\frac{b_n}{\sigma^2}\right) \\
&\quad \times \int \frac{1}{(2\pi\sigma^2)^{k/2}|V_n|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \beta_n)^\top V_n^{-1}(\beta - \beta_n)\right\} d\beta \\
&= \frac{b_n^{a_n}}{\Gamma(a_n)}(\sigma^2)^{-a_n-1}\exp\left(-\frac{b_n}{\sigma^2}\right) \times 1 \\
&= \frac{b_n^{a_n}}{\Gamma(a_n)}(\sigma^2)^{-a_n-1}\exp\left(-\frac{b_n}{\sigma^2}\right)
\end{aligned}
$$

La expresión anterior corresponde a la función de densidad de una distribución gamma-inversa con parámetros (a_n, b_n). Por lo tanto

$$
\sigma^2 | y \sim \text{Inv-Gamma}(a_n, b_n).
$$

Por otro lado, la distribución posterior marginal de β se puede encontrar a partir de la definición, esto es,

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

donde $m = a_n + k/2$ y $\lambda = b_n + \frac{1}{2}(\beta-\beta_n)^\top V_n^{-1}(\beta-\beta_n)$,

$$
\begin{aligned}
&= \frac{b_n^{a_n}}{(2\pi)^{k/2}|V_n|^{1/2}\Gamma(a_n)} \int_0^\infty (\sigma^2)^{-(m+1)} \exp\left\{-\frac{\lambda}{\sigma^2}\right\} d\sigma^2 \\
&= \frac{b_n^{a_n}}{(2\pi)^{k/2}|V_n|^{1/2}\Gamma(a_n)} \times \frac{\Gamma(m)}{\lambda^m}
\end{aligned}
$$

usando la función de densidad gamma, $\int_0^\infty x^{-(\alpha+1)}e^{-\beta/x}dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$ con $\alpha = m, \beta = \lambda$,

$$
\begin{aligned}
&= \frac{b_n^{a_n}}{(2\pi)^{k/2}|V_n|^{1/2}\Gamma(a_n)} \times \frac{\Gamma(a_n+k/2)}{\left[b_n + \frac{1}{2}(\beta-\beta_n)^\top V_n^{-1}(\beta-\beta_n)\right]^{a_n+k/2}} \\
&= \frac{\Gamma(a_n+k/2)}{\Gamma(a_n)} \frac{b_n^{a_n}}{(2\pi)^{k/2}|V_n|^{1/2}} \left[b_n + \frac{1}{2}(\beta-\beta_n)^\top V_n^{-1}(\beta-\beta_n)\right]^{-(a_n+k/2)} \\
&= \frac{\Gamma\left(a_n+k/2\right)}{\Gamma\left(a_n\right)} \frac{1}{(2\pi)^{k/2}\left|V_n\right|^{1/2} b_n^{-a_n}} \left[b_n\left(1 + \frac{1}{2b_n}\left(\beta-\beta_n\right)^\top V_n^{-1}\left(\beta-\beta_n\right)\right)\right]^{-(a_n+k/2)} \\
&= \frac{\Gamma\left(a_n+k/2\right)}{\Gamma\left(a_n\right)} \frac{1}{(2\pi)^{k/2}\left|V_n\right|^{1/2}} b_n^{-k/2} \left[1 + \frac{1}{2b_n}\left(\beta-\beta_n\right)^\top V_n^{-1}\left(\beta-\beta_n\right)\right]^{-(a_n+k/2)} \\
&= \frac{\Gamma\left(\frac{2a_n+k}{2}\right)}{\Gamma\left(\frac{2a_n}{2}\right)(2a_n\pi)^{k/2}\left|\frac{b_n}{a_n}V_n\right|^{1/2}} \left[1 + \frac{1}{2a_n}\left(\beta-\beta_n\right)^\top \left(\frac{b_n}{a_n}V_n\right)^{-1}\left(\beta-\beta_n\right)\right]^{-(2a_n+k)/2}
\end{aligned}
$$

La expresión anterior corresponde a la función de densidad de una distribución t de Student multivariada con $2a_n$ grados de libertad, localización β_n y matriz de escala $\frac{a_n}{b_n}V_n$.

...

Ejercicio

3 Conexión con regularización

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

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

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

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

Solución:
Sean $p(\boldsymbol{Y} \mid \boldsymbol{\beta})$ la función de densidad de probabilidad de $\boldsymbol{Y}$ dada $\boldsymbol{\beta}$ la cual es una normal multivariada con media $X\boldsymbol{\beta}$ y varianza $\sigma^2 I_n$, y $p(\boldsymbol{\beta})$ la función de densidad de probabilidad de $\boldsymbol{\beta}$ la cual es una normal multivariada con media $\mathbf{0}$ y varianza $\tau^2 I_p$. Observe que si la función de densidad de los datos se le quitan las constantes, se tiene que esta es proporcional a,

$$\begin{aligned}
p(\boldsymbol{Y} \mid \boldsymbol{\beta}) &\propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{Y} - X\boldsymbol{\beta}\|^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^\top(\boldsymbol{Y} - X\boldsymbol{\beta})\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}^\top\boldsymbol{Y} - 2\boldsymbol{\beta}^\top X^\top\boldsymbol{Y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}\right)\right)
\end{aligned}$$

Del mismo modo, se tiene la distribución a priori es proporcional a,

$$p(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right).$$

De lo anterior se sigue la distribución aposterior es proporcional a,

$$\begin{aligned}
p(\boldsymbol{\beta} \mid \boldsymbol{Y}) &\propto p(\boldsymbol{Y} \mid \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}^\top\boldsymbol{Y} - 2\boldsymbol{\beta}^\top X^\top\boldsymbol{Y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}\right) - \frac{1}{2\tau^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{Y}^\top\boldsymbol{Y} + \frac{1}{\sigma^2}\boldsymbol{\beta}^\top X^\top\boldsymbol{Y} - \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta} - \frac{1}{2\tau^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right) \\
&\propto \exp\left(\boldsymbol{\beta}^\top\left(\frac{X^\top\boldsymbol{Y}}{\sigma^2}\right) - \frac{1}{2}\boldsymbol{\beta}^\top\left(\frac{X^\top X}{\sigma^2} + \frac{I_p}{\tau^2}\right)\boldsymbol{\beta}\right)
\end{aligned}$$

Si se define a

$$\begin{aligned}
\Sigma^{-1} &= \frac{X^\top X}{\sigma^2} + \frac{I_p}{\tau^2} \\
\boldsymbol{\mu} &= \Sigma\left(\frac{X^\top\boldsymbol{Y}}{\sigma^2}\right)
\end{aligned}$$

se tiene que la expresión anterior se puede reescribir como,

$$\begin{aligned}
p(\boldsymbol{\beta} \mid \boldsymbol{Y}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\top\Sigma^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\Sigma^{-1}\boldsymbol{\mu}\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})\right)
\end{aligned}$$

...

Observe que la expresión anterior corresponde a el kernel de una distribución normal multivariada con media $\boldsymbol{\mu}$ y varianza Σ. De aquí, observe que la media explicita de la distribución a posterior de $\boldsymbol{\beta}$ está dada por,

$$\begin{aligned}\boldsymbol{\mu} &= \left(\frac{X^\top X}{\sigma^2} + \frac{I_p}{\tau^2}\right)^{-1} \frac{X^\top \boldsymbol{Y}}{\sigma^2} \\ &= \left(X^\top X + \frac{\sigma^2}{\tau^2} I_p\right)^{-1} X^\top \boldsymbol{Y},\end{aligned}$$

que coincide con el estimador Ridge $\hat{\beta}$. Observe que en este caso, la relación entre α, σ^2 y τ^2 es

$$\alpha = \frac{\sigma^2}{\tau^2}$$

la penalización dependerá de la discrepancia que hay entre la varianza a priori y la varianza de los datos.

...

Ejercicio

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

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

el estimador MAP corresponde a la regresión Lasso:

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

Solución:

En este caso, considere el modelo regresión lineal dado por, $y = X\beta + \epsilon$ con errores $\epsilon \sim N(0, \sigma^2 I)$, la verosimilitud es,

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

Además, considere una distribución apriori para β dada por,

$$p(\beta|\lambda) = \prod_{j=1}^{p} \frac{\lambda}{2}\exp(-\lambda|\beta_j|) \propto \exp\left(-\lambda\|\beta\|_1\right).$$

Una vez que se tiene definida la función de verosimilitud y la distribución a priori, se puede obtener la distribución a posterior dada por,

$$\begin{aligned}
p(\beta \mid y) &\propto p(y \mid \beta)\cdot p(\beta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right)\cdot\prod_{j=1}^{p}\frac{\lambda}{2}\exp\left(-\lambda|\beta_j|\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right)\cdot\exp\left(-\lambda\sum_{j=1}^{p}|\beta_j|\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \lambda\sum_{j=1}^{p}|\beta_j|\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \lambda\|\beta\|_1\right)
\end{aligned}$$

$$p(\beta|y) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \lambda\sum_{j=1}^{p}|\beta_j|\right).$$

Por último, observe que el estimador MAP (Maximum A Posteriori) se obtiene maximizando la distribución posterior,

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

además, note que maximizar $p(\beta|y)$ es equivalente a minimizar el negativo del log-posterior,

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

...

Multiplicando por $2\sigma^2$ y definiendo $\lambda' = 2\sigma^2\lambda$,

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

que es exactamente la formulación de la regresión Lasso.

...

Ejercicio

4 Extensiones: errores no normales

1. Modelos alternativos: Proponga un modelo de regresión donde el error ε no siga una distribución Normal. Ejemplos:
- $\varepsilon \sim \text{Laplace}(0, b)$ (robusto a outliers).
- $\varepsilon \sim$ 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.).

Solución:
Para realizar este análisis, considere un modelo de regresión lineal dado por, dado por, $y = X\beta + \varepsilon$ con errores $\varepsilon \sim$ Student- $t(\nu)$. En este caso la verosimilitud toma la forma,

$$
\begin{aligned}
p(y|X, \beta, \sigma^2, \nu) &= \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(y_i - x_i^\top \beta)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \\
&= \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\right]^n \prod_{i=1}^{n} \left(1 + \frac{(y_i - x_i^\top \beta)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}
\end{aligned}
$$

Donde,

- $\nu > 0$ son los grados de libertad
- σ^2 es el parámetro de escala
- $\Gamma(\cdot)$ es la función Gamma
- $x_i^\top$ es la fila i-ésima de la matriz de diseño X.

Observe que esta distribución tiene colas más pesadas, lo que lleva a que residuos de mayor tamaño tengan mayor probabilidad, esto produce que los parámetros no se vean tan afectados por outliers y se tengan estimaciones más robustas.

Se sabe que para una verosimilitud de este tipo, no hay distribuciones a priori tales que se pueda tener un modelo conjugado con esta verosimilitud. Sin embargo, se puede recurrir a modelos semi conjugados, los cuales pueden facilitar las cuentas y la implementación de algoritmos de muestreo. Al no poder llegar a una expresión cerrada de la distribució a posterior, se debe recurrir a otros metodos de inferencia. Usualmente se utiliza MCMC, como Gibbs sampling o Metropolis Hasting. Particularmente, si se plantea,

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

, entonces se puede obtener la distribución t de Student con ν grados de libertad y escala σ^2, marginalizando sobre λ_i. A partir de esto, condicionando a $\boldsymbol{\lambda}, \boldsymbol{\beta}$ sigue una distribución normal.

...

En este caso en particular, se puede implemetentar un muestreo de Gibbs, muestreando a β a partir de la normal condicional, σ^2 a partir de una gamma inversa (como en el caso usual) y a λ_i a partir de la gamma propuesta, ν se puede muestrear a partir de un algoritmo Metrópolis Hasting.

...