Notas de Clase :

Análisis de Regresión

Felipe Osorio

Departamento de Matemática, Universidad Técnica Federico Santa María

Índice general

Prefacio

Estas notas de clase están asociadas a los contenidos de la asignatura *MAT-266: Análisis de Regresión*, dictada en el programa de Ingeniería Civil Matemática de la Universidad Técnica Federico Santa María. Aunque el documento se encuentra en una etapa bastante preliminar, espero ir puliendo el mismo para que se pueda convertir en un apunte que sirva de apoyo para los estudiantes de la asignatura.

Las notas se encuentran divididas en tres partes, con preliminares conteniendo resultados de la distribución normal y su conexión con formas cuadráticas. Luego, se presenta la inferencia en el modelo de regresión lineal y posteriormente hay una serie de resultados que permiten la crítica del proceso de modelación así como procedimientos alternativos en caso de que algunos de los supuestos básicos no sean satisfechos. El objetivo de este texto es proveer de una introducción rigurosa al tópico de regresión presentando también aplicaciones prácticas así como destacar los elementos necesarios para la implementación computacional de tales técnicas.

Agradezco al profesor Manuel Galea por haberme introducido en este tópico así como por su constante apoyo durante toda mi carrera académica. Adicionalmente, debo destacar el apoyo de los estudiantes que han participado de alguna versión de este curso, pues producto de sus comentarios y sugerencias este documento se ha visto notablemente mejorado.

Felipe Osorio
Valparaíso, Agosto 2021.

Capítulo 1

Preliminares

1.1. Vectores Aleatorios

El propósito de esta sección es introducir algunas propiedades elementales de vectores aleatorios útiles a lo largo de este curso. Se asume que el lector es familiar con el concepto de variable aleatoria unidimensional.

Un vector aleatorio n-dimensional $\boldsymbol{X}$ es una función (medible) desde el espacio de probabilidad Ω a $\mathbb{R}^n$, esto es

$$\boldsymbol{X} : \Omega \to \mathbb{R}^n.$$

Por convención asumiremos que el vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ es un vector columna.

Definición 1.1 (Función de distribución). Para $\boldsymbol{X}$ distribuído en $\mathbb{R}^n$, la *función de distribución* de $\boldsymbol{X}$ es una función $F : \mathbb{R}^n \to [0, 1]$, tal que

$$F(\boldsymbol{x}) = \mathsf{P}(\boldsymbol{X} \leq \boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^n \tag{1.1}$$

y denotamos $\boldsymbol{X} \sim F$ o $\boldsymbol{X} \sim F_X$.

La función en (1.1) debe ser entendida como

$$F(\boldsymbol{x}) = \mathsf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n),$$

que corresponde a la probabilidad del evento $\bigcap_{k=1}^{n} \{X_k \leq x_k\}$.

Propiedad 1.2. La función de distribución acumulada tiene las siguientes propiedades:

(a) $F(\boldsymbol{x})$ es función monótona creciente y contínua a la derecha en cada uno de los componentes de $\boldsymbol{X}$,
(b) $0 \leq F(\boldsymbol{x}) \leq 1$,
(c) $F(-\infty, x_2, \ldots, x_n) = \cdots = F(x_1, \ldots, x_{n-1}, -\infty) = 0$,
(d) $F(+\infty, \ldots, +\infty) = 1$.

Sea F la función de distribución del vector aleatorio $\boldsymbol{X}$. Entonces, existe una función no-negativa f tal que

$$F(\boldsymbol{x}) = \int_{-\infty}^{x} f(\boldsymbol{u}) \, \mathsf{d}\boldsymbol{u}, \qquad \boldsymbol{x} \in \mathbb{R}^n,$$

en este caso decimos que $\boldsymbol{X}$ es un vector aleatorio contínuo con *función de densidad* f. Por el teorema fundamental del Cálculo, tenemos que

$$f(\boldsymbol{x}) = \frac{\partial^n F(\boldsymbol{x})}{\partial x_1 \cdots \partial x_n}.$$

Además, considere $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, para $\boldsymbol{x}, \boldsymbol{y}$ vectores en $\bar{\mathbb{R}}^n$, entonces

$$\boldsymbol{x} \leq \boldsymbol{y} \quad \text{esto es,} \quad x_i \leq y_i, \quad \text{para } i = 1, \ldots, n.$$

Esto permite definir un rectángulo n-dimensional en $\mathbb{R}^n$ como

$$I = (\boldsymbol{a}, \boldsymbol{b}] = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a} < \boldsymbol{x} \leq \boldsymbol{b}\}$$

para todo $\boldsymbol{a}, \boldsymbol{b} \in \bar{\mathbb{R}}^n$. Entonces, también por el teorema fundamental del Cálculo, tenemos que si

$$f(\boldsymbol{x}) = \frac{\partial^n F(\boldsymbol{x})}{\partial x_1 \cdots \partial x_n}.$$

existe y es continua (casi en toda parte) sobre un rectángulo I, entonces

$$\mathsf{P}(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{x})\, \mathsf{d}\boldsymbol{x}, \quad \forall A \subset I.$$

Naturalmente la función de densidad debe satisfacer

$$\int_{\mathbb{R}^n} f(\boldsymbol{x})\, \mathsf{d}\boldsymbol{x} = 1.$$

Considere el vector aleatorio n-dimensional $\boldsymbol{X}$ particionado como $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$ donde $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$ son vectores $n_1 \times 1$ y $n_2 \times 1$, respectivamente, con $n = n_1 + n_2$. Tenemos que $\boldsymbol{X}_i \sim F_i$, $i = 1, 2$, de este modo $\boldsymbol{X}$ se denomina la *conjunta* de $\boldsymbol{X}_1, \boldsymbol{X}_2$ mientras que los $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$ son llamados *marginales* de $\boldsymbol{X}$.

Note que, las funciones de distribución marginal pueden ser recuperadas desde la distribución conjunta mediante

$$F_1(\boldsymbol{s}) = F(\boldsymbol{s}, +\boldsymbol{\infty}), \qquad F_2(\boldsymbol{t}) = F(+\boldsymbol{\infty}, \boldsymbol{t}), \qquad \forall \boldsymbol{s} \in \mathbb{R}^{n_1}, \boldsymbol{t} \in \mathbb{R}^{n_2}.$$

Cuando $\boldsymbol{X}$ es absolutamente contínua con función de densidad $f(\boldsymbol{x}) = f(\boldsymbol{x}_1, \boldsymbol{x}_2)$, entonces la función de densidad de $\boldsymbol{X}_i$ también es absolutamente contínua y puede ser obtenida como

$$f_1(\boldsymbol{s}) = \int_{\mathbb{R}^{n_2}} f(\boldsymbol{s}, \boldsymbol{u})\, \mathsf{d}\boldsymbol{u}, \quad f_2(\boldsymbol{t}) = \int_{\mathbb{R}^{n_1}} f(\boldsymbol{u}, \boldsymbol{t})\, \mathsf{d}\boldsymbol{u}, \quad \forall \boldsymbol{s} \in \mathbb{R}^{n_1}, \boldsymbol{t} \in \mathbb{R}^{n_2},$$

el resultado anterior es análogo para el caso de distribuciones discretas. Si $\boldsymbol{X}$ es absolutamente contínuo y $f_1(\boldsymbol{x}_1) > 0$, entonces la *densidad condicional* de $\boldsymbol{X}_2$ dado $\boldsymbol{X}_1 = \boldsymbol{x}_1$ es

$$f_{X_2|X_1=x_1}(\boldsymbol{u}) = \frac{f_X(\boldsymbol{x}_1, \boldsymbol{u})}{f_1(\boldsymbol{x}_1)},$$

con función de distribución de $\boldsymbol{X}_2$ condicional a $\boldsymbol{X}_1 = \boldsymbol{x}_1$ dada por

$$F_{X_2|X_1=x_1}(\boldsymbol{u}) = \int_{-\infty}^{u} f_{X_2|X_1=x_1}(\boldsymbol{t})\, \mathsf{d}\boldsymbol{t},$$

tenemos además que

$$f_{X_2|X_1=x_1}(\boldsymbol{u}) = \frac{f_X(\boldsymbol{x}_1, \boldsymbol{u})}{\int_{\mathbb{R}^{n_2}} f_X(\boldsymbol{x}_1, \boldsymbol{t})\, \mathsf{d}\boldsymbol{t}}.$$

1.2. Operadores de esperanza y covarianza

Considere $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ vector aleatorio n-dimensional con función de densidad f. Entonces la esperanza de cualquier función $\boldsymbol{g}$ de $\boldsymbol{X}$ está dada por

$$\mathsf{E}(\boldsymbol{g}(\boldsymbol{X})) = \int_{\mathbb{R}^n} \boldsymbol{g}(\boldsymbol{t}) f(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t},$$

siempre que la integral (n-dimensional) exista.

Más generalmente, sea $\boldsymbol{Z} = (Z_{ij})$ una función matricial $m \times n$, entonces podemos definir el operador de esperanza de una matriz aleatoria como

$$\mathsf{E}(\boldsymbol{Z}(\boldsymbol{X})) = (\mathsf{E}(Z_{ij})), \qquad Z_{ij} = Z_{ij}(\boldsymbol{X}). \tag{1.2}$$

De la definición en (1.2) se desprenden una serie de resultados útiles con relación al operador de esperanza. Por ejemplo, sea $\boldsymbol{A} = (a_{ij})$ una matriz de constantes, entonces

$$\mathsf{E}(\boldsymbol{A}) = \boldsymbol{A}.$$

Resultado 1.3. *Sea $\boldsymbol{A} = (a_{ij})$, $\boldsymbol{B} = (b_{ij})$ y $\boldsymbol{C} = (c_{ij})$ matrices de constantes $l \times m$, $n \times p$ y $l \times p$, respectivamente. Entonces*

$$\mathsf{E}(\boldsymbol{AZB} + \boldsymbol{C}) = \boldsymbol{A}\,\mathsf{E}(\boldsymbol{Z})\boldsymbol{B} + \boldsymbol{C}.$$

Demostración. Sea $\boldsymbol{Y} = \boldsymbol{AZB} + \boldsymbol{C}$, entonces

$$Y_{ij} = \sum_{r=1}^{m} \sum_{s=1}^{n} a_{ir} Z_{rs} b_{sj} + c_{ij},$$

de este modo

$$\begin{aligned} \mathsf{E}(\boldsymbol{AZB} + \boldsymbol{C}) = (\mathsf{E}(Y_{ij})) &= \left(\sum_{r=1}^{m} \sum_{s=1}^{n} a_{ir}\, \mathsf{E}(Z_{rs}) b_{sj} + c_{ij} \right) \\ &= \boldsymbol{A}\,\mathsf{E}(\boldsymbol{Z})\boldsymbol{B} + \boldsymbol{C}. \end{aligned}$$

□

Un caso particular importante corresponde a la esperanza de una transformación lineal. Considere el vector aleatorio n-dimensional, $\boldsymbol{Y} = \boldsymbol{AX}$, donde $\boldsymbol{X}$ es vector aleatorio $m \times 1$, entonces $\mathsf{E}(\boldsymbol{AX}) = \boldsymbol{A}\,\mathsf{E}(\boldsymbol{X})$. Esta propiedad puede ser extendida para sumas de vectores aleatorios, como

$$\mathsf{E}\left(\sum_i \boldsymbol{A}_i \boldsymbol{X}_i \right) = \sum_i \boldsymbol{A}_i\, \mathsf{E}(\boldsymbol{X}_i),$$

de manera similar tenemos que

$$\mathsf{E}\left(\sum_i \alpha_i \boldsymbol{Z}_i \right) = \sum_i \alpha_i\, \mathsf{E}(\boldsymbol{Z}_i),$$

donde α_i son constantes y los $\boldsymbol{Z}_i$ son matrices aleatorias.

Definición 1.4 (Matriz de covarianza). Sean $\boldsymbol{X}$ e $\boldsymbol{Y}$ vectores aleatorios m y n-dimensionales, respectivamente. Se define la *matriz de covarianza* entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ como la matriz $m \times n$,

$$\mathsf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = (\mathsf{Cov}(X_i, Y_j)).$$

Podemos apreciar, a partir de la definición de covarianza que

$$\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\}.$$

En efecto, sean $\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{X})$ y $\boldsymbol{\eta} = \mathsf{E}(\boldsymbol{Y})$. Entonces,

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y}) &= (\mathsf{Cov}(X_i,Y_j)) = (\mathsf{E}(X_i-\mu_i)(Y_j-\eta_j)) \\ &= \mathsf{E}([(X_i-\mu_i)(Y_j-\eta_j)]) = \mathsf{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\eta})^\top].\end{aligned}$$

Tenemos ademas el siguiente resultado

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y}) &= \mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\} \\ &= \mathsf{E}(\boldsymbol{X}\boldsymbol{Y}^\top - \mathsf{E}(\boldsymbol{X})\boldsymbol{Y}^\top - \boldsymbol{X}\,\mathsf{E}^\top(\boldsymbol{Y}) + \mathsf{E}(\boldsymbol{X})\,\mathsf{E}^\top(\boldsymbol{Y})) \\ &= \mathsf{E}(\boldsymbol{X}\boldsymbol{Y}^\top) - \mathsf{E}(\boldsymbol{X})\,\mathsf{E}^\top(\boldsymbol{Y}).\end{aligned}$$

Se define la *matriz de dispersión (varianza)*, como $\mathsf{Cov}(\boldsymbol{X}) = \mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})$. De este modo, tenemos

$$\mathsf{Cov}(\boldsymbol{X}) = (\mathsf{Cov}(X_i,X_j)) = \mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))^\top\},$$

y, de la misma manera que para el caso de la matriz de covarianza,

$$\mathsf{Cov}(\boldsymbol{X}) = \mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top) - \mathsf{E}(\boldsymbol{X})\,\mathsf{E}^\top(\boldsymbol{X}).$$

EJEMPLO 1.5. Sea $\boldsymbol{a}$ vector de constantes $n \times 1$, entonces

$$\mathsf{Cov}(\boldsymbol{X}-\boldsymbol{a}) = \mathsf{Cov}(\boldsymbol{X}).$$

En efecto, note que

$$\boldsymbol{X}-\boldsymbol{a}-\mathsf{E}(\boldsymbol{X}-\boldsymbol{a}) = \boldsymbol{X}-\mathsf{E}(\boldsymbol{X}),$$

por tanto, tenemos

$$\mathsf{Cov}(\boldsymbol{X}-\boldsymbol{a},\boldsymbol{X}-\boldsymbol{a}) = \mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})$$

RESULTADO 1.6. *Si $\boldsymbol{X}$ e $\boldsymbol{Y}$ son vectores aleatorios m y n-dimensionales, respectivamente y $\boldsymbol{A} \in \mathbb{R}^{l\times m}$, $\boldsymbol{B} \in \mathbb{R}^{p\times n}$, entonces*

$$\mathsf{Cov}(\boldsymbol{A}\boldsymbol{X},\boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}^\top.$$

DEMOSTRACIÓN. Sean $\boldsymbol{U} = \boldsymbol{A}\boldsymbol{X}$ y $\boldsymbol{V} = \boldsymbol{B}\boldsymbol{Y}$, entonces

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{A}\boldsymbol{X},\boldsymbol{B}\boldsymbol{Y}) &= \mathsf{Cov}(\boldsymbol{U},\boldsymbol{V}) = \mathsf{E}\{(\boldsymbol{U}-\mathsf{E}(\boldsymbol{U}))(\boldsymbol{V}-\mathsf{E}(\boldsymbol{V}))^\top\} \\ &= \mathsf{E}\{(\boldsymbol{A}\boldsymbol{X}-\boldsymbol{A}\,\mathsf{E}(\boldsymbol{X}))(\boldsymbol{B}\boldsymbol{Y}-\boldsymbol{B}\,\mathsf{E}(\boldsymbol{Y}))^\top\} \\ &= \mathsf{E}\{\boldsymbol{A}(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\boldsymbol{B}^\top\} \\ &= \boldsymbol{A}\,\mathsf{E}\{(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathsf{E}(\boldsymbol{Y}))^\top\}\boldsymbol{B}^\top \\ &= \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}^\top.\end{aligned}$$

□

Tenemos el siguiente caso particular,

$$\mathsf{Cov}(\boldsymbol{A}\boldsymbol{X}) = \mathsf{Cov}(\boldsymbol{A}\boldsymbol{X},\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{X})\boldsymbol{A}^\top = \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{A}^\top.$$

EJEMPLO 1.7. Considere $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{U}$ y $\boldsymbol{V}$ vectores aleatorios n-dimensionales y $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ y $\boldsymbol{D}$ matrices de órdenes apropiados, entonces

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{AX}+\boldsymbol{BY},\boldsymbol{CU}+\boldsymbol{DV}) &= \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{U})\boldsymbol{C}^\top + \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{V})\boldsymbol{D}^\top \\ &\quad + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y},\boldsymbol{U})\boldsymbol{C}^\top + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y},\boldsymbol{V})\boldsymbol{D}^\top.\end{aligned}$$

tomando $\boldsymbol{U}=\boldsymbol{X}$, $\boldsymbol{V}=\boldsymbol{Y}$, $\boldsymbol{C}=\boldsymbol{A}$ y $\boldsymbol{D}=\boldsymbol{B}$, tenemos

$$\begin{aligned}\mathsf{Cov}(\boldsymbol{AX}+\boldsymbol{BY}) &= \mathsf{Cov}(\boldsymbol{AX}+\boldsymbol{BY},\boldsymbol{AX}+\boldsymbol{BY}) \\ &= \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{A}^\top + \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}^\top \\ &\quad + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y},\boldsymbol{X})\boldsymbol{A}^\top + \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y})\boldsymbol{B}^\top.\end{aligned}$$

RESULTADO 1.8. *Toda matriz de dispersión es simétrica y semidefinida positiva*

DEMOSTRACIÓN. La simetría de la matriz de dispersión es obvia. Para mostrar que $\mathsf{Cov}(\boldsymbol{X})$ es semidefinida positiva, sea $\boldsymbol{Z}=\boldsymbol{X}-\mathsf{E}(\boldsymbol{X})$, y considere la variable aleatoria $Y=\boldsymbol{a}^\top\boldsymbol{Z}$, para $\boldsymbol{a}\in\mathbb{R}^n$ un vector arbitrareo. Entonces,

$$\begin{aligned}\boldsymbol{a}^\top\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{a} &= \boldsymbol{a}^\top\,\mathsf{E}(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))^\top\boldsymbol{a} \\ &= \mathsf{E}(\boldsymbol{a}^\top(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))(\boldsymbol{X}-\mathsf{E}(\boldsymbol{X}))^\top\boldsymbol{a}) \\ &= \mathsf{E}(\boldsymbol{a}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{a}) = \mathsf{E}(Y^2)\geq 0\end{aligned}$$

y por tanto, $\mathsf{Cov}(\boldsymbol{X})$ es semidefinida positiva.

Ahora, suponga que $\mathsf{Cov}(\boldsymbol{X})$ es semidefinida positiva de rango r ($r\leq n$). Luego $\mathsf{Cov}(\boldsymbol{X})=\boldsymbol{B}\boldsymbol{B}^\top$ donde $\boldsymbol{B}\in\mathbb{R}^{n\times r}$ de rango r. Sea $\boldsymbol{Y}$ vector aleatorio r-dimensional con $\mathsf{E}(\boldsymbol{Y})=\boldsymbol{0}$ y $\mathsf{Cov}(\boldsymbol{Y})=\boldsymbol{I}$. Haciendo $\boldsymbol{X}=\boldsymbol{BY}$, sigue que $\mathsf{E}(\boldsymbol{X})=\boldsymbol{0}$ y

$$\mathsf{Cov}(\boldsymbol{X})=\mathsf{Cov}(\boldsymbol{BY})=\boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Y})\boldsymbol{B}^\top=\boldsymbol{B}\boldsymbol{B}^\top.$$

Es decir, corresponde a una matriz de covarianza. □

RESULTADO 1.9. *Sea $\boldsymbol{X}$ vector aleatorio n-dimensional y considere la transformación lineal $\boldsymbol{Y}=\boldsymbol{AX}+\boldsymbol{b}$, donde $\boldsymbol{A}$ es una matriz de constantes $m\times n$ y $\boldsymbol{b}$ es vector de constantes $m\times 1$. Entonces*

$$\mathsf{E}(\boldsymbol{Y})=\boldsymbol{A}\,\mathsf{E}(\boldsymbol{X})+\boldsymbol{b},\qquad \mathsf{Cov}(\boldsymbol{Y})=\boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{A}^\top.$$

EJEMPLO 1.10. Sea $\boldsymbol{X}$ vector aleatorio n-dimensional con media $\mathsf{E}(\boldsymbol{X})=\boldsymbol{\mu}$ y matriz de dispersión $\mathsf{Cov}(\boldsymbol{X})=\boldsymbol{\Sigma}$. Sea

$$\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$$

la descomposición espectral de $\boldsymbol{\Sigma}$, donde $\boldsymbol{U}$ es matriz ortogonal y $\boldsymbol{\Lambda}=\mathrm{diag}(\boldsymbol{\lambda})$, y considere la siguiente transformación

$$\boldsymbol{Z}=\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top(\boldsymbol{X}-\boldsymbol{\mu})$$

de este modo, obtenemos que

$$\mathsf{E}(\boldsymbol{Z})=\boldsymbol{0}\qquad\text{y}\qquad\mathsf{Cov}(\boldsymbol{Z})=\boldsymbol{I}.$$

En efecto, la transformación $\boldsymbol{Z}=\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})$ también satisface que $\mathsf{E}(\boldsymbol{Z})=\boldsymbol{0}$ y $\mathsf{Cov}(\boldsymbol{Z})=\boldsymbol{I}$.

Suponga que $\boldsymbol{Z}$ es una matriz aleatoria $n \times p$ cuyas filas son vectores aleatorios independientes $p \times 1$, cada uno con la misma matriz de covarianza $\boldsymbol{\Sigma}$. Considere la partición

$$\boldsymbol{Z}^\top = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n),$$

donde $\mathsf{Cov}(\boldsymbol{Z}_i) = \boldsymbol{\Sigma}$, para $i = 1, \ldots, n$. Tenemos que

$$\operatorname{vec}(\boldsymbol{Z}^\top) = \begin{pmatrix} \boldsymbol{Z}_1 \\ \vdots \\ \boldsymbol{Z}_n \end{pmatrix},$$

y dado que todos los $\boldsymbol{Z}_i$ son independientes con la misma matriz de covarianza, podemos escribir

$$\mathsf{Cov}(\operatorname{vec}(\boldsymbol{Z}^\top)) = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma} \end{pmatrix} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}.$$

Ahora suponga que llevamos a cabo la transformación lineal $\boldsymbol{Y} = \boldsymbol{AZB}$, donde $\boldsymbol{A} \in \mathbb{R}^{r \times n}$, $\boldsymbol{B} \in \mathbb{R}^{p \times q}$ son matrices de constantes. Entonces $\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{A}\,\mathsf{E}(\boldsymbol{Z})\boldsymbol{B}$, mientras que

$$\operatorname{vec}(\boldsymbol{Y}^\top) = (\boldsymbol{A} \otimes \boldsymbol{B}^\top) \operatorname{vec}(\boldsymbol{Z}^\top),$$

de modo que

$$\mathsf{E}(\operatorname{vec}(\boldsymbol{Y}^\top)) = (\boldsymbol{A} \otimes \boldsymbol{B}^\top)\, \mathsf{E}(\operatorname{vec}(\boldsymbol{Z})^\top).$$

Lo que lleva a calcular fácilmente la matriz de covarianza

$$\begin{aligned} \mathsf{Cov}(\operatorname{vec}(\boldsymbol{Y}^\top)) &= (\boldsymbol{A} \otimes \boldsymbol{B}^\top)\, \mathsf{Cov}(\operatorname{vec}(\boldsymbol{Z}^\top))(\boldsymbol{A} \otimes \boldsymbol{B}^\top)^\top \\ &= (\boldsymbol{A} \otimes \boldsymbol{B}^\top)(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma})(\boldsymbol{A}^\top \otimes \boldsymbol{B}) \\ &= \boldsymbol{AA}^\top \otimes \boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}. \end{aligned}$$

Definición 1.11 (Matriz de correlación). Sea $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ vector aleatorio con media $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$. Se define la matriz de correlaciones como $\boldsymbol{R} = (\rho_{ij})$, donde

$$\rho_{ij} = \frac{\mathsf{Cov}(X_i, X_j)}{\{\mathsf{var}(X_i)\, \mathsf{var}(X_j)\}^{1/2}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \qquad i, j = 1, \ldots, p.$$

Note que, para $\boldsymbol{\Sigma}$ matriz de covarianza del vector aleatorio $\boldsymbol{X}$ y con $\boldsymbol{D} = \operatorname{diag}(\sigma_{11}, \ldots, \sigma_{pp})$ $(= \operatorname{diag}(\boldsymbol{\Sigma}))$ podemos escribir

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{D}^{-1/2}.$$

Cada elemento de la diagonal de $\boldsymbol{R}$ es igual a 1, mientras que sus elementos fuera de la diagonal están entre -1 y 1. Además se desprende desde la definición que $\boldsymbol{R}$ es una matriz semidefinida positiva.

Resultado 1.12. *Sea $\boldsymbol{X}$ vector aleatorio p-dimensional con $\mathsf{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ y $\mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Sea $\boldsymbol{A}$ una matriz $p \times p$. Entonces*

$$\mathsf{E}(\boldsymbol{X}^\top \boldsymbol{AX}) = \operatorname{tr}(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{A\mu}.$$

DEMOSTRACIÓN. Tenemos

$$\begin{aligned}\mathsf{E}(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}) &= \mathsf{E}(\operatorname{tr} \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}) = \mathsf{E}(\operatorname{tr} \boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top)\\ &= \operatorname{tr} \mathsf{E}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top) = \operatorname{tr} \boldsymbol{A}\, \mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top)\\ &= \operatorname{tr} \boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}.\end{aligned}$$

□

Considere el siguiente caso especial: sea $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{a}$, entonces $\mathsf{Cov}(\boldsymbol{Y}) = \mathsf{Cov}(\boldsymbol{X})$ y tenemos

$$\mathsf{E}[(\boldsymbol{X} - \boldsymbol{a})^\top \boldsymbol{A}(\boldsymbol{X} - \boldsymbol{a})] = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \boldsymbol{a})^\top \boldsymbol{A}(\boldsymbol{\mu} - \boldsymbol{a}).$$

EJEMPLO 1.13. Sea $\mathbf{1}_n = (1, \ldots, 1)^\top$ vector n-dimensional cuyos componentes son todos 1. Note que, $\mathbf{1}_n^\top \mathbf{1}_n = n$. Considere el vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, entonces

$$\boldsymbol{X}^\top \boldsymbol{X} = \sum_{i=1}^n X_i^2, \qquad \mathbf{1}^\top \boldsymbol{X} = \sum_{i=1}^n X_i.$$

De este modo, tenemos

$$\begin{aligned}\sum_{i=1}^n (X_i - \overline{X})^2 &= \sum_{i=1}^n X_i^2 - n\overline{X}^2 = \boldsymbol{X}^\top \boldsymbol{X} - n\Big(\frac{1}{n}\mathbf{1}^\top \boldsymbol{X}\Big)^2\\ &= \boldsymbol{X}^\top \boldsymbol{X} - n\Big(\frac{1}{n}\mathbf{1}^\top \boldsymbol{X}\Big)\Big(\frac{1}{n}\mathbf{1}^\top \boldsymbol{X}\Big) = \boldsymbol{X}^\top \boldsymbol{X} - \frac{1}{n}\boldsymbol{X}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{X}\\ &= \boldsymbol{X}^\top \Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}_n\Big)\boldsymbol{X}, \qquad \boldsymbol{J}_n = \mathbf{1}_n \mathbf{1}_n^\top\end{aligned}$$

Llamaremos a $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{J}_n$ la *matriz de centrado*. Suponga que $X_1, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuídas con media μ y varianza σ^2. Sigue que,

$$\mathsf{E}(\boldsymbol{X}) = \mu \mathbf{1}_n, \qquad \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n,$$

pues $\mathsf{Cov}(X_i, X_j) = 0$ $(i \neq j)$. Por tanto, podemos usar el Resultado (1.12) para calcular la esperanza de la variable aleatoria,

$$Q = \sum_{i=1}^n (X_i - \overline{X})^2 = \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X},$$

obteniendo

$$\mathsf{E}(Q) = \sigma^2 \operatorname{tr}(\boldsymbol{C}) + \mu^2 \mathbf{1}^\top \boldsymbol{C}\mathbf{1}.$$

Es fácil verificar que

$$\begin{aligned}\operatorname{tr}(\boldsymbol{C}) &= \operatorname{tr}\Big(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\Big) = \operatorname{tr}(\boldsymbol{I}) - \frac{1}{n}\operatorname{tr}(\mathbf{1}\mathbf{1}^\top) = n - \frac{1}{n}\mathbf{1}^\top \mathbf{1} = n - 1,\\ \boldsymbol{C}\mathbf{1} &= \Big(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\Big)\mathbf{1} = \mathbf{1} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0},\end{aligned}$$

de donde sigue que $\mathsf{E}(Q) = \sigma^2(n-1)$.

RESULTADO 1.14. *Si $\boldsymbol{X}$ es vector aleatorio $n \times 1$. Entonces su distribución está determinada por las distribuciones de las funciones lineales $\boldsymbol{a}^\top \boldsymbol{X}$, para todo $\boldsymbol{a} \in \mathbb{R}^n$.*

DEMOSTRACIÓN. La función característica de $\boldsymbol{a}^\top \boldsymbol{X}$ es

$$\varphi_{a^\top X}(t) = \mathsf{E}\{\exp(it\boldsymbol{a}^\top \boldsymbol{X})\},$$

de modo que

$$\varphi_{a^\top X}(1) = \mathsf{E}\{\exp(i\boldsymbol{a}^\top \boldsymbol{X})\} \quad (= \varphi_X(\boldsymbol{a})).$$

Es considerada como una función de $\boldsymbol{a}$, esto es, la función característica (conjunta) de $\boldsymbol{X}$. El resultado sigue notando que una distribución en $\mathbb{R}^n$ está completamente determinada por su función característica. □

La función característica permite un método bastante operativo para el cálculo del k-ésimo momento de un vector aleatorio $\boldsymbol{X}$. En efecto,

$$\begin{aligned}\boldsymbol{\mu}_k(\boldsymbol{X}) &= \begin{cases} \mathsf{E}(\boldsymbol{X} \otimes \boldsymbol{X}^\top \otimes \cdots \otimes \boldsymbol{X}^\top), & k \text{ par}, \\ \mathsf{E}(\boldsymbol{X} \otimes \boldsymbol{X}^\top \otimes \cdots \otimes \boldsymbol{X}^\top \otimes \boldsymbol{X}), & k \text{ impar}, \end{cases} \\ &= \begin{cases} i^{-k} \dfrac{\partial^k \varphi(\boldsymbol{t})}{\partial \boldsymbol{t} \partial \boldsymbol{t}^\top \cdots \partial \boldsymbol{t}^\top}\Big|_{t=0}, & k \text{ par}, \\ i^{-k} \dfrac{\partial^k \varphi(\boldsymbol{t})}{\partial \boldsymbol{t} \partial \boldsymbol{t}^\top \cdots \partial \boldsymbol{t}^\top \partial \boldsymbol{t}}\Big|_{t=0}, & k \text{ impar}. \end{cases}\end{aligned}$$

1.3. Independencia de vectores aleatorios

Sea $\boldsymbol{Z} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$ con $\boldsymbol{X}$, $\boldsymbol{Y}$ vectores aleatorios n y q-dimensionales, respectivamente. Se dicen independientes si y sólo si

$$F(\boldsymbol{x}, \boldsymbol{y}) = G(\boldsymbol{x})H(\boldsymbol{y}),$$

donde $F(\boldsymbol{z})$, $G(\boldsymbol{x})$ y $H(\boldsymbol{y})$ son las funciones de distribución de $\boldsymbol{Z}$, $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente.

Si $\boldsymbol{Z}$, $\boldsymbol{X}$ e $\boldsymbol{Y}$ tienen densidades $f(\boldsymbol{z})$, $g(\boldsymbol{x})$ y $h(\boldsymbol{y})$, respectivamente. Entonces $\boldsymbol{X}$ e $\boldsymbol{Y}$ son independientes si

$$f(\boldsymbol{z}) = g(\boldsymbol{x})h(\boldsymbol{y}).$$

En cuyo caso, obtenemos como resultado

$$f(\boldsymbol{x}|\boldsymbol{y}) = g(\boldsymbol{x}).$$

RESULTADO 1.15. *Sean $\boldsymbol{X}$ e $\boldsymbol{Y}$ dos vectores aleatorios independientes. Entonces para funciones cualquiera κ y τ, tenemos*

$$\mathsf{E}\{\kappa(\boldsymbol{X})\tau(\boldsymbol{Y})\} = \mathsf{E}\{\kappa(\boldsymbol{X})\}\,\mathsf{E}\{\tau(\boldsymbol{Y})\},$$

si las esperanzas existen.

DEMOSTRACIÓN. En efecto, es fácil notar que

$$\begin{aligned}\mathsf{E}\{\kappa(\boldsymbol{X})\tau(\boldsymbol{Y})\} &= \int\int \kappa(\boldsymbol{x})\tau(\boldsymbol{y})g(\boldsymbol{x})h(\boldsymbol{y})\,\mathsf{d}\boldsymbol{x}\,\mathsf{d}\boldsymbol{y} \\ &= \Big(\int \kappa(\boldsymbol{x})g(\boldsymbol{x})\,\mathsf{d}\boldsymbol{x}\Big)\Big(\int \tau(\boldsymbol{y})h(\boldsymbol{y})\,\mathsf{d}\boldsymbol{y}\Big) \\ &= \mathsf{E}\{\kappa(\boldsymbol{X})\}\,\mathsf{E}\{\tau(\boldsymbol{Y})\}.\end{aligned}$$

□

1.4. Cambios de variable

Considere la función $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$, el *Jacobiano* se define como el valor absoluto del determinante de $\mathsf{D}\boldsymbol{f}(\boldsymbol{x})$ y es denotado por

$$J(\boldsymbol{y} \to \boldsymbol{x}) = |\mathsf{D}\boldsymbol{f}(\boldsymbol{x})|_+ = \operatorname{abs}(\det(\mathsf{D}\boldsymbol{f}(\boldsymbol{x}))),$$

donde $\boldsymbol{y} = \boldsymbol{f}(x)$. Note que si $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{y})$ y $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$, entonces tenemos

$$\begin{aligned} J(\boldsymbol{z} \to \boldsymbol{x}) &= J(\boldsymbol{z} \to \boldsymbol{y}) \cdot J(\boldsymbol{y} \to \boldsymbol{x}) \\ J(\boldsymbol{y} \to \boldsymbol{x}) &= \{J(\boldsymbol{x} \to \boldsymbol{y})\}^{-1} \end{aligned}$$

El siguiente resultado presenta una de aplicación del Jacobiano de una transformación para obtener la función de densidad de una transformación de un vector aleatorio.

Proposición 1.16 (Transformación de vectores aleatorios continuos). *Sea $\boldsymbol{X}$ vector aleatorio n-dimensional con densidad $f_X(\boldsymbol{x})$ y soporte $S = \{\boldsymbol{x} : f_X(\boldsymbol{x}) > 0\}$. Para $\boldsymbol{g} : S \to \mathbb{R}^n$ diferenciable e invertible, sea $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$. Entonces la densidad de $\boldsymbol{Y}$ está dada por*

$$\begin{aligned} f_Y(\boldsymbol{y}) &= |\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{y})|_+ \, f_X(\boldsymbol{g}^{-1}(\boldsymbol{y})) \\ &= \{J(\boldsymbol{y} \to \boldsymbol{x})\}^{-1} f_X(\boldsymbol{g}^{-1}(\boldsymbol{y})). \end{aligned}$$

Ejemplo 1.17. Sea $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$, $\boldsymbol{X} \in \mathbb{R}^{n\times q}$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ y $\boldsymbol{B} \in \mathbb{R}^{q\times q}$. Entonces

$$\mathsf{d}\,\boldsymbol{Y} = \boldsymbol{A}(\mathsf{d}\,\boldsymbol{X})\boldsymbol{B},$$

vectorizando obtenemos

$$\operatorname{vec}\mathsf{d}\,\boldsymbol{Y} = (\boldsymbol{B}^\top \otimes \boldsymbol{A}) \operatorname{vec}\mathsf{d}\,\boldsymbol{X},$$

esto es, $\mathsf{D}\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{B}^\top \otimes \boldsymbol{A}$, por tanto

$$J(\boldsymbol{Y} \to \boldsymbol{X}) = |\boldsymbol{B}^\top \otimes \boldsymbol{A}|_+ = |\boldsymbol{A}|_+^q |\boldsymbol{B}^\top|_+^n = |\boldsymbol{A}|_+^q |\boldsymbol{B}|_+^n$$

1.5. Distribución normal multivariada

La distribución normal multivariada ocupa un rol central en inferencia multivariada así como en modelación estadística. En esta sección introducimos la distribución normal multivariada mediante tres definiciones equivalentes.

Una variable aleatoria (uni-dimensional) Z tiene una distribución normal con media cero y varianza uno si su función de densidad es de la forma

$$f(z) = (2\pi)^{-1/2} \exp\big(-\tfrac{1}{2}z^2\big), \quad z \in \mathbb{R},$$

en cuyo caso escribimos $Z \sim \mathsf{N}(0,1)$. Más generalmente una variable aleatoria $Y \in \mathbb{R}$ tiene distribución normal con media $\mu \in \mathbb{R}$ y varianza $\sigma^2 \geq 0$ si

$$Y \overset{\mathsf{d}}{=} \mu + \sigma Z, \qquad Z \sim \mathsf{N}(0,1),$$

en cuyo caso escribimos $Y \sim \mathsf{N}(\mu, \sigma^2)$. Cuando $\sigma^2 = 0$, la distribución $\mathsf{N}(\mu, \sigma^2)$ se interpreta como una distribución degenerada en μ. Si $Y \sim \mathsf{N}(\mu, \sigma^2)$, entonces su función característica adopta la forma

$$\varphi(t) = \exp\big(it\mu - \tfrac{1}{2}\sigma^2 t^2\big), \quad t \in \mathbb{R}.$$

Sea $Z_1, \ldots, Z_n$ variables aleatorias independientes cada una con distribución $\mathsf{N}(0,1)$ y considere el vector aleatorio $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$. De este modo, la densidad conjunta de $\boldsymbol{Z}$ es dada por

$$\begin{aligned} f(\boldsymbol{z}) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\big(-\tfrac{1}{2} z_i^2\big) = (2\pi)^{-n/2} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n} z_i^2\Big) \\ &= (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2}\|\boldsymbol{z}\|^2\big), \end{aligned}$$

y anotamos $\boldsymbol{Z} \sim \mathsf{N}_n(\boldsymbol{0}, \boldsymbol{I})$.

Definición 1.18. Un vector aleatorio p-dimensional, $\boldsymbol{X}$ tiene distribución normal con vector de medias $\boldsymbol{\mu} \in \mathbb{R}^p$ y matriz de covarianza $\mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} \geq \boldsymbol{0}$ sólo si, para todo vector $\boldsymbol{t}$ la variable aleatoria (uni-dimensional) $\boldsymbol{t}^\top \boldsymbol{X}$ es normal univariada, en cuyo caso escribimos $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Observación 1.19. Note que en la definición anterior *no* se ha hecho supuestos respecto de la independencia de los componentes de $\boldsymbol{X}$.

Resultado 1.20. *Suponga que* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y considere la transformación lineal* $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ *donde* $\boldsymbol{A} \in \mathbb{R}^{m\times p}$ *con* $\mathrm{rg}(\boldsymbol{A}) = m$. *Entonces* $\boldsymbol{Y} \sim \mathsf{N}_m(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$.

Demostración. Sea $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ y simplemente note que

$$\boldsymbol{t}^\top \boldsymbol{Y} = \boldsymbol{t}^\top \boldsymbol{A}\boldsymbol{X} + \boldsymbol{t}^\top \boldsymbol{b} = (\boldsymbol{A}^\top \boldsymbol{t})^\top \boldsymbol{X} + \boldsymbol{t}^\top \boldsymbol{b} = \boldsymbol{h}^\top \boldsymbol{X} + c,$$

por la Definición 1.18 tenemos que $\boldsymbol{h}^\top \boldsymbol{X}$ es normal y como c es una constante, sigue que $\boldsymbol{t}^\top \boldsymbol{Y}$ tiene distribución normal multivariada. □

A partir del resultado anterior sigue que todas las distribuciones marginales de $\boldsymbol{X}$ también son normalmente distribuídas. En particular, también permite apreciar que la distribución normal satisface la siguiente propiedad relativa a la simetría multivariada:

$$\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p) \Longrightarrow \boldsymbol{Q}\boldsymbol{Z} \overset{\mathrm{d}}{=} \boldsymbol{Z}, \quad \forall \boldsymbol{Q} \in \mathcal{O}_p.$$

Resultado 1.21. *Si* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *entonces la función característica de* $\boldsymbol{X}$ *es dada por*

$$\varphi_X(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top \boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma}\boldsymbol{t}).$$

Demostración. Sabemos que la función característica de un vector aleatorio, satisface

$$\varphi_X(\boldsymbol{t}) = \mathsf{E}\{\exp(i\boldsymbol{t}^\top \boldsymbol{X})\} = \varphi_{t^\top X}(1),$$

donde la función característica de la variable aleatoria uni-dimensional $Y = \boldsymbol{t}^\top \boldsymbol{X}$ es evaluada en 1. Como $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ sólo si $\boldsymbol{t}^\top \boldsymbol{X} \sim \mathsf{N}_1(\boldsymbol{t}^\top \boldsymbol{\mu}, \boldsymbol{t}^\top \boldsymbol{\Sigma}\boldsymbol{t})$, tenemos

$$\varphi_X(\boldsymbol{t}) = \exp\big(i\boldsymbol{t}^\top \boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma}\boldsymbol{t}\big).$$

En efecto, sea $\boldsymbol{\Sigma}$ matriz de covarianza $p \times p$ semidefinida positiva de rango r y sea $Z_1, \ldots, Z_r$ variables aleatorias IID $\mathsf{N}(0,1)$. Entonces el vector $\boldsymbol{Z} = (Z_1, \ldots, Z_r)^\top$ tiene función característica

$$\begin{aligned} \varphi_Z(\boldsymbol{t}) &= \mathsf{E}\{\exp(i\boldsymbol{t}^\top \boldsymbol{Z})\} = \prod_{j=1}^{r} \mathsf{E}\{\exp(it_j Z_j)\} \\ &= \prod_{j=1}^{r} \exp\big(-\tfrac{1}{2}t_j^2\big) = \exp\big(-\tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{t}\big). \end{aligned}$$

Considere

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z},$$

donde $\boldsymbol{B} \in \mathbb{R}^{p\times r}$ con $\operatorname{rg}(\boldsymbol{B}) = r$, tal que $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$ y $\boldsymbol{\mu} \in \mathbb{R}^p$. De este modo, $\boldsymbol{X}$ tiene función característica

$$\begin{aligned}\varphi_X(\boldsymbol{t}) &= \mathsf{E}\{\exp(i\boldsymbol{t}^\top \boldsymbol{X})\} = \mathsf{E}\{\exp(i\boldsymbol{t}^\top(\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z}))\} \\ &= \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\,\mathsf{E}\{\exp(i\boldsymbol{t}^\top\boldsymbol{B}\boldsymbol{Z})\} = \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\varphi_Z(\boldsymbol{h}), \quad \boldsymbol{h} = \boldsymbol{B}^\top\boldsymbol{t} \\ &= \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\exp(-\tfrac{1}{2}\boldsymbol{t}^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}).\end{aligned}$$

□

Observación 1.22. El Resultado 1.20 puede ser demostrado de manera bastante simple usando la función característica (ver Ejercicio 1.3).

Resultado 1.23. *Si* $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$. *Entonces*

$$\mathsf{E}(\boldsymbol{Z}) = \boldsymbol{0}, \qquad \mathsf{Cov}(\boldsymbol{Z}) = \boldsymbol{I}.$$

Demostración. Para mostrar el resultado deseado, podemos calcular el primer y segundo diferencial de la función característica del vector aleatorio $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$. Debemos calcular,

$$\mathsf{d}\varphi_Z(\boldsymbol{t}) = -\varphi_Z(\boldsymbol{t})\boldsymbol{t}^\top\,\mathsf{d}\boldsymbol{t},$$

y

$$\begin{aligned}\mathsf{d}^2\,\varphi_Z(\boldsymbol{t}) &= -\,\mathsf{d}\varphi_Z(\boldsymbol{t})\boldsymbol{t}^\top\,\mathsf{d}\boldsymbol{t} - \varphi_Z(\boldsymbol{t})(\mathsf{d}\boldsymbol{t})^\top\,\mathsf{d}\boldsymbol{t} \\ &= \varphi_Z(\boldsymbol{t})(\mathsf{d}\boldsymbol{t})^\top\boldsymbol{t}\boldsymbol{t}^\top\,\mathsf{d}\boldsymbol{t} - \varphi_Z(\boldsymbol{t})(\mathsf{d}\boldsymbol{t})^\top\,\mathsf{d}\boldsymbol{t} \\ &= \varphi_Z(\boldsymbol{t})(\mathsf{d}\boldsymbol{t})^\top(\boldsymbol{t}\boldsymbol{t}^\top - \boldsymbol{I})\,\mathsf{d}\boldsymbol{t},\end{aligned}$$

de ahí que

$$\frac{\partial\varphi_Z(\boldsymbol{t})}{\partial\boldsymbol{t}} = -\varphi_Z(\boldsymbol{t})\boldsymbol{t}, \qquad \frac{\partial^2\varphi_Z(\boldsymbol{t})}{\partial\boldsymbol{t}\,\partial\boldsymbol{t}^\top} = \varphi_Z(\boldsymbol{t})(\boldsymbol{t}\boldsymbol{t}^\top - \boldsymbol{I}).$$

Ahora, el vector de medias y matriz de covarianzas están dadas por

$$\begin{aligned}\mathsf{E}(\boldsymbol{Z}) &= i^{-1}\frac{\partial\varphi_Z(\boldsymbol{t})}{\partial\boldsymbol{t}}\Big|_{t=0} = \boldsymbol{0}, \\ \mathsf{E}(\boldsymbol{Z}\boldsymbol{Z}^\top) &= i^{-2}\frac{\partial^2\varphi_Z(\boldsymbol{t})}{\partial\boldsymbol{t}\,\partial\boldsymbol{t}^\top}\Big|_{t=0} = \boldsymbol{I} = \mathsf{Cov}(\boldsymbol{Z}).\end{aligned}$$

□

Observación 1.24. Considere

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z}, \qquad \boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top,$$

con $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$. Usando los Resultados 1.20 y 1.23, sigue que

$$\mathsf{E}(\boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{B}\,\mathsf{E}(\boldsymbol{Z}) = \boldsymbol{\mu}, \qquad \mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{B}\,\mathsf{Cov}(\boldsymbol{Z})\boldsymbol{B}^\top = \boldsymbol{\Sigma}.$$

Resultado 1.25. *Si* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *entonces la distribución marginal de cualquier subconjunto de* k $(< p)$ *componentes de* $\boldsymbol{X}$ *es normal* k*-variada.*

DEMOSTRACIÓN. Considere la siguiente partición:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \tag{1.3}$$

donde $\boldsymbol{X}_1$ y $\boldsymbol{\mu}_1$ son vectores $k \times 1$ y $\boldsymbol{\Sigma}_{11}$ es $k \times k$. Aplicando el Resultado 1.20 con

$$\boldsymbol{A} = (\boldsymbol{I}_k, \boldsymbol{0}) \in \mathbb{R}^{k \times p} \qquad \text{y} \qquad \boldsymbol{b} = \boldsymbol{0},$$

sigue inmediatamente que $\boldsymbol{X}_1 \sim \mathsf{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. □

Una consecuencia de este resultado es que la distribución marginal de *cada* componente de $\boldsymbol{X}$ es normal univariada.

OBSERVACIÓN 1.26. La inversa del Resultado 1.25 *no* es verdad en general. Es decir, que cada componente de un vector aleatorio tenga distribución normal no implica que todo el vector siga una distribución normal multivariada.

RESULTADO 1.27. *Si* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y* $\boldsymbol{X}$, $\boldsymbol{\mu}$ *y* $\boldsymbol{\Sigma}$ *son particionadas como en la Ecuación (1.3). Entonces los vectores* $\boldsymbol{X}_1$ *y* $\boldsymbol{X}_2$ *son independientes si y sólo si* $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

DEMOSTRACIÓN. Note que $\mathsf{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12}$, así la independencia entre $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$ implica que $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$. Suponga ahora que $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$. Entonces la función característica

$$\begin{aligned} \varphi_X(\boldsymbol{t}) &= \exp(i\boldsymbol{t}^\top \boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}) \\ &= \exp(i\boldsymbol{t}_1^\top \boldsymbol{\mu}_1 + i\boldsymbol{t}_2^\top \boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{t}_1^\top \boldsymbol{\Sigma}_{11} \boldsymbol{t}_1 - \tfrac{1}{2}\boldsymbol{t}_2^\top \boldsymbol{\Sigma}_{22} \boldsymbol{t}_2) \\ &= \exp(i\boldsymbol{t}_1^\top \boldsymbol{\mu}_1 - \tfrac{1}{2}\boldsymbol{t}_1^\top \boldsymbol{\Sigma}_{11} \boldsymbol{t}_1) \exp(i\boldsymbol{t}_2^\top \boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{t}_2^\top \boldsymbol{\Sigma}_{22} \boldsymbol{t}_2) \\ &= \varphi_{X_1}(\boldsymbol{t}_1) \varphi_{X_2}(\boldsymbol{t}_2), \end{aligned}$$

es decir, $\boldsymbol{X}_1 \sim \mathsf{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ es independiente de $\boldsymbol{X}_2 \sim \mathsf{N}_{p-k}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. □

DEFINICIÓN 1.28. Si $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\boldsymbol{\Sigma}$ es definida positiva, entonces la densidad de $\boldsymbol{X}$ asume la forma

$$f_X(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}, \qquad \boldsymbol{x} \in \mathbb{R}^p.$$

DEMOSTRACIÓN. Sea $Z_1, \ldots, Z_p$ variables aleatorias $\mathsf{IID}\ \mathsf{N}(0,1)$. Tenemos que la densidad conjunta de $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ es

$$f_Z(\boldsymbol{z}) = (2\pi)^{-p/2} \exp(-\tfrac{1}{2}\|\boldsymbol{z}\|^2).$$

Considere $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Z}$ con $\boldsymbol{\mu} \in \mathbb{R}^p$ y $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$, con $\boldsymbol{B}$ matriz de rango completo. Entonces, tenemos la transformación inversa

$$\boldsymbol{Z} = \boldsymbol{g}^{-1}(\boldsymbol{X}) = \boldsymbol{B}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}),$$

y $\mathsf{d}\boldsymbol{Z} = \mathsf{d}\boldsymbol{g}^{-1}(\boldsymbol{X}) = \boldsymbol{B}^{-1}\,\mathsf{d}\boldsymbol{X}$, con matriz jacobiana $\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{X}) = \boldsymbol{B}^{-1}$, como

$$|\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{X})|_+ = |\boldsymbol{B}|^{-1} = |\boldsymbol{B}\boldsymbol{B}^\top|^{-1/2},$$

obtenemos

$$\begin{aligned} f_X(\boldsymbol{x}) &= |\mathsf{D}\boldsymbol{g}^{-1}(\boldsymbol{x})|_+ f_Z(\boldsymbol{g}^{-1}(\boldsymbol{x})) \\ &= (2\pi)^{-p/2} |\boldsymbol{B}\boldsymbol{B}^\top|^{-1/2} \exp\{\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{B}^{-\top} \boldsymbol{B}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}, \end{aligned}$$

notando que $\boldsymbol{\Sigma}^{-1} = \boldsymbol{B}^{-\top}\boldsymbol{B}^{-1}$ sigue el resultado deseado. □

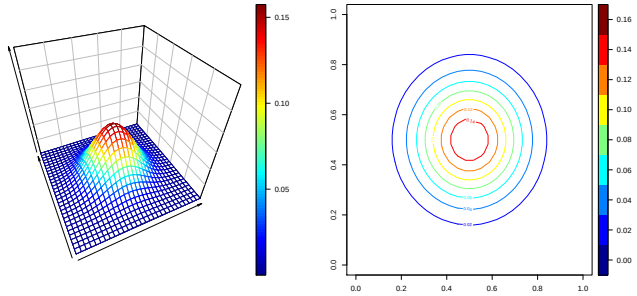
EJEMPLO 1.29. Sea $\boldsymbol{X} \sim \mathsf{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$ donde

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad -1 < \rho < 1.$$

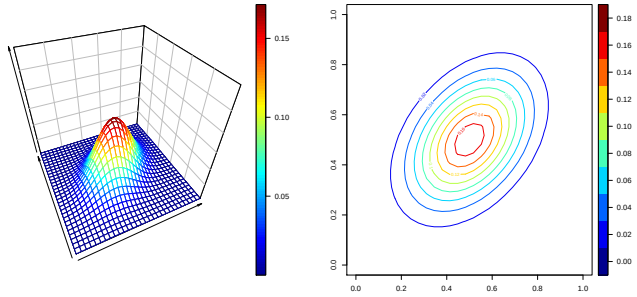
En cuyo caso, la función de densidad es dada por:

$$f(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2(1-\rho^2)}(x_1^2 + x_2^2 - 2\rho x_1 x_2)\Big\}.$$

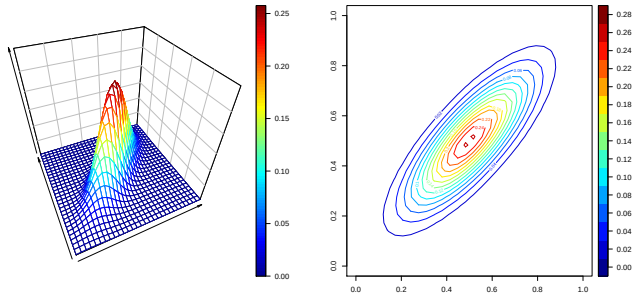
A continuación se presenta la función de densidad para los casos $\rho = 0.0, 0.4$ y 0.8.



(a) $\rho = 0.0$ (b) contornos



(c) $\rho = 0.4$ (d) contornos



(e) $\rho = 0.8$ (f) contornos

FIGURA 1. Densidad de $\boldsymbol{X} \sim \mathsf{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$ para $\rho = 0.0, 0.4$ y 0.8.

Es fácil apreciar que la función de densidad es constante sobre el elipsoide

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \lambda,$$

en $\mathbb{R}^p$ para todo $\lambda > 0$. Este elipsoide tiene centro $\boldsymbol{\mu}$, mientras que $\boldsymbol{\Sigma}$ determina su forma y orientación. Además, la variable aleatoria

$$(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{Z}^\top \boldsymbol{Z} = \sum_{i=1}^{p} Z_i^2, \tag{1.4}$$

sigue una distribución chi-cuadrado con p grados de libertad y la cantidad $D = \{(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\}^{1/2}$ se conoce como *distancia de Mahalanobis* de $\boldsymbol{X}$ a $\boldsymbol{\mu}$.

Observación 1.30. Para la existencia de densidad hemos asumido que $\boldsymbol{\Sigma} > \mathbf{0}$. En el caso de que $\boldsymbol{\Sigma} \geq \mathbf{0}$ decimos que $\boldsymbol{X}$ sigue una distribución normal singular.

Para introducir una definición de la función de densidad asociada a una variable con distribución normal singular, note que $X \sim \mathsf{N}(\mu, \sigma^2)$ con $\sigma^2 = 0 \Leftrightarrow x = \mu$ con probabilidad 1 (pues si $\sigma^2 = 0$, $\mathsf{P}(X = \mu) = \lim_{n\to\infty} \mathsf{P}(|X-\mu| < 1/n) = 0$, $\forall n$).

Considere $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\operatorname{rg}(\boldsymbol{\Sigma}) = r < p$. Entonces, podemos escribir

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = (\boldsymbol{U}_1, \boldsymbol{U}_2)\begin{pmatrix}\boldsymbol{\Lambda}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{U}_1^\top\\ \boldsymbol{U}_2^\top\end{pmatrix} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1^\top,$$

donde $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$. De este modo, es claro que

$$\boldsymbol{U}^\top\boldsymbol{\Sigma}\boldsymbol{U} \quad \Longrightarrow \quad \boldsymbol{U}_2^\top\boldsymbol{\Sigma}\boldsymbol{U}_2 = \mathbf{0},$$

es decir, tenemos que $\boldsymbol{U}_2^\top(\boldsymbol{Y}-\boldsymbol{\mu}) = \mathbf{0}$ con probabilidad 1. Mientras que

$$\boldsymbol{U}_1^\top(\boldsymbol{Y}-\boldsymbol{\mu}) \sim \mathsf{N}_r(\mathbf{0}, \boldsymbol{\Lambda}_1).$$

Además $\boldsymbol{\Sigma}^- = \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top = \boldsymbol{U}_1(\boldsymbol{U}_1^\top\boldsymbol{\Sigma}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1^\top$. Así, $\boldsymbol{Y}$ tiene la siguiente densidad normal (singular)

$$\begin{aligned} f_Y(\boldsymbol{y}) &= |2\pi\,\boldsymbol{U}_1^\top\boldsymbol{\Sigma}\boldsymbol{U}_1|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{U}_1^\top(\boldsymbol{y}-\boldsymbol{\mu}))^\top(\boldsymbol{U}_1^\top\boldsymbol{\Sigma}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1^\top(\boldsymbol{y}-\boldsymbol{\mu})\}\\ &= (2\pi)^{-r/2}|\boldsymbol{\Lambda}_1|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top(\boldsymbol{y}-\boldsymbol{\mu})\}. \end{aligned}$$

El siguiente resultado presenta la distribución condicional de un vector aleatorio con distribución normal multivariada.

Resultado 1.31. *Sea $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y particione $\boldsymbol{X}$, $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ como:*

$$\boldsymbol{X} = \begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix},$$

donde $\boldsymbol{X}_1$ y $\boldsymbol{\mu}_1$ son vectores $k \times 1$, mientras que $\boldsymbol{\Sigma}_{11}$ es matriz $k \times k$. Sea $\boldsymbol{\Sigma}_{22}^-$ una inversa generalizada de $\boldsymbol{\Sigma}_{22}$, esto es, una matriz que satisface

$$\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^-\boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_{22},$$

y sea $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-\boldsymbol{\Sigma}_{21}$. Entonces

(a) $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-\boldsymbol{X}_2 \sim \mathsf{N}_k(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11\cdot 2})$ y es independiente de $\boldsymbol{X}_2$.

(b) La distribución condicional

$$(\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2) \sim \mathsf{N}_k(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11\cdot 2}).$$

DEMOSTRACIÓN. Considere la transformación lineal

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_k & -\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I}_{p-k} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} = \boldsymbol{C}\boldsymbol{X},$$

sigue que $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top)$, donde

$$\begin{aligned} \boldsymbol{C}\boldsymbol{\mu} &= \begin{pmatrix} \boldsymbol{I}_k & -\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I}_{p-k} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{B}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix} \\ \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top &= \begin{pmatrix} \boldsymbol{I}_k & -\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I}_{p-k} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ -\boldsymbol{B}^\top & \boldsymbol{I}_{p-k} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{B}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{B}^\top + \boldsymbol{B}\boldsymbol{\Sigma}_{22}\boldsymbol{B}^\top & \boldsymbol{\Sigma}_{12} - \boldsymbol{B}\boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{22}\boldsymbol{B}^\top & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \end{aligned}$$

De este modo, nuestro interés es escoger $\boldsymbol{\Sigma}_{12} - \boldsymbol{B}\boldsymbol{\Sigma}_{22} = \boldsymbol{0}$. Es decir, $\boldsymbol{\Sigma}_{12} = \boldsymbol{B}\boldsymbol{\Sigma}_{22}$. Por otro lado, notando que

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{\Sigma}_{22} = \boldsymbol{B}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{\Sigma}_{22} = \boldsymbol{B}\boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_{12},$$

sigue que $\boldsymbol{\Sigma}_{12}\boldsymbol{B}^\top = \boldsymbol{B}\boldsymbol{\Sigma}_{22}\boldsymbol{B}^\top$ (y análogamente $\boldsymbol{B}\boldsymbol{\Sigma}_{21} = \boldsymbol{B}\boldsymbol{\Sigma}_{22}\boldsymbol{B}^\top$). Esto es, si $\boldsymbol{B}$ es escogida como $\boldsymbol{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}$, entonces $\boldsymbol{Y}_1$ y $\boldsymbol{Y}_2$ son independientes con distribución conjunta

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{X}_2 \\ \boldsymbol{X}_2 \end{pmatrix} \sim \mathsf{N}_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11\cdot 2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

Esto muestra la parte (a). Para notar la parte (b), note que las densidades de $\boldsymbol{Y}_1$ y $\boldsymbol{Y}_2$ están dadas por

$$\begin{aligned} g(\boldsymbol{y}_1; \boldsymbol{\delta}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2}) &= |2\pi\boldsymbol{\Sigma}_{11\cdot 2}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{\delta}_{1\cdot 2})^\top \boldsymbol{\Sigma}_{11\cdot 2}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\delta}_{1\cdot 2})\} \\ f_2(\boldsymbol{y}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) &= |2\pi\boldsymbol{\Sigma}_{22}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2)\}, \end{aligned}$$

y la densidad conjunta para $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)^\top$ adopta la forma

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = g(\boldsymbol{y}_1; \boldsymbol{\delta}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2})\, f_2(\boldsymbol{y}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}).$$

Como

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f_{1|2}(\boldsymbol{x}_1; \boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{x}_2)\, f_2(\boldsymbol{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}),$$

entonces, la densidad condicional de $\boldsymbol{X}_1$ dado $\boldsymbol{X}_2 = \boldsymbol{x}_2$ debe ser $g(\boldsymbol{y}_1; \boldsymbol{\delta}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2})$. Además, es fácil notar que la forma cuadrática

$$\begin{aligned} q(\boldsymbol{y}_1; \boldsymbol{\mu}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2}) &= (\boldsymbol{y}_1 - \boldsymbol{\delta}_{1\cdot 2})^\top \boldsymbol{\Sigma}_{11\cdot 2}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\delta}_{1\cdot 2}) \\ &= (\boldsymbol{x}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{x}_2 - \boldsymbol{\delta}_{1\cdot 2})^\top \boldsymbol{\Sigma}_{11\cdot 2}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}\boldsymbol{x}_2 - \boldsymbol{\delta}_{1\cdot 2}) \\ &= (\boldsymbol{x}_1 - \boldsymbol{\mu}_{1\cdot 2})^\top \boldsymbol{\Sigma}_{11\cdot 2}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_{1\cdot 2}), \end{aligned}$$

donde

$$\boldsymbol{\mu}_{1\cdot 2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2),$$

lo que muestra el resultado. □

OBSERVACIÓN 1.32. La esperanza de la distribución condicional de $\boldsymbol{X}_1$ dado $\boldsymbol{X}_2$, es decir

$$\mathsf{E}(\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2),$$

se denomina *función de regresión* de $\boldsymbol{X}_1$ sobre $\boldsymbol{X}_2$ con coeficientes de regresión $\boldsymbol{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-}$. Esta es una función lineal de $\boldsymbol{X}_2$ y la matriz de covarianza $\boldsymbol{\Sigma}_{11\cdot 2}$ no depende de $\boldsymbol{X}_2$.

RESULTADO 1.33. *Sea* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *y considere* $\boldsymbol{Y}_1 = \boldsymbol{A}_1\boldsymbol{X}$, $\boldsymbol{Y}_2 = \boldsymbol{A}_2\boldsymbol{X}$ *dos funciones lineales del vector aleatorio* $\boldsymbol{X}$. *La covarianza entre* $\boldsymbol{Y}_1$ *y* $\boldsymbol{Y}_2$ *es dada por*

$$\mathsf{Cov}(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \boldsymbol{A}_1 \,\mathsf{Cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{A}_2^\top = \boldsymbol{A}_1\boldsymbol{\Sigma}\boldsymbol{A}_2^\top$$

Este resultado permite obtener una condición para la independencia entre dos formas lineales en variables aleatorias normales, estos es $\boldsymbol{Y}_1$ y $\boldsymbol{Y}_2$ serán independientes si y sólo si $\boldsymbol{A}_1\boldsymbol{\Sigma}\boldsymbol{A}_2^\top = \boldsymbol{0}$.

EJEMPLO 1.34. Considere $X_1, \ldots, X_n$ una muestra aleatoria desde $\mathsf{N}(\mu, \sigma^2)$ y sea $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$ el vector de datos centrados con $Z_i = X_i - \overline{X}$, $i = 1, \ldots, n$, donde $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Podemos escribir

$$\overline{X} = \frac{1}{n}\boldsymbol{1}^\top\boldsymbol{X}, \qquad \boldsymbol{Z} = \boldsymbol{C}\boldsymbol{X},$$

con $\boldsymbol{C} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$ la matriz de centrado. Tenemos que $\boldsymbol{X} \sim \mathsf{N}_n(\mu\boldsymbol{1}, \sigma^2\boldsymbol{I}_n)$ y $\overline{X}$ con $\boldsymbol{Z}$ son independientes pues $\boldsymbol{C}\boldsymbol{1} = \boldsymbol{0}$.

EJEMPLO 1.35. Sea $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ y considere las transformaciones $\boldsymbol{Y}_1 = \boldsymbol{A}\boldsymbol{X}$ y $\boldsymbol{Y}_2 = (\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})^\top\boldsymbol{X}$. De este modo

$$\mathsf{Cov}(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \mathsf{Cov}(\boldsymbol{A}\boldsymbol{X}, (\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})^\top\boldsymbol{X}) = \sigma^2\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A}) = \boldsymbol{0},$$

pues $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$ y $\boldsymbol{Y}_1$ con $\boldsymbol{Y}_2$ son independientes.

1.6. Alternativas a la distribución normal multivariada

La distribución normal multivariada es de importancia fundamental en la teoría clásica de modelos lineales así como para análisis multivariado. A pesar de su uso amplio, es bien sabido que la inferencia estadística basada en la distribución normal es vulnerable a la presencia de datos atípicos, esto ha motivado considerar distribuciones alternativas que eviten este tipo de limitaciones. En esta dirección, varios autores han sugerido utilizar la clase de distribuciones elípticas (ver, por ejemplo, Fang et al., 1990; Arellano, 1994) particularmente debido al hecho de incluir distribuciones con colas más pesadas que la normal, tales como la t de Student, exponencial potencia y normal contaminada, entre otras. Una subclase importante de la familia de distribuciones elípticas es la clase de distribuciones de mezcla de escala normal (Andrews y Mallows, 1974) la que tiene propiedades similares a la distribución normal, es relativamente simple de trabajar y permite proponer procedimientos para estimación robusta. A continuación se presenta la definición y algunos ejemplos de distribuciones en la clase elíptica.

DEFINICIÓN 1.36. Sea $\boldsymbol{U}$ vector aleatorio $p \times 1$ con *distribución uniforme* sobre el conjunto

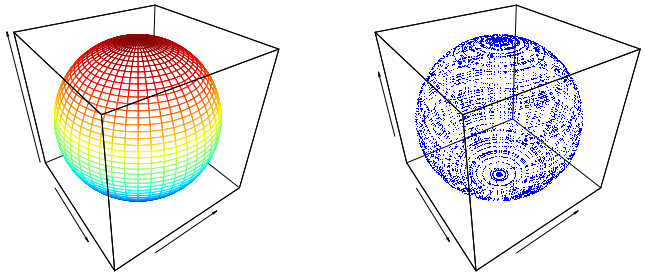
$$\mathcal{S}_p = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\| = 1\}, \tag{1.5}$$

esto es $\mathcal{S}_p$ denota la *superficie de la esfera unitaria* en $\mathbb{R}^p$. En cuyo caso anotamos $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$.

PROPIEDAD 1.37. Si $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$, entonces $\boldsymbol{U} = (U_1, \ldots, U_p)^\top \sim \mathsf{U}(\mathcal{S}_p)$, donde

$$\boldsymbol{U} = \frac{\boldsymbol{Z}}{\|\boldsymbol{Z}\|}.$$

El resultado anterior es muy relevante pues permite definir la densidad de un vector aleatorio $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$ y ofrece un procedimiento muy simple para generar observaciones sobre la esfera unitaria. Considere el siguiente gráfico,



(a) Esfera unitaria (b) Datos simulados

FIGURA 2. Esfera unitaria y datos generados sobre la superficie $\mathcal{S}_p$.

DEFINICIÓN 1.38. Un vector aleatorio $p \times 1$, $\boldsymbol{X}$ se dice que tiene simetría esférica si para cualquier $\boldsymbol{Q} \in \mathcal{O}_p$, sigue que

$$\boldsymbol{Q}\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{X}.$$

EJEMPLO 1.39. Sea $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$, entonces es bastante obvio que $\boldsymbol{Q}\boldsymbol{U} \overset{\mathsf{d}}{=} \boldsymbol{U}$.

EJEMPLO 1.40. Suponga $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Tenemos que

$$\boldsymbol{Q}\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{I}),$$

para $\boldsymbol{Q} \in \mathcal{O}_p$, es decir $\boldsymbol{Q}\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{X}$ tiene simetría esférica.

DEFINICIÓN 1.41. Un vector aleatorio p-dimensional tiene *distribución esférica* sólo si su función característica satisface

(a) $\varphi(\boldsymbol{Q}^\top\boldsymbol{t}) = \varphi(\boldsymbol{t})$, para todo $\boldsymbol{Q} \in \mathcal{O}_p$.
(b) Existe una función $\psi(\cdot)$ de una variable escalar tal que $\varphi(\boldsymbol{t}) = \psi(\boldsymbol{t}^\top\boldsymbol{t})$.

En este caso escribimos $\boldsymbol{X} \sim \mathsf{S}_p(\psi)$.

EJEMPLO 1.42. Sea $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$, tenemos que

$$\varphi(\boldsymbol{t}) = \exp\{-\tfrac{1}{2}(t_1^2 + \cdots + t_p^2)\} = \exp(-\tfrac{1}{2}\boldsymbol{t}^\top\boldsymbol{t}).$$

RESULTADO 1.43. *Sea $\psi(\boldsymbol{t}^\top\boldsymbol{t})$ la función característica del vector aleatorio $\boldsymbol{X}$. Entonces $\boldsymbol{X}$ tiene representación estocástica*

$$\boldsymbol{X} \overset{\mathsf{d}}{=} R\boldsymbol{U},$$

donde $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$ y $R \sim F(\boldsymbol{X})$ son independientes.

RESULTADO 1.44. *Suponga que $\boldsymbol{X} \overset{\mathsf{d}}{=} R\boldsymbol{U} \sim \mathsf{S}_p(\psi)$ $(\mathsf{P}(\boldsymbol{X} = \boldsymbol{0}) = 0)$, entonces*

$$\|\boldsymbol{X}\| \overset{\mathsf{d}}{=} R, \qquad \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \overset{\mathsf{d}}{=} \boldsymbol{U}.$$

Además $\|\boldsymbol{X}\|$ y $\boldsymbol{X}/\|\boldsymbol{X}\|$ son independientes.

RESULTADO 1.45. *El vector de medias y la matriz de covarianza de* $\boldsymbol{U} \sim \mathsf{U}(\mathcal{S}_p)$ *son:*

$$\mathsf{E}(\boldsymbol{U}) = \boldsymbol{0}, \qquad \mathsf{Cov}(\boldsymbol{U}) = \frac{1}{p}\boldsymbol{I}_p,$$

respectivamente.

DEMOSTRACIÓN. Sea $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$, tenemos que $\boldsymbol{X} \overset{\mathsf{d}}{=} \|\boldsymbol{X}\|\boldsymbol{U}$, con $\|\boldsymbol{X}\|$ independiente de $\boldsymbol{U}$. Sabemos que $\|\boldsymbol{X}\|^2 \sim \chi^2(p)$. Dado que

$$\mathsf{E}(\boldsymbol{X}) = \boldsymbol{0},\ \mathsf{E}(\|\boldsymbol{X}\|) > 0, \quad \text{y} \quad \mathsf{E}(\|\boldsymbol{X}\|^2) = p,\ \mathsf{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p,$$

el resultado sigue. □

RESULTADO 1.46. *Si* $\boldsymbol{X} \overset{\mathsf{d}}{=} R\boldsymbol{U} \sim \mathsf{S}_p(g)$ *y* $\mathsf{E}(R^2) < \infty$. *Entonces,*

$$\mathsf{E}(\boldsymbol{X}) = \boldsymbol{0}, \qquad \mathsf{Cov}(\boldsymbol{X}) = \frac{\mathsf{E}(R^2)}{p}\boldsymbol{I}_p,$$

respectivamente.

DEMOSTRACIÓN. En efecto, como R y $\boldsymbol{U}$ son independientes, sigue que

$$\mathsf{E}(\boldsymbol{X}) = \mathsf{E}(R)\,\mathsf{E}(\boldsymbol{U}) = \boldsymbol{0},$$

$$\mathsf{Cov}(\boldsymbol{X}) = \mathsf{E}(R^2)\,\mathsf{E}(\boldsymbol{U}\boldsymbol{U}^\top) = \mathsf{E}(R^2)\,\mathsf{Cov}(\boldsymbol{U}) = \frac{\mathsf{E}(R^2)}{p}\boldsymbol{I}_p,$$

siempre que $\mathsf{E}(R) < \infty$ y $\mathsf{E}(R^2) < \infty$. □

DEFINICIÓN 1.47. Un vector aleatorio $p \times 1$, $\boldsymbol{X}$ tiene *distribución de contornos elípticos* con parámetros $\boldsymbol{\mu} \in \mathbb{R}^p$ y $\boldsymbol{\Sigma} \geq \boldsymbol{0}$ si

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{Y}, \qquad \boldsymbol{Y} \sim \mathsf{S}_k(\psi),$$

donde $\boldsymbol{B} \in \mathbb{R}^{k\times p}$ es matriz de rango completo tal que, $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{\Sigma}$ con $\mathrm{rg}(\boldsymbol{\Sigma}) = k$ y escribimos $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \psi)$.

OBSERVACIÓN 1.48. La función característica de $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \psi)$ es de la forma

$$\varphi(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top\boldsymbol{\mu})\psi(\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}).$$

Note además que la representación estocástica de $\boldsymbol{X}$ es dada por

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + R\,\boldsymbol{B}\boldsymbol{U},$$

donde $R \geq 0$ es independiente de $\boldsymbol{U}$ y $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{\Sigma}$.

RESULTADO 1.49. *Suponga que* $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \psi)$ *y* $\mathsf{E}(R^2) < \infty$. *Entonces*

$$\mathsf{E}(\boldsymbol{X}) = \boldsymbol{\mu}, \qquad \mathsf{Cov}(\boldsymbol{X}) = \frac{\mathsf{E}(R^2)}{p}\boldsymbol{\Sigma}.$$

DEFINICIÓN 1.50. Se dice que el vector $\boldsymbol{X}$ tiene distribución de contornos elípticos si su función de densidad es de la forma

$$f(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2} g((\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})), \qquad \boldsymbol{x} \in \mathbb{R}^p,$$

donde $g : \mathbb{R} \to [0, \infty)$ es función decreciente, llamada *función generadora de densidad*, tal que:

$$\int_0^\infty u^{p/2-1} g(u)\,\mathsf{d}u < \infty,$$

y escribimos $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; g)$.

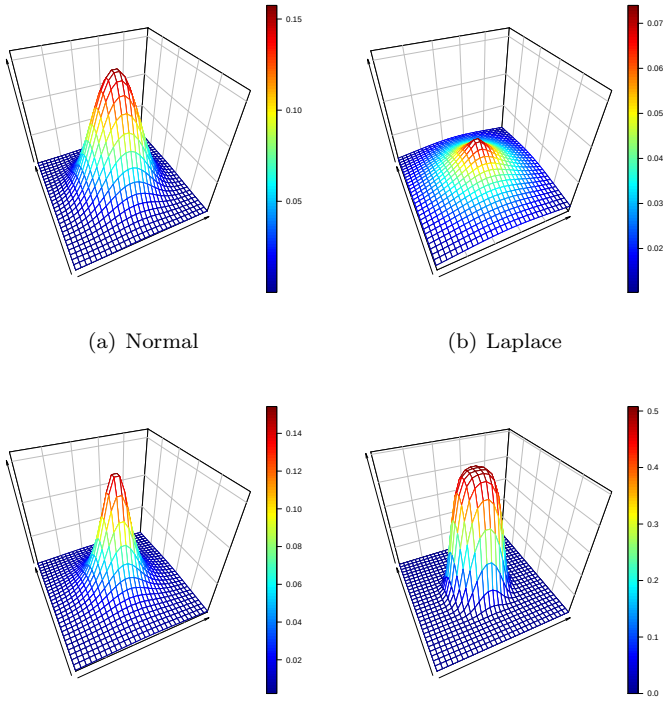
OBSERVACIÓN 1.51. Asuma que $\boldsymbol{X} \sim \mathsf{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \psi)$ con $\mathrm{rg}(\boldsymbol{\Sigma}) = k$. Entonces,

$$U = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^- (\boldsymbol{X} - \boldsymbol{\mu}) \overset{\mathsf{d}}{=} R^2,$$

donde $\boldsymbol{\Sigma}^-$ es una inversa generalizada de $\boldsymbol{\Sigma}$.

EJEMPLO 1.52. En la siguiente figura se presenta la densidad asociadas a las siguientes funciones g:

- Normal: $g(u) = c_1 \exp(-u/2)$.
- Laplace: $g(u) = c_2 \exp(-\sqrt{u}/2)$.
- Cauchy: $g(u) = c_3(1+u)^{-(p+1)/2}$.
- Exponencial potencia (PE): $g(u) = c_4 \exp(-u^\lambda/2)$, $\lambda = 2$.



(a) Normal (b) Laplace

(c) Cauchy (d) PE, $\lambda = 2$

FIGURA 3. Funciones de densidad del vector $\boldsymbol{X} \sim \mathsf{EC}_2(\boldsymbol{0}, \boldsymbol{I}; g)$ para las distribuciones normal, Laplace, Cauchy y exponencial potencia con $\lambda = 2$.

EJEMPLO 1.53 (Distribución t de Student). La función generadora de densidad de un vector aleatorio con distribución t de Student asume la forma

$$g(u) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{p/2}} \left(1 + \frac{u}{\nu}\right)^{-(\nu+p)/2}, \qquad \nu > 0.$$

Para la distribución t de Student, tenemos que $R^2/p \sim F_{p,\nu}$. Además, la función característica de $\boldsymbol{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ es dada por

$$\varphi(\boldsymbol{t}) = \frac{\|\sqrt{\nu}\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}\|^{\nu/2}}{2^{\nu/2-1}\Gamma(\nu/2)} \exp\{i\boldsymbol{t}^\top\boldsymbol{\mu}\} K_{\nu/2}(\|\sqrt{\nu}\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}\|), \qquad \boldsymbol{t} \in \mathbb{R}^p,$$

donde $K_\nu(x)$ denota la función de Bessel modificada de segundo tipo. Un caso particular importante corresponde a la distribución Cauchy, cuando $\nu = 1$, mientras que la distribución normal corresponde al caso límite $\nu \to \infty$.

Ejemplo 1.54 (Distribución Exponencial Potencia). Para la distribución Exponencial Potencia (Gómez et al., 1988), la función generadora de densidades es dada por

$$g(u) = \frac{p\Gamma(\frac{p}{2})\pi^{-p/2}}{\Gamma(1+\frac{p}{2\lambda})2^{1+\frac{p}{2\lambda}}} \exp(-u^\lambda/2), \qquad \lambda > 0.$$

y es usual utilizar la notación $\boldsymbol{X} \sim \mathsf{PE}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda)$. En este caso tenemos que la variable aleatoria positiva R tiene densidad

$$h(r) = \frac{p}{\Gamma(1+\frac{p}{2\lambda})2^{\frac{p}{2\lambda}}} r^{p-1} \exp(-r^{2\lambda}/2), \qquad r > 0.$$

Note también que $R^{2\lambda} \sim \mathsf{Gama}(\frac{1}{2}, \frac{p}{2\lambda})$. Debemos destacar que esta clase de distribuciones contiene la distribución normal como un caso particular cuando $\lambda = 1$. Mientras que tiene colas más pesadas que la normal si $\lambda < 1$ y colas más livianas para el caso $\lambda > 1$. Otro caso particular de interés es la distribución Laplace, que es recuperada cuando $\lambda = 1/2$.

Definición 1.55. Sea $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma}$ matriz $p \times p$ definida positiva y H función de distribución de la variable aleatoria positiva W. Entonces, se dice que el vector aleatorio $\boldsymbol{X}$ sigue una *distribución de mezcla de escala normal* si su función de densidad asume la forma

$$f(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \int_0^\infty w^{p/2} \exp(-wu/2)\, \mathsf{dH}(w),$$

donde $u = (\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ y anotamos $\boldsymbol{X} \sim \mathsf{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathsf{H})$.

Ejemplo 1.56 (Distribución Slash). Un vector aleatorio $\boldsymbol{X}$ tiene distribución Slash si su función de densidad es de la forma:

$$f(\boldsymbol{x}) = \nu(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \int_0^1 w^{p/2+\nu-1} \exp(-wu/2)\, \mathsf{d}w.$$

En este caso, tenemos que $h(w) = \nu w^{\nu-1}$, para $w \in (0,1)$ y $\nu > 0$. Es decir $W \sim \mathsf{Beta}(\nu, 1)$.

Observación 1.57. Un vector aleatorio $\boldsymbol{X} \sim \mathsf{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathsf{H})$ admite la representación

$$\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + W^{-1/2}\boldsymbol{Z},$$

donde $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ y $W \sim \mathsf{H}(\boldsymbol{\nu})$ son independientes. También podemos utilizar la siguiente estructura jerárquica:

$$\boldsymbol{X}|W \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/W), \qquad W \sim \mathsf{H}(\boldsymbol{\nu}).$$

Esta representación permite, por ejemplo

$$\begin{aligned}\mathsf{E}(\boldsymbol{X}) &= \mathsf{E}(\mathsf{E}(\boldsymbol{X}|W)) = \boldsymbol{\mu}\\ \mathsf{Cov}(\boldsymbol{X}) &= \mathsf{E}(\mathsf{Cov}(\boldsymbol{X}|W)) + \mathsf{Cov}(\mathsf{E}(\boldsymbol{X}|W)) = \mathsf{E}(W^{-1})\boldsymbol{\Sigma},\end{aligned}$$

siempre que $\mathsf{E}(W^{-1}) < \infty$.

EJEMPLO 1.58 (Distribución t de Student). Para $\boldsymbol{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, con $\nu > 0$, podemos escribir

$$\boldsymbol{X}|W \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\omega), \qquad W \sim \mathsf{Gamma}(\nu/2, \nu/2),$$

es decir, la función de densidad asociado a la variable de mezcla, es dada por

$$h(\omega; \nu) = \frac{(\nu/2)^{\nu/2}\omega^{\nu/2-1}}{\Gamma(\nu/2)} \exp(-\nu\omega/2).$$

EJEMPLO 1.59 (Distribución normal contaminada). Considere $\boldsymbol{X} \sim \mathsf{CN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon, \gamma)$ Little (1988) donde $0 \leq \epsilon \leq 1$ denota el *porcentaje de contaminación* y $0 < \gamma < 1$ corresponde a un *factor de inflación de escala*. En este caso, la variable de mezcla tiene densidad

$$h(\omega; \boldsymbol{\delta}) = \begin{cases} \epsilon, & \omega = \gamma \\ 1-\epsilon & \omega = 1 \end{cases},$$

con $\boldsymbol{\delta} = (\epsilon, \gamma)^\top$. Podemos notar que la función de densidad adopta la forma:

$$f(\boldsymbol{x}) = (1-\epsilon)|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp(-u/2) + \epsilon\gamma^{p/2}|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp(-\lambda u/2).$$

1.7. Algunas distribuciones no centrales

Las distribuciones chi-cuadrado, F, t de Student no central son derivadas desde la distribución normal multivariada y son útiles para desarrollar la inferencia en modelo de regresión lineal.

RESULTADO 1.60. *Sea $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$ y sea $U = \boldsymbol{Z}^\top\boldsymbol{Z}$. Entonces $U \sim \chi^2(p)$, con función de densidad*

$$f(u) = \frac{1}{2^{p/2}\Gamma(p/2)} u^{p/2-1} \exp(-u/2), \qquad u > 0.$$

DEMOSTRACIÓN. Como U es una función de variables aleatorias normales, entonces su función característica asume la forma

$$\begin{aligned}\varphi_U(t) = \mathsf{E}\{\exp(itU)\} &= \int_{\mathbb{R}^p} \exp(itu)(2\pi)^{-p/2} \exp(-\tfrac{1}{2}\boldsymbol{z}^\top\boldsymbol{z})\, \mathsf{d}\boldsymbol{z}\\ &= (2\pi)^{-p/2} \int_{\mathbb{R}^p} \exp(-\tfrac{1}{2}(1-2it)\boldsymbol{z}^\top\boldsymbol{z})\, \mathsf{d}\boldsymbol{z} = (1-2it)^{-p/2},\end{aligned}$$

que corresponde a la función característica de una variable aleatoria chi-cuadrado con p grados de libertad. □

DEFINICIÓN 1.61 (Distribución chi-cuadrado no central). Si $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{I})$, entonces $U = \boldsymbol{Y}^\top\boldsymbol{Y}$ tiene *distribución chi-cuadrado no central* con p grados de libertad y parámetro de no centralidad $\lambda = \boldsymbol{\mu}^\top\boldsymbol{\mu}/2$, en cuyo caso anotamos $U \sim \chi^2(p; \lambda)$.

RESULTADO 1.62. *Sea* $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{I})$ *donde* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p) \neq \mathbf{0}$ *y sea* $U = \boldsymbol{Y}^\top \boldsymbol{Y}$. *Entonces la función característica de* U *es dada por*

$$\varphi_U(t) = (1 - 2it)^{-p/2} \exp\Big(\frac{2it\lambda}{1 - 2it}\Big),$$

con $\lambda = \boldsymbol{\mu}^\top \boldsymbol{\mu}/2$.

DEMOSTRACIÓN. Como $Y_1, \ldots, Y_p$ son variables aleatorias independientes, tenemos

$$\begin{aligned}\varphi_U(t) &= \mathsf{E}\Big\{ \exp\Big(t \sum_{j=1}^{n} Y_j^2\Big)\Big\} = \mathsf{E}\Big\{ \prod_{j=1}^{p} \exp(tY_j^2)\Big\} = \prod_{j=1}^{p} \mathsf{E}\{\exp(tY_j^2)\} \\ &= \prod_{j=1}^{p} \varphi_{Y_j^2}(t).\end{aligned}$$

Ahora, la función característica asociada a la variable aleatoria Y_j^2 es dada por

$$\begin{aligned}\varphi_{Y_j^2}(t) &= \int_{-\infty}^{\infty} \exp(ity_j^2)(2\pi)^{-1/2} \exp\{-\tfrac{1}{2}(y_j - \mu_j)^2\}\, \mathsf{d}y_j \\ &= \exp\Big\{\frac{\mu_j^2}{2}\Big(\frac{1}{1-2it}\Big) - \frac{\mu_j^2}{2}\Big\} \int_{-\infty}^{\infty} (2\pi)^{-1/2} \exp\Big\{ -\frac{(1-2it)}{2}\Big(y_j - \frac{\mu_j}{1-2it}\Big)^2\Big\}\, \mathsf{d}y_j,\end{aligned}$$

de este modo,

$$\varphi_{Y_j^2}(t) = (1 - 2it)^{-1/2} \exp\Big\{\frac{\mu_j^2}{2}\Big(\frac{2it}{1 - 2it}\Big)\Big\},$$

y por tanto la función característica de la variable $U = \sum_{j=1}^{p} Y_j^2$, asume la forma

$$\varphi_U(t) = (1 - 2it)^{-p/2} \exp\Big(\frac{2it\lambda}{1 - 2it}\Big), \qquad \lambda = \boldsymbol{\mu}^\top \boldsymbol{\mu}/2.$$

□

OBSERVACIÓN 1.63. Es interesante notar que la función característica de la variable $U = \boldsymbol{Y}^\top \boldsymbol{Y}$, puede ser escrita como

$$\begin{aligned}\varphi_U(t) &= (1 - 2it)^{-p/2} \exp\Big(\frac{\lambda}{1 - 2it} - \lambda\Big) \\ &= (1 - 2it)^{-p/2} e^{-\lambda} \sum_{k=0}^{\infty} \frac{\{\lambda/(1 - 2it)\}^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} (1 - 2it)^{-(p+2k)/2}.\end{aligned}$$

Es decir, la función característica de U es un *promedio ponderado con pesos Poisson* de funciones características de variables aleatorias chi-cuadrado con $p + 2k$ grados de libertad.

Usando la relación entre funciones características y sus correspondientes funciones de densidad, sigue que la chi-cuadrado no central tiene la siguiente representación de mezcla

$$U|Z \sim \chi^2(p + 2z), \qquad Z \sim \mathsf{Poisson}(\lambda), \tag{1.6}$$

con densidad

$$f(u) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \frac{1}{2^{p/2+k}\Gamma(\frac{p}{2} + k)} u^{p/2+k-1} \exp(-u/2), \quad u > 0.$$

La representación en (1.6) es muy útil para obtener los momentos de una variable aleatoria con distribución chi-cuadrado no central. En efecto, el valor esperado de $U \sim \chi^2(p;\lambda)$ es dado por

$$\mathsf{E}(U) = \mathsf{E}\{\mathsf{E}(U|Z)\} = \mathsf{E}\{p+2Z\} = p + 2\,\mathsf{E}(Z) = p + 2\lambda,$$

mientras que la varianza de U puede ser calculada como

$$\begin{aligned}\mathsf{var}(U) &= E\{\mathsf{var}(U|Z)\} + \mathsf{var}\{E(U|Z)\} \\ &= \mathsf{E}\{2(p+2Z)\} + \mathsf{var}(p+2Z) \\ &= 2p + 4\lambda + 4\lambda = 2p + 8\lambda.\end{aligned}$$

Resultado 1.64. *Si* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *donde* $\boldsymbol{\Sigma}$ *es matriz no singular. Entonces*

(a) $(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2(p)$.

(b) $\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2(p;\lambda)$, *donde* $\lambda = \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

Demostración. La idea de la demostración se basa en transformar los componentes de $\boldsymbol{X}$ en variables aleatorias normales independientes. Considere $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$ con $\boldsymbol{B}$ no singular. Para probar (a), tome

$$\boldsymbol{Z} = \boldsymbol{B}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}),$$

luego $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$ y de este modo

$$(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{Z}^\top \boldsymbol{Z} \sim \chi^2(p;0).$$

Para probar (b), sea $\boldsymbol{Y} = \boldsymbol{B}^{-1}\boldsymbol{X}$, luego

$$\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{B}^{-1}\boldsymbol{\mu}, \boldsymbol{I}),$$

y

$$\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X} = \boldsymbol{Y}^\top \boldsymbol{B}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{Y}^\top \boldsymbol{Y},$$

que por definición tiene una distribución chi-cuadrado no central, con parámetro de no centralidad

$$\lambda = \frac{1}{2}(\boldsymbol{B}^{-1}\boldsymbol{\mu})^\top(\boldsymbol{B}^{-1}\boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

□

Definición 1.65 (Distribución F no central). Sea $X_1 \sim \chi^2(\nu_1;\lambda)$ y $X_2 \sim \chi^2(\nu_2)$ variables aleatorias independientes. Entonces,

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} \sim \mathsf{F}(\nu_1, \nu_2, \lambda),$$

es decir F sigue una *distribución F no central* con ν_1 y ν_2 grados de libertad y parámetro de no centralidad λ.

Definición 1.66 (Distribución Beta no central). Considere $U_1 \sim \chi^2(\nu_1, \lambda)$, $U_2 \sim \chi^2(\nu_2)$ tal que U_1 y U_2 son variables aleatorias independientes. Entonces,

$$G = \frac{U_1}{U_1 + U_2} \sim \mathsf{Beta}(\nu_1, \nu_2, \lambda),$$

esto es, G sigue una *distribución Beta no central* con parámetros de forma y escala ν_1 y ν_2, respectivamente y parámetro de no centralidad λ.

DEFINICIÓN 1.67 (Distribución t de Student no central). Si $Y \sim \mathsf{N}(\mu, \sigma^2)$ y $U/\sigma^2 \sim \chi^2(\nu)$ son independientes, entonces

$$T = \frac{Y}{\sqrt{U/\nu}} \sim t_\nu(\lambda), \quad \lambda = \mu/\sigma,$$

es llamada una variable aleatoria con *distribución t de Student no central* con ν grados de libertad y parámetro de no centralidad λ.

Note también que si $Z \sim \mathsf{N}(0,1)$, $U \sim \chi^2(\nu)$, δ es una constante, y Z es independiente de U, entonces

$$T = \frac{Z+\delta}{\sqrt{U/\nu}} \sim t_\nu(\delta).$$

Además el cuadrado de una variable aleatoria t no central se distribuye como una variable aleatoria F no central con parémetro de no centralidad $\delta = \lambda^2/2$. De este modo,

$$t^2_\nu(\lambda) \overset{\mathsf{d}}{=} \mathsf{F}(1, \nu, \lambda^2/2).$$

1.8. Distribución de formas cuadráticas

Para motivar ideas, sabemos que si $\boldsymbol{Z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$, entonces $U = \boldsymbol{Z}^\top \boldsymbol{Z} \sim \chi^2(p)$ pues corresponde a la suma de variables aleatorias $\mathsf{IID}\ \mathsf{N}(0,1)$. El objetivo de esta sección es proveer condiciones bajo las cuales variables aleatorias de la forma $U = \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$ con $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ siguen una distribución chi-cuadrado no central así como establecer la independencia entre dos o más formas cuadráticas.

RESULTADO 1.68. *Si $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{I})$ y $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ es matriz simétrica. Entonces $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} \sim \chi^2(k;\theta)$ si y sólo si $\boldsymbol{A}$ es idempotente, en cuyo caso los grados de libertad y el parámetro de no centralidad están dados por*

$$k = \operatorname{rg}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{A}), \qquad y \qquad \theta = \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu},$$

respectivamente.

DEMOSTRACIÓN. Suponga que $\boldsymbol{A}$ es idempotente de rango k. Entonces existe una matriz ortogonal $\boldsymbol{P}$ tal que

$$\boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P} = \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Sea $\boldsymbol{Y} = \boldsymbol{P}^\top \boldsymbol{X}$, entonces $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{P}^\top \boldsymbol{\mu}, \boldsymbol{I})$, y

$$\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} = \boldsymbol{Y}^\top \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{Y} = \sum_{i=1}^{k} Y_i^2,$$

que sigue una distribución chi-cuadrado con k grados de libertad. Para el parámetro de no centralidad θ, note que

$$\begin{aligned} \mathsf{E}\{\chi^2(k;\theta)\} = k + 2\theta = \mathsf{E}(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}) &= \operatorname{tr}(\mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{A}) \\ &= \operatorname{tr}((\boldsymbol{I} + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{A}) = k + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}, \end{aligned}$$

y de ahí que $\theta = \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$.

Ahora, suponga que $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} \sim \chi^2(k;\theta)$. Si $\boldsymbol{A}$ tiene rango r, entonces para $\boldsymbol{P}$ matriz ortogonal $p \times p$,

$$\boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

con $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$, donde $\lambda_1, \ldots, \lambda_r$ son los valores propios no nulos de $\boldsymbol{A}$. Sea $\boldsymbol{Y} = \boldsymbol{P}^\top \boldsymbol{X}$, entonces

$$\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} = \boldsymbol{Y}^\top \boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P}\boldsymbol{Y} = \sum_{j=1}^{r} \lambda_j Y_j^2 = U.$$

Tenemos que $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{\delta}, \boldsymbol{I})$ con $\boldsymbol{\delta} = \boldsymbol{P}^\top \boldsymbol{\mu}$, de modo que $Y_j^2 \sim \chi^2(1; \delta_j^2/2)$ con función característica

$$\varphi_{Y_j^2}(t) = (1-2it)^{-1/2} \exp\Big(\frac{it\delta_j^2}{1-2it}\Big),$$

por la independencia de $Y_1, \ldots, Y_r$ sigue que

$$\begin{aligned} \varphi_U(t) &= \prod_{j=1}^{r} (1-2it\lambda_j)^{-1/2} \exp\Big(\frac{it\lambda_j\delta_j^2}{1-2it\lambda_j}\Big) \\ &= \exp\Big(it \sum_{j=1}^{r} \frac{\lambda_j \delta_j^2}{1-2it\lambda_j}\Big) \prod_{j=1}^{r} (1-2it\lambda_j)^{-1/2}. \end{aligned}$$

Como $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} \sim \chi_k^2(\theta)$ tiene función característica

$$\varphi_{X^\top AX}(t) = (1-2it)^{-k/2} \exp\Big(\frac{2it\theta}{1-2it}\Big),$$

entonces desde las dos expresiones anteriores debemos tener $r = k$, $\lambda_j = 1$, $\forall j$ y $\theta = \sum_j \delta_j^2/2$. Consecuentemente $\boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P}$ tiene la forma

$$\boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P} = \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

que es idempotente. Luego

$$\boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P} = (\boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P})(\boldsymbol{P}^\top \boldsymbol{A}\boldsymbol{P}) = \boldsymbol{P}^\top \boldsymbol{A}^2 \boldsymbol{P} \quad \Longrightarrow \quad \boldsymbol{A}^2 = \boldsymbol{A}.$$

□

RESULTADO 1.69. *Si $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ donde $\boldsymbol{\Sigma}$ es no singular y $\boldsymbol{X}$, $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son particionados como*

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

donde $\boldsymbol{X}_1$, $\boldsymbol{\mu}_1$ son $k \times 1$ y $\boldsymbol{\Sigma}_{11}$ es $k \times k$. Entonces

$$U = (\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) - (\boldsymbol{X}_1-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_1) \sim \chi^2(p-k).$$

DEMOSTRACIÓN. Considere $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$, donde $\boldsymbol{B}$ es no singular y particione $\boldsymbol{B}$ como

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{pmatrix}, \qquad \boldsymbol{B}_1 \in \mathbb{R}^{k \times p}.$$

Luego,

$$\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top = \begin{pmatrix} \boldsymbol{B}_1\boldsymbol{B}_1^\top & \boldsymbol{B}_1\boldsymbol{B}_2^\top \\ \boldsymbol{B}_2\boldsymbol{B}_1^\top & \boldsymbol{B}_2\boldsymbol{B}_2^\top \end{pmatrix},$$

de donde sigue que $\boldsymbol{\Sigma}_{11} = \boldsymbol{B}_1\boldsymbol{B}_1^\top$. Ahora, sea $\boldsymbol{Z} = \boldsymbol{B}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I})$. De este modo,

$$\begin{pmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{pmatrix} \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{X}_2 - \boldsymbol{\mu}_2 \end{pmatrix}.$$

Entonces

$$\begin{aligned} U &= \boldsymbol{Z}^\top\boldsymbol{Z} - \boldsymbol{Z}^\top\boldsymbol{B}_1^\top(\boldsymbol{B}_1\boldsymbol{B}_1^\top)^{-1}\boldsymbol{B}_1\boldsymbol{Z} = \boldsymbol{Z}^\top(\boldsymbol{I} - \boldsymbol{B}_1^\top(\boldsymbol{B}_1\boldsymbol{B}_1^\top)^{-1}\boldsymbol{B}_1)\boldsymbol{Z} \\ &= \boldsymbol{Z}^\top(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Z}, \qquad \text{con } \boldsymbol{H}_1 = \boldsymbol{B}_1^\top(\boldsymbol{B}_1\boldsymbol{B}_1^\top)^{-1}\boldsymbol{B}_1. \end{aligned}$$

Note que $\boldsymbol{H}_1$ es simétrica e idempotente y por tanto también lo es $\boldsymbol{C} = \boldsymbol{I} - \boldsymbol{H}_1$. De donde sigue que $U \sim \chi^2(\nu)$, con $\nu = \operatorname{rg}(\boldsymbol{C}) = p - k$. □

El Resultado 1.68 se puede generalizar al caso que $\boldsymbol{X}$ tiene una matriz de covarianza arbitraria. Suponga que $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Una condición para que $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}$ tenga una distribución chi-cuadrado es

$$\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{\Sigma}\boldsymbol{A},$$

en cuyo caso los grados de libertad son $k = \operatorname{rg}(\boldsymbol{A}\boldsymbol{\Sigma})$. Si $\boldsymbol{\Sigma}$ es no singular, la condición resulta $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{A}$.

RESULTADO 1.70. *Si $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ donde $\boldsymbol{\Sigma}$ tiene rango k ($\leq p$) y si $\boldsymbol{A}$ es una inversa generalizada de $\boldsymbol{\Sigma}$ ($\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$), entonces $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} \sim \chi^2(k)$.*

DEMOSTRACIÓN. Considere $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$ donde $\boldsymbol{B}$ es una matriz no singular $p \times p$ tal que

$$\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top = \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Particionando $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)^\top$ donde $\boldsymbol{Y}_1$ es un vector $k \times 1$ sigue que $\boldsymbol{Y}_1 \sim \mathsf{N}_k(\boldsymbol{0}, \boldsymbol{I})$ y $\boldsymbol{Y}_2 = \boldsymbol{0}$ con probabilidad 1. Es decir, tenemos que

$$\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{0})^\top, \quad \text{con probabilidad 1.}$$

Ahora, note que

$$\begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^\top$$

pues $\boldsymbol{A}$ es una inversa generalizada de $\boldsymbol{\Sigma}$. De este modo,

$$\begin{aligned} \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} &= \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top \\ &= \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1} \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}. \end{aligned}$$

Luego, con probabilidad uno,

$$\begin{aligned} \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} = \boldsymbol{Y}^\top\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{Y} &= (\boldsymbol{Y}_1^\top, \boldsymbol{0})\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1}\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{0} \end{pmatrix} \\ &= (\boldsymbol{Y}_1^\top, \boldsymbol{0}) \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1} \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{0} \end{pmatrix} \\ &= (\boldsymbol{Y}_1^\top, \boldsymbol{0}) \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{Y}_1^\top\boldsymbol{Y}_1 \sim \chi^2(k). \end{aligned}$$

□

RESULTADO 1.71. *Si* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *donde* $\boldsymbol{\Sigma}$ *es no singular, y* $\boldsymbol{A}$ *es una matriz simétrica* $p \times p$. *Entonces* $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} \sim \chi^2(k; \lambda)$, *donde* $k = \mathrm{rg}(\boldsymbol{A})$, $\lambda = \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}/2$ *si y sólo si* $\boldsymbol{A}\boldsymbol{\Sigma}$ *es matriz idempotente.*

DEMOSTRACIÓN. Considere $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$, donde $\boldsymbol{B}$ es una matriz no singular $p \times p$ tal que $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top = \boldsymbol{I}_p$. Entonces

$$\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} = \boldsymbol{Y}^\top \boldsymbol{B}^{-\top} \boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{Y},$$

donde $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{I})$. Desde el Resultado 1.68 sigue que $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$ tiene distribución chi-cuadrado sólo si $\boldsymbol{B}^{-\top} \boldsymbol{A}\boldsymbol{B}^{-1}$ es idempotente. Esto es equivalente a mostrar que $\boldsymbol{A}\boldsymbol{\Sigma}$ es idempotente.

Si $\boldsymbol{A}\boldsymbol{\Sigma}$ es idempotente, tenemos

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{B}^{-\top}\boldsymbol{A}, \qquad (\boldsymbol{\Sigma} = \boldsymbol{B}^{-1}\boldsymbol{B}^{-\top})$$

así, pre- y post-multiplicando por $\boldsymbol{B}^{-\top}$ y $\boldsymbol{B}^{-1}$, obtenemos

$$\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1} = (\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1})(\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1}),$$

y por tanto es idempotente.

Por otro lado, si $\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1}$ es idempotente, entonces

$$\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1} = (\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1})(\boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{B}^{-1}) = \boldsymbol{B}^{-\top}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{B}^{-1},$$

es decir $\boldsymbol{A} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}$ y de ahí que $\boldsymbol{A}\boldsymbol{\Sigma}$ es idempotente. □

EJEMPLO 1.72. Sea $X_1, \ldots, X_n$ variables aleatorias $\mathsf{IID}\ \mathsf{N}(\theta, \sigma^2)$, en este caso podemos definir $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ tal que $\boldsymbol{X} \sim \mathsf{N}_n(\theta\mathbf{1}_n, \sigma^2\boldsymbol{I})$. Considere la forma cuadrática

$$Q = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X},$$

con $\boldsymbol{C} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ y $\boldsymbol{A} = \boldsymbol{C}/\sigma^2$. De esta manera

$$\boldsymbol{A}\boldsymbol{\Sigma} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top,$$

que es idempotente. Además

$$\mathrm{rg}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{C}) = \mathrm{tr}\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) = n - 1,$$

y

$$\lambda = \frac{\theta^2}{2}\mathbf{1}^\top\boldsymbol{A}\mathbf{1} = \frac{\theta^2}{2\sigma^2}\mathbf{1}^\top\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{1} = 0.$$

Finalmente,

$$Q = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1).$$

RESULTADO 1.73. *Sea* $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $Q_1 = \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}$ *y* $Q_2 = \boldsymbol{X}^\top\boldsymbol{B}\boldsymbol{X}$. *Entonces* Q_1 *y* Q_2 *son independientes si y sólo si* $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \mathbf{0}$.

DEMOSTRACIÓN. Tenemos $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{T}^\top$, y defina $\boldsymbol{G}_1 = \boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{T}$, $\boldsymbol{G}_2 = \boldsymbol{T}^\top\boldsymbol{B}\boldsymbol{T}$. Note que si $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \mathbf{0}$, entonces

$$\boldsymbol{G}_1\boldsymbol{G}_2 = (\boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{T})(\boldsymbol{T}^\top\boldsymbol{B}\boldsymbol{T}) = \boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{T} = \mathbf{0}.$$

Debido a la simetría de $\boldsymbol{G}_1$ y $\boldsymbol{G}_2$, sigue que

$$\boldsymbol{0} = (\boldsymbol{G}_1\boldsymbol{G}_2)^\top = \boldsymbol{G}_2^\top\boldsymbol{G}_1^\top = \boldsymbol{G}_2\boldsymbol{G}_1.$$

Como $\boldsymbol{G}_1\boldsymbol{G}_2 = \boldsymbol{G}_2\boldsymbol{G}_1$ existe una matriz ortogonal $\boldsymbol{P}$ que simultáneamente diagonaliza $\boldsymbol{G}_1$ y $\boldsymbol{G}_2$, esto es:

$$\boldsymbol{P}^\top\boldsymbol{G}_1\boldsymbol{P} = \boldsymbol{P}^\top\boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{T}\boldsymbol{P} = \boldsymbol{D}_1,$$
$$\boldsymbol{P}^\top\boldsymbol{G}_2\boldsymbol{P} = \boldsymbol{P}^\top\boldsymbol{T}^\top\boldsymbol{B}\boldsymbol{T}\boldsymbol{P} = \boldsymbol{D}_2.$$

De este modo,

$$\boldsymbol{0} = \boldsymbol{G}_1\boldsymbol{G}_2 = \boldsymbol{P}\boldsymbol{D}_1\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{D}_2\boldsymbol{P}^\top = \boldsymbol{P}\boldsymbol{D}_1\boldsymbol{D}_2\boldsymbol{P}^\top$$

lo que es verdad si $\boldsymbol{D}_1\boldsymbol{D}_2 = \boldsymbol{0}$. Como $\boldsymbol{D}_1$ y $\boldsymbol{D}_2$ son diagonales, sus elementos diagonales deben ocurrir en posiciones diferentes. Es decir, podemos escribir

$$\boldsymbol{D}_1 = \begin{pmatrix} \boldsymbol{M}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{D}_2 = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_2 \end{pmatrix}.$$

Sea $\boldsymbol{Y} = \boldsymbol{P}^\top\boldsymbol{T}^{-1}\boldsymbol{X}$, entonces

$$Q_1 = \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{T}^{-\top}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{T}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{T}^{-1}\boldsymbol{X} = \boldsymbol{Y}^\top\boldsymbol{D}_1\boldsymbol{Y},$$
$$Q_2 = \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{T}^{-\top}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{T}^\top\boldsymbol{B}\boldsymbol{T}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{T}^{-1}\boldsymbol{X} = \boldsymbol{Y}^\top\boldsymbol{D}_2\boldsymbol{Y}.$$

Además,

$$\mathsf{Cov}(\boldsymbol{Y}) = \mathsf{Cov}(\boldsymbol{P}^\top\boldsymbol{T}^{-1}\boldsymbol{X}) = \boldsymbol{P}^\top\boldsymbol{T}^{-1}\,\mathsf{Cov}(\boldsymbol{X})\boldsymbol{T}^{-\top}\boldsymbol{P} = \boldsymbol{I}.$$

En efecto, $\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{P}^\top\boldsymbol{T}^{-1}\boldsymbol{\mu}, \boldsymbol{I})$. Ahora, particionando adecuadamente $\boldsymbol{Y}$, sigue que

$$\boldsymbol{Y}^\top\boldsymbol{D}_1\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)\begin{pmatrix} \boldsymbol{M}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} = \boldsymbol{Y}_1^\top\boldsymbol{M}_1\boldsymbol{Y}_1,$$
$$\boldsymbol{Y}^\top\boldsymbol{D}_2\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top)\begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_2 \end{pmatrix}\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} = \boldsymbol{Y}_2^\top\boldsymbol{M}_2\boldsymbol{Y}_2,$$

y la independencia entre Q_1 y Q_2 sigue desde la independencia entre $\boldsymbol{Y}_1$ y $\boldsymbol{Y}_2$. □

RESULTADO 1.74. *Sea $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $Q = \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}$ y $\boldsymbol{U} = \boldsymbol{B}\boldsymbol{X}$. Entonces Q y $\boldsymbol{U}$ son independientes si y sólo si $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$.*

EJEMPLO 1.75. Considere $X_1, \ldots, X_n$ muestra aleatoria desde $\mathsf{N}(\theta, \sigma^2)$, así

$$\boldsymbol{X} = (X_1, \ldots, X_n)^\top \sim \mathsf{N}_n(\theta\boldsymbol{1}, \sigma^2\boldsymbol{I}_n).$$

Tenemos

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\boldsymbol{1}^\top\boldsymbol{X}, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}.$$

Como $\boldsymbol{C}\boldsymbol{1} = \boldsymbol{0}$ sigue la independencia entre $\overline{X}$ y S^2.

Considere los siguientes dos lemas, los que permitirán mostrar el resultado principal de esta sección.

LEMA 1.76. *Sean $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$ matrices $m \times m$ simétricas e idempotentes y suponga que*

$$\boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_k = \boldsymbol{I}_m.$$

Entonces $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{0}$ para todo $i \neq j$.

DEMOSTRACIÓN. Considere cualquiera de esas matrices, digamos $\boldsymbol{A}_h$ y denote su rango por r. Como $\boldsymbol{A}_h$ es simétrica e idempotente, existe una matriz ortogonal $\boldsymbol{P}$ tal que

$$\boldsymbol{P}^\top \boldsymbol{A}_h \boldsymbol{P} = \begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Para $j \neq h$, defina $\boldsymbol{B}_j = \boldsymbol{P}^\top \boldsymbol{A}_j \boldsymbol{P}$, y note que

$$\boldsymbol{I}_m = \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{P}^\top \Big(\sum_{j=1}^{k} \boldsymbol{A}_j \Big) \boldsymbol{P} = \sum_{j=1}^{k} \boldsymbol{P}^\top \boldsymbol{A}_j \boldsymbol{P} = \begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} + \sum_{j \neq h} \boldsymbol{B}_j.$$

O equivalentemente,

$$\sum_{j \neq h} \boldsymbol{B}_j = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m-r} \end{pmatrix}.$$

Claramente, dado que $\boldsymbol{A}_j$ es simétrica e idempotente, sigue que $\boldsymbol{B}_j$ también lo es. De modo que, sus elementos diagonales son no negativos. Además, $(\boldsymbol{B}_j)_{ll} = 0$, para $l = 1, \ldots, r$. Así, sigue que $\boldsymbol{B}_j$ debe ser de la forma

$$\boldsymbol{B}_j = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_j \end{pmatrix},$$

donde $\boldsymbol{C}_j$ es matriz $(m-r)\times(m-r)$, simétrica e idempotente. Ahora, para cualquier $j \neq h$

$$\boldsymbol{P}^\top \boldsymbol{A}_h \boldsymbol{A}_j \boldsymbol{P} = (\boldsymbol{P}^\top \boldsymbol{A}_h \boldsymbol{P})(\boldsymbol{P}^\top \boldsymbol{A}_j \boldsymbol{P}) = \begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_j \end{pmatrix} = \boldsymbol{0},$$

lo que es verdad, sólo si $\boldsymbol{A}_h \boldsymbol{A}_j = \boldsymbol{0}$, pues $\boldsymbol{P}$ es no singular. Notando que h es arbitrareo, la prueba es completa. □

LEMA 1.77. *Sean $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$ matrices simétricas de orden $m \times m$ y defina*

$$\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_k.$$

Considere las siguientes afirmaciones,

(a) $\boldsymbol{A}_i$ es idempotente, para $i = 1, \ldots, k$.
(b) $\boldsymbol{A}$ es idempotente.
(c) $\boldsymbol{A}_i \boldsymbol{A}_j = \boldsymbol{0}$, para $i \neq j$.

Entonces, si dos condiciones son satisfechas, la tercera condición debe ser verdadera.

DEMOSTRACIÓN. Primero mostraremos que (a) y (b) implica (c). Como $\boldsymbol{A}$ es simétrica e idempotente, existe una matriz ortogonal $\boldsymbol{P}$ tal que

$$\boldsymbol{P}^\top \boldsymbol{A} \boldsymbol{P} = \boldsymbol{P}^\top (\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_k) \boldsymbol{P} = \begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \tag{1.7}$$

donde $r = \operatorname{rg}(\boldsymbol{A})$.

Sea $\boldsymbol{B}_i = \boldsymbol{P}^\top \boldsymbol{A}_i \boldsymbol{P}$, para $i = 1, \ldots, k$, y note que $\boldsymbol{B}_i$ es simétrica e idempotente. Es decir, $\boldsymbol{B}_i$ debe ser de la forma

$$\boldsymbol{B}_i = \begin{pmatrix} \boldsymbol{C}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

donde la matriz $r \times r$, $\boldsymbol{C}_i$ debe ser simétrica e idempotente. Por (1.7), tenemos

$$\boldsymbol{C}_1 + \cdots + \boldsymbol{C}_k = \boldsymbol{I}_r.$$

Por el Lema 1.76, sigue que $\boldsymbol{C}_i\boldsymbol{C}_j = \boldsymbol{0}$ para $i \neq j$, de donde obtenemos $\boldsymbol{B}_i\boldsymbol{B}_j = \boldsymbol{0}$ y de ahí que $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{0}$, para $i \neq j$.
Que (a) y (c) implican (b), sigue de notar

$$\boldsymbol{A}^2 = \Big(\sum_{i=1}^{k} \boldsymbol{A}_i\Big)^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} \boldsymbol{A}_i\boldsymbol{A}_j = \sum_{i=1}^{k} \boldsymbol{A}_i^2 + \sum\sum_{i\neq j} \boldsymbol{A}_i\boldsymbol{A}_j$$
$$= \sum_{i=1}^{k} \boldsymbol{A}_i = \boldsymbol{A}.$$

Finalmente, para probar que (b) y (c) implican (a). Suponga que (c) es verdad, entonces $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{A}_j\boldsymbol{A}_i$ para todo $i \neq j$ y las matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$ pueden ser diagonalizadas simultáneamente. Esto es, existe una matriz ortogonal $\boldsymbol{Q}$ tal que

$$\boldsymbol{Q}^\top \boldsymbol{A}_i \boldsymbol{Q} = \boldsymbol{D}_i, \qquad i = 1, \ldots, k,$$

donde cada una de las matrices $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_k$ es diagonal. Además,

$$\boldsymbol{D}_i\boldsymbol{D}_j = \boldsymbol{Q}^\top \boldsymbol{A}_i \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{A}_j \boldsymbol{Q} = \boldsymbol{Q}^\top \boldsymbol{A}_i\boldsymbol{A}_j \boldsymbol{Q} = \boldsymbol{0}, \qquad i \neq j. \tag{1.8}$$

Como $\boldsymbol{A}$ es simétrica e idempotente, también lo es la matriz diagonal

$$\boldsymbol{Q}^\top \boldsymbol{A}\boldsymbol{Q} = \boldsymbol{D}_1 + \cdots \boldsymbol{D}_k,$$

y cada elemento diagonal de $\boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q}$ debe ser 0 o 1, y por (1.8), lo mismo es válido para los elementos diagonales de $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_k$.
De este modo, $\boldsymbol{D}_i$ es simétrica e idempotente y de ahí que también lo es

$$\boldsymbol{A}_i = \boldsymbol{Q}\boldsymbol{D}_i\boldsymbol{Q}^\top, \qquad i = 1, \ldots, k,$$

lo que termina la prueba. □

Observación 1.78. Suponga que las condiciones del Lema 1.77 son satisfechas. Entonces (a) implica que $\operatorname{rg}(\boldsymbol{A}_i) = \operatorname{tr}(\boldsymbol{A}_i)$, y desde (b), sigue que

$$\operatorname{rg}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{A}) = \operatorname{tr}\Big(\sum_{i=1}^{k} \boldsymbol{A}_i\Big) = \sum_{i=1}^{k} \operatorname{tr}(\boldsymbol{A}_i) = \sum_{i=1}^{k} \operatorname{rg}(\boldsymbol{A}_i).$$

Resultado 1.79 (Teorema de Cochran). *Sea $\boldsymbol{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma} > \boldsymbol{0}$. Suponga que $\boldsymbol{A}_i$, es una matriz simétrica de orden $p \times p$ con rango r_i, para $i = 1, \ldots, k$, y*

$$\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_k,$$

es de rango r. Considere las condiciones:

- *(a) $\boldsymbol{A}_i\boldsymbol{\Sigma}$ es idempotente, para $i = 1, \ldots, k$.*
- *(b) $\boldsymbol{A}\boldsymbol{\Sigma}$ es idempotente.*
- *(c) $\boldsymbol{A}_i\boldsymbol{\Sigma}\boldsymbol{A}_j = \boldsymbol{0}$, para $i \neq j$.*
- *(d) $r = \sum_{i=1}^{k} r_i$.*

si dos de (a), (b) y (c) se satisfacen, o si (b) y (d) son satisfechas. Entonces,

- *(i) $\boldsymbol{X}^\top \boldsymbol{A}_i \boldsymbol{X} \sim \chi^2(r_i; \lambda_i)$, con $\lambda_i = \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}/2$, para $i = 1, \ldots, k$.*
- *(ii) $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} \sim \chi^2(r; \lambda)$, con $\lambda = \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}/2$.*
- *(iii) $\boldsymbol{X}^\top \boldsymbol{A}_1\boldsymbol{X}, \boldsymbol{X}^\top \boldsymbol{A}_2\boldsymbol{X}, \ldots, \boldsymbol{X}^\top \boldsymbol{A}_k\boldsymbol{X}$ son mutuamente independientes.*

DEMOSTRACIÓN. Tenemos que $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{T}^\top$ y las condiciones (a)-(d), pueden ser expresadas como:

(a) $\boldsymbol{T}^\top \boldsymbol{A}_i \boldsymbol{T}$ es idempotente, para $i = 1, \ldots, k$.
(b) $\boldsymbol{T}^\top \boldsymbol{A} \boldsymbol{T}$ es idempotente.
(c) $(\boldsymbol{T}^\top \boldsymbol{A}_i \boldsymbol{T})(\boldsymbol{T}^\top \boldsymbol{\Sigma} \boldsymbol{A}_j \boldsymbol{T}) = \boldsymbol{0}$, para $i \neq j$.
(d) $\mathrm{rg}(\boldsymbol{T}^\top \boldsymbol{A} \boldsymbol{T}) = \sum_{i=1}^{k} \mathrm{rg}(\boldsymbol{T}^\top \boldsymbol{A}_i \boldsymbol{T})$.

Como $\boldsymbol{T}^\top \boldsymbol{A}_1 \boldsymbol{T}, \boldsymbol{T}^\top \boldsymbol{A}_2 \boldsymbol{T}, \ldots, \boldsymbol{T}^\top \boldsymbol{A}_k \boldsymbol{T}$ y $\boldsymbol{T}^\top \boldsymbol{A} \boldsymbol{T}$ satisfacen las condiciones del Lema 1.77 (y de la Observación 1.78). Entonces, las condiciones (a)-(d) se satisfacen.

Sabemos que (a) implica (i) y (b) implica (ii). Mientras que, Resultado 1.73 con (c), garantiza (iii), lo que completa la prueba. □

Ejercicios

1.1 Sean $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ vectores aleatorios independientes con $\boldsymbol{X}_i \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, para $i = 1, \ldots, n$. Obtenga la distribución de

$$\sum_{i=1}^{n} \alpha_i \boldsymbol{X}_i,$$

con $\alpha_1, \ldots, \alpha_n$ constantes fijas.

1.2 Si $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ son independientes cada uno con $\boldsymbol{X}_i \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Muestre que la distribución del vector de medias

$$\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i,$$

es $\mathsf{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.

1.3 Demuestre el Resultado 1.20, usando la función característica de un vector aleatorio normal.

1.4 Sean $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuídas $\mathsf{N}(\mu, \sigma^2)$ y defina

$$Q = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2,$$

¿Es Q un estimador insesgado de σ^2?

1.5 Sea $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y defina

$$\boldsymbol{Y} = \boldsymbol{T}^\top \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{X} - \boldsymbol{\mu}), \qquad \boldsymbol{u} = \boldsymbol{T}^\top \boldsymbol{\Sigma}^{1/2} \boldsymbol{A} \boldsymbol{\mu}.$$

con $\boldsymbol{T}$ ortogonal y $\boldsymbol{A} = \boldsymbol{A}^\top$. Obtenga la distribución de $\boldsymbol{Y}$ y calcule $\mathsf{var}(\boldsymbol{u}^\top \boldsymbol{Y})$.

1.6 Considere $\boldsymbol{Z}$ matriz aleatoria $n \times p$ con función característica

$$\varphi_Z(\boldsymbol{T}) = \mathsf{E}\{\exp(i\,\mathrm{tr}(\boldsymbol{T}^\top \boldsymbol{Z}))\} = \exp\{-\tfrac{1}{2}\,\mathrm{tr}(\boldsymbol{T}^\top \boldsymbol{T})\}.$$

con $\boldsymbol{T} \in \mathbb{R}^{n\times p}$. Obtenga la función característica de

$$\boldsymbol{Y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\boldsymbol{\Theta}^{1/2} + \boldsymbol{\mu},$$

donde $\boldsymbol{\mu} \in \mathbb{R}^{n\times p}$ y $\boldsymbol{\Sigma}$, $\boldsymbol{\Theta}$ son matrices semidefinidas positivas $n \times n$ y $p \times p$, respectivamente.

1.7 Sea $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$ con $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ tal que $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{D}$ es matriz diagonal $p \times p$ y $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Considere

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{Z}.$$

donde λ es un escalar positivo.
(a) Obtenga la distribución de $\widehat{\boldsymbol{\alpha}}$,
(b) Muestre que

$$\boldsymbol{\alpha} - \mathsf{E}(\widehat{\boldsymbol{\alpha}}) = \lambda(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{\alpha}.$$

1.8 Sea $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ y considere $\boldsymbol{b} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$, $\boldsymbol{u} = (\boldsymbol{D}^{-1} + \boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$, donde $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$ y $\boldsymbol{D}$ es matriz no singular $q \times q$.
(a) Halle la distribución de $\boldsymbol{b}$ y $\boldsymbol{u}$,
(b) ¿Son $\boldsymbol{b}$ y $\boldsymbol{u}$ independientes?

1.9 Considere

$$\begin{pmatrix}\boldsymbol{Y}\\ \boldsymbol{b}\end{pmatrix} \sim \mathsf{N}_{n+q}\left(\begin{pmatrix}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\top + \boldsymbol{R} & \boldsymbol{Z}\boldsymbol{D}\\ \boldsymbol{D}\boldsymbol{Z}^\top & \boldsymbol{D}\end{pmatrix}\right),$$

donde $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$ y $\boldsymbol{R}$, $\boldsymbol{D}$ son matrices no singulares $n \times n$ y $q \times q$, respectivamente. Determine la distribución de $\boldsymbol{b}|\boldsymbol{Y}$.

1.10 Sea $U_i \sim \chi^2(n_i; \lambda_i)$, $i = 1, \ldots, K$ variables aleatorias independientes. Muestre que

$$U = \sum_{i=1}^{K} U_i \sim \chi^2(n; \lambda),$$

donde $n = \sum_{i=1}^K n_i$ y $\lambda = \sum_{i=1}^K \lambda_i$.

1.11 Sea $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$, donde $\boldsymbol{X}$ es matriz $n \times p$ con $\mathrm{rg}(\boldsymbol{X}) = p$ y $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. Defina

$$Q = \frac{(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})^\top[\boldsymbol{G}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{G}^\top]^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})}{\sigma^2},$$

donde $\boldsymbol{G} \in \mathbb{R}^{m\times p}$ con $\mathrm{rg}(\boldsymbol{G}) = m$ y $\boldsymbol{g}$ es vector m-dimensional. Determine la distribución de Q.

1.12 Sea $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ y considere las formas cuadráticas

$$Q_1 = \frac{\widehat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{\sigma^2}, \qquad Q_2 = \frac{(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{\sigma^2},$$

donde $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ con $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ y $\mathrm{rg}(\boldsymbol{X}) = p$.
(a) Halle la distribución de Q_i, $i = 1, 2$.
(b) Sea $Q = Q_1 + Q_2$, mostrar la independencia conjunta de Q_1 y Q_2.

1.13 Considere $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ con $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ y $\boldsymbol{\beta} \in \mathbb{R}^p$ y sea $Q = Q_1 + Q_2$, donde

$$Q_1 = \frac{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}{\sigma^2}, \qquad Q_2 = \frac{\boldsymbol{Y}^\top(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}}{\sigma^2},$$

con $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Muestre que Q_1 y Q_2 tienen distribuciones chi-cuadrado independientes.

1.14 Considere

$$\boldsymbol{Y} = (\boldsymbol{I}_p \otimes \boldsymbol{1}_n)\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

donde $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_p^\top)^\top$ con $\boldsymbol{Y}_i$ vector n-dimensional, para $i = 1, \ldots, n$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ y $\boldsymbol{\epsilon} \sim \mathsf{N}_{np}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{np})$. Sean

$$Q_1 = \frac{\boldsymbol{Y}^\top(\boldsymbol{I}_p \otimes \frac{1}{n}\boldsymbol{J}_n)\boldsymbol{Y}}{\sigma^2}, \quad \text{y} \quad Q_2 = \frac{\boldsymbol{Y}^\top(\boldsymbol{I}_p \otimes \boldsymbol{C})\boldsymbol{Y}}{\sigma^2},$$

donde $\boldsymbol{J}_n = \boldsymbol{1}_n\boldsymbol{1}_n^\top$ y $\boldsymbol{C} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n$.

(a) Halle la distribución de Q_k, $k = 1, 2$.

(b) ¿Son Q_1 y Q_2 independientes?

Capítulo 2

Inferencia en el Modelo Lineal

En este capítulo se describe la inferencia en modelos lineales. Primeramente introducimos algunas definiciones y supuestos en los que se basan los modelos de regresión.

2.1. Definición de un modelo lineal

DEFINICIÓN 2.1. Considere la variable aleatoria Y, decimos que sigue un *modelo lineal* si

$$\mathsf{E}(Y) = \sum_{i=1}^{p} x_j \beta_j = \boldsymbol{x}^\top \boldsymbol{\beta},$$

donde $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ representa un vector de p *variables regresoras*, mientras que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ denota un vector de parámetros desconocidos, conocidos como *coeficientes de regresión*.

Suponga que tenemos n observaciones recolectadas desde Y, entonces podemos considerar el modelo

$$\mathsf{E}(Y_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \qquad i = 1, \ldots, n.$$

Es conveniente escribir lo anterior como:

$$\mathsf{E}(\boldsymbol{Y}) = \begin{pmatrix} \boldsymbol{x}_1^\top \boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n^\top \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} \boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta},$$

donde $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{n \times p}$ se denomina *matriz de diseño*.

OBSERVACIÓN 2.2. Cuando

$$\mathsf{Cov}(\boldsymbol{Y}) = \sum_{r=1}^{K} \phi_r \boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}(\boldsymbol{\phi}),$$

donde $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)^\top$ son parámetros desconocidos y $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ son matrices (simétricas) conocidas, decimos que $\boldsymbol{Y}$ sigue un *modelo lineal general*.

Típicamente, asumiremos que $Y_1, \ldots, Y_n$ son independientes con varianza constante, en cuyo caso

$$\mathsf{Cov}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I},$$

y decimos que $\boldsymbol{Y}$ sigue un *modelo lineal simple*.

DEFINICIÓN 2.3. Se dice que el vector $\boldsymbol{Y}$ sigue un *modelo lineal* si

$$\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}, \qquad \mathsf{Cov}(\boldsymbol{Y}) = \sum_{r=1}^{K} \phi_r \boldsymbol{\Sigma}_r.$$

Note que la definición anterior puede ser expresada de forma equivalente como:

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

con

$$\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \qquad \mathsf{Cov}(\boldsymbol{\epsilon}) = \sum_{r=1}^{K} \phi_r \boldsymbol{\Sigma}_r.$$

Observación 2.4. Los supuestos de momentos dados en la Definición 2.3 suelen ser llamados *condiciones de Gauss-Markov*. Aunque es usual que sean expresados en términos del modelo lineal simple, como

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

con

$$\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \qquad \mathsf{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Definición 2.5. Se dice que el vector $\boldsymbol{Y}$ sigue un *modelo lineal normal* si

$$\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\phi})),$$

o bien

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\phi})).$$

Mientras que sigue un *modelo normal simple*, si

$$\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}). \tag{2.1}$$

Observación 2.6. Se debe destacar que el modelo en (2.1) puede ser escrito como

$$Y_i \overset{\mathsf{ind}}{\sim} \mathsf{N}_1(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2), \qquad i = 1, \ldots, n. \tag{2.2}$$

En efecto, la inferencia estadística para los modelos definidos en Ecuaciones (2.1) y (2.2) son equivalentes.[1]

Supuesto 1. *El modelo lineal descrito por la ecuación:*

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

tiene los siguientes supuestos:

- A1: $\boldsymbol{X}$ *es una matriz no aleatoria* $n \times p$ *con* $n > p$.
- A2: *La matriz* $\boldsymbol{X}$ *tiene rango* p*, es decir,* $\boldsymbol{X}$ *es rango columna completo.*
- A3: *El vector aleatorio* n*-dimensional* $\boldsymbol{Y}$ *tiene elementos que son observables.*
- A4: *El vector aleatorio no observable* $\boldsymbol{\epsilon}$ *satisface*

$$\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \qquad \mathsf{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}, \quad \sigma^2 > 0.$$

El Supuesto A4 puede ser re-establecido para incorporar la suposición de normalidad, esto es,

Supuesto 2. *Considere:*

- A4*: *El vector aleatorio* $\boldsymbol{\epsilon}$ *satisface* $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$*, o equivalentemente*

$$\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n).$$

[1] Es decir, **bajo normalidad** estos dos modelos son equivalentes. En efecto, esto **no** es verdad en general.

2.2. Estimación de parámetros en el modelo de regresión lineal

Primeramente abordaremos la estimación máximo verosímil bajo normalidad, revisaremos propiedades de los estimadores y abordaremos la conexión con el método de mínimos cuadrados. En efecto, es fácil notar que la función de verosimilitud proveniente del modelo $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$, adopta la forma:

$$L(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \Big),$$

con $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$. De este modo, la función de log-verosimilitud es dada por

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \\ &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} Q(\boldsymbol{\beta}), \end{aligned}$$

donde

$$Q(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2,$$

se denomina *suma de cuadrados del error*. Diferenciando con relación a $\boldsymbol{\beta}$ y σ^2, obtenemos

$$\mathsf{d}_\beta\, \ell(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\, \mathsf{d}_\beta\, \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{X}\, \mathsf{d}\boldsymbol{\beta},$$

y

$$\mathsf{d}_{\sigma^2}\, \ell(\boldsymbol{\theta}) = -\frac{n}{2\sigma^2}\, \mathsf{d}\sigma^2 + \frac{1}{2\sigma^4}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\, \mathsf{d}\sigma^2.$$

Es decir,[2]

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2. \end{aligned}$$

Desde la condición de primer orden, tenemos las *ecuaciones de verosimilitud*:

$$\begin{aligned} \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) &= \boldsymbol{0} \\ n\widehat{\sigma}^2 - \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 &= 0. \end{aligned}$$

Resolviendo con relación a $\boldsymbol{\beta}$ obtenemos las *ecuaciones normales*

$$\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{Y},$$

dado que $\mathrm{rg}(\boldsymbol{X}) = \mathrm{rg}(\boldsymbol{X}^\top\boldsymbol{X}) = p$, entonces el sistema anterior admite solución única dada por:[3]

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}, \\ \widehat{\sigma}^2 &= \frac{1}{n}Q(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2. \end{aligned}$$

Se define el vector de *valores predichos* como

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

donde $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

---

[2] Usando el primer Teorema de identificación dado en Magnus y Neudecker (2007), pag. 98.

[3] Note que $\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^+\boldsymbol{Y}$ con $\boldsymbol{X}^+ = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ la inversa Moore-Penrose de $\boldsymbol{X}$.

Es fácil notar que $\boldsymbol{H}$ es simétrica e idempotente, en cuyo caso

$$\mathrm{rg}(\boldsymbol{H}) = \mathrm{tr}(\boldsymbol{H}) = \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \mathrm{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}) = p.$$

El vector de diferencias entre $\boldsymbol{Y}$ y $\widehat{\boldsymbol{Y}}$ se denomina el *vector de residuos*, es decir

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

Además, tenemos que $\boldsymbol{I}-\boldsymbol{H}$ también es simétrica e idempotente. Con esta notación podemos escribir

$$\begin{aligned} Q(\widehat{\boldsymbol{\beta}}) = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 &= \boldsymbol{e}^\top\boldsymbol{e} = \boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \\ &= \sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})^2, \end{aligned}$$

que es conocido como *suma de cuadrados residual*.

Resultado 2.7. *Considere el modelo:*

$$\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n),$$

con los supuestos A1 a A4⋆. Entonces, tenemos que:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \sim \mathsf{N}_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}), \tag{2.3}$$

de donde sigue que $\widehat{\boldsymbol{\beta}}$ es insesgado y $\mathsf{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. Además, la variable aleatoria

$$\frac{Q(\widehat{\boldsymbol{\beta}})}{\sigma^2} = \frac{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}{\sigma^2} \sim \chi^2(n-p).$$

Finalmente $\widehat{\boldsymbol{\beta}}$ y $s^2 = Q(\widehat{\boldsymbol{\beta}})/(n-p)$ son independientes.

Demostración. Notando que $\widehat{\boldsymbol{\beta}}$ es una transformación lineal de un vector aleatorio normal y

$$\begin{aligned} \mathsf{E}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\,\mathsf{E}(\boldsymbol{Y}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}, \\ \mathsf{Cov}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\,\mathsf{Cov}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \end{aligned}$$

y el resultado en Ecuación (2.3) sigue. Por otro lado,

$$\frac{Q(\widehat{\boldsymbol{\beta}})}{\sigma^2} = \frac{(n-p)s^2}{\sigma^2} = \frac{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}{\sigma^2},$$

sigue una distribución chi-cuadrado pues $\boldsymbol{I} - \boldsymbol{H}$ es matriz idempotente con

$$\mathrm{rg}(\boldsymbol{I} - \boldsymbol{H}) = \mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}) = \mathrm{tr}(\boldsymbol{I}) - \mathrm{tr}(\boldsymbol{H}) = n - p.$$

Además, el parámetro de no centralidad es dado por

$$\lambda = \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = 0.$$

En efecto,

$$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X} \quad \Rightarrow \quad (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}.$$

La independencia entre $\widehat{\boldsymbol{\beta}}$ y s^2 sigue desde $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$, lo que concluye la prueba. □

OBSERVACIÓN 2.8. Es fácil notar que

$$\begin{aligned}\mathsf{E}\{Q(\widehat{\boldsymbol{\beta}})\} &= \mathsf{E}\{\boldsymbol{Y}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}\} = \sigma^2\,\mathrm{tr}(\boldsymbol{I}-\boldsymbol{H}) + \boldsymbol{\beta}^\top\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta}\\ &= \sigma^2(n-p),\end{aligned}$$

es decir,

$$\mathsf{E}(\widehat{\sigma}^2) = \Big(\frac{n-p}{n}\Big)\sigma^2,$$

lo que permite sugerir el *estimador insesgado*:

$$s^2 = \frac{1}{n-p}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2.$$

RESULTADO 2.9. *El vector de valores predichos y el vector de residuos son independientemente distribuídos como*

$$\widehat{\boldsymbol{Y}} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H}), \qquad \boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I}-\boldsymbol{H})).$$

DEMOSTRACIÓN. Considere

$$\begin{pmatrix}\widehat{\boldsymbol{Y}}\\ \boldsymbol{e}\end{pmatrix} = \begin{pmatrix}\boldsymbol{H}\\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\boldsymbol{Y},$$

luego,

$$\mathsf{E}\begin{pmatrix}\widehat{\boldsymbol{Y}}\\ \boldsymbol{e}\end{pmatrix} = \begin{pmatrix}\boldsymbol{H}\\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{pmatrix}.$$

Además,

$$\begin{aligned}\mathsf{Cov}\begin{pmatrix}\widehat{\boldsymbol{Y}}\\ \boldsymbol{e}\end{pmatrix} &= \sigma^2\begin{pmatrix}\boldsymbol{H}\\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\big(\boldsymbol{H}^\top, (\boldsymbol{I}-\boldsymbol{H})^\top\big)\\ &= \sigma^2\begin{pmatrix}\boldsymbol{H}\boldsymbol{H}^\top & \boldsymbol{H}(\boldsymbol{I}-\boldsymbol{H})^\top\\ (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{H}^\top & (\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})^\top\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}\boldsymbol{H}^2 & \boldsymbol{0}\\ \boldsymbol{0} & (\boldsymbol{I}-\boldsymbol{H})^2\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}\boldsymbol{H} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}-\boldsymbol{H}\end{pmatrix},\end{aligned}$$

lo que permite establecer el resultado. □

OBSERVACIÓN 2.10. Debemos resaltar que $\boldsymbol{H}$ y $\boldsymbol{I}-\boldsymbol{H}$ son matrices de rango incompleto y por tanto $\widehat{\boldsymbol{Y}}$ y $\boldsymbol{e}$ siguen distribuciones normales *singulares*.

A continuación vamos a suponer que el vector de respuestas y la matriz de diseño dependen del tamaño muestral, n. Considere el siguiente problema

$$\min_{\beta} Q_n(\boldsymbol{\beta}) = \min_{\beta}\|\boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\|^2,$$

cuya solución, $\widehat{\boldsymbol{\beta}}_n$ es llamado estimador *mínimos cuadrados (LS)*, dado por:

$$\widehat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n^\top\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top\boldsymbol{Y}_n,$$

Es habitual asumir el modelo

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

con $\mathsf{E}(\boldsymbol{\epsilon}_n) = \mathbf{0}$ y $\mathsf{Cov}(\boldsymbol{\epsilon}_n) = \sigma^2 \boldsymbol{I}_n$. Esto lleva a

$$\mathsf{E}(\widehat{\boldsymbol{\beta}}_n) = (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top \mathsf{E}(\boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\epsilon}_n) = \boldsymbol{\beta}$$
$$\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_n) = \mathsf{Cov}((\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top (\boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\epsilon}_n)) = \sigma^2 (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}.$$

SUPUESTO 3. *Considere el modelo,*

$$\boldsymbol{Y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\epsilon}_n, \tag{2.4}$$

y suponga las condiciones:

B1: $\mathsf{E}(\boldsymbol{\epsilon}_n) = \mathbf{0}$, $\mathsf{Cov}(\boldsymbol{\epsilon}_n) = \sigma^2 \boldsymbol{I}$.

B2: *Sea* $h_{kk} = \boldsymbol{x}_{k,n}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{k,n}$ *el* k*-ésimo elemento de la diagonal de* $\boldsymbol{H}_n$ *y considere*

$$\max_{1 \le k \le n} h_{kk} \to 0, \quad \textit{conforme } n \to \infty.$$

B3: $\lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}_n^\top \boldsymbol{X}_n = \boldsymbol{K}$ *es una matriz no singular (no estocástica).*

Usando el supuesto B3, tenemos

$$\lim_{n\to\infty} \mathsf{Cov}(\widehat{\boldsymbol{\beta}}_n) = \sigma^2 \lim_{n\to\infty} \frac{1}{n} \Big(\frac{\boldsymbol{X}_n^\top \boldsymbol{X}_n}{n} \Big)^{-1} = \sigma^2 \lim_{n\to\infty} \frac{1}{n} \boldsymbol{K}^{-1} = \mathbf{0}.$$

Esto implica que $\widehat{\boldsymbol{\beta}}_n \xrightarrow{\mathsf{2nd}} \boldsymbol{\beta}$. Es decir, $\widehat{\boldsymbol{\beta}}_n$ es un estimador consistente de $\boldsymbol{\beta}$.

RESULTADO 2.11 (Distribución asintótica del estimador LS). *Considere el modelo* (2.4) *bajo los supuestos B1 a B3. Entonces*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \sigma^2 \boldsymbol{K}^{-1}).$$

DEMOSTRACIÓN. Ver Sen y Singer (1993), pág. 279. □

RESULTADO 2.12 (Teorema de Gauss-Markov). *Suponga* $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *y sea* $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ *el estimador mínimos cuadrados. Asuma el supuesto B1. El estimador* $\widehat{\gamma} = \boldsymbol{h}^\top \widehat{\boldsymbol{\beta}}$ *de* $\gamma = \boldsymbol{h}^\top \boldsymbol{\beta}$ *es el mejor estimador lineal e insesgado (BLUE).*

DEMOSTRACIÓN. Sea $\boldsymbol{c}^\top \boldsymbol{Y}$ cualquier otro estimador lineal e insesgado de $\gamma = \boldsymbol{h}^\top \boldsymbol{\beta}$. Dado que $\boldsymbol{c}^\top \boldsymbol{Y}$ es insesgado, sigue que

$$\boldsymbol{h}^\top \boldsymbol{\beta} = \mathsf{E}(\boldsymbol{c}^\top \boldsymbol{Y}) = \boldsymbol{c}^\top \mathsf{E}(\boldsymbol{Y}) = \boldsymbol{c}^\top \boldsymbol{X} \boldsymbol{\beta}, \qquad \forall\, \boldsymbol{\beta},$$

luego, tenemos

$$\boldsymbol{c}^\top \boldsymbol{X} = \boldsymbol{h}^\top.$$

Ahora,

$$\mathsf{var}(\boldsymbol{c}^\top \boldsymbol{Y}) = \boldsymbol{c}^\top \mathsf{Cov}(\boldsymbol{Y}) \boldsymbol{c} = \sigma^2 \boldsymbol{c}^\top \boldsymbol{c}, \tag{2.5}$$

mientras que

$$\mathsf{var}(\boldsymbol{h}^\top \widehat{\boldsymbol{\beta}}) = \boldsymbol{h}^\top \mathsf{Cov}(\widehat{\boldsymbol{\beta}}) \boldsymbol{h} = \sigma^2 \boldsymbol{h}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{h} = \sigma^2 \boldsymbol{c}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{c}, \tag{2.6}$$

desde (2.5) y (2.6), tenemos

$$\mathsf{var}(\boldsymbol{c}^\top \boldsymbol{Y}) - \mathsf{var}(\boldsymbol{h}^\top \widehat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{c}^\top \boldsymbol{c} - \boldsymbol{c}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{c})$$
$$\sigma^2 \boldsymbol{c}^\top (\boldsymbol{I} - \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}) \boldsymbol{c} \ge 0,$$

el resultado sigue pues $\boldsymbol{I} - \boldsymbol{H}$ es semidefinida positiva. □

2.3. Aspectos numéricos de estimación LS en regresión lineal

El estimador mínimos cuadrados (o de máximo verosimilitud bajo normalidad) para el modelo en lineal con los Supuestos A1-A4, puede ser expresado como la solución del problema:

$$\min_{\beta\in\mathbb{R}^p} Q(\boldsymbol{\beta}), \quad \text{con} \quad Q(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2,$$

lo que lleva a las ecuaciones de estimación $\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$, o bien

$$\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{Y}. \tag{2.7}$$

Métodos habituales para obtener $\widehat{\boldsymbol{\beta}}$, son:

- Métodos directos basados en la factorización Cholesky, el operador Sweep (Goodnight, 1979) y descomposiciones QR y SVD.
- Un poco menos común en regresión, es el uso del método gradientes conjugados (CG) (McIntosh, 1982).

Observación 2.13. Algunas características de los procedimientos descritos anteriormente son relevantes de ser destacadas, a saber:

- Cholesky y Sweep requieren formar las matrices:

$$\boldsymbol{X}^\top\boldsymbol{X}, \boldsymbol{X}^\top\boldsymbol{Y}, \qquad \text{y} \qquad \begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{Y} \\ \boldsymbol{Y}^\top\boldsymbol{X} & \boldsymbol{Y}^\top\boldsymbol{Y} \end{pmatrix},$$

  respectivamente.
- QR y SVD descomponen la matriz de diseño $\boldsymbol{X}$ y resuelven sistemas lineales (triangular y diagonal, respectivamente) mucho más pequeños ($n \gg p$).
- Una implementación cuidadosa de CG sólo requiere productos matriz-vector/operaciones entre vectores y tan sólo $4p$ posiciones para almacenamiento.
- Existe código confiable y con excelente desempeño para álgebra lineal numérica en las bibliotecas BLAS, LAPACK, rutinas que pueden ser invocadas desde (por ejemplo) R y Matlab.

A continuación introducimos algunas ideas sobre los procedimientos numéricos que serán aplicados al modelo de regresión lineal. Primeramente, considere el siguiente resultado

Resultado 2.14 (Factorización Cholesky). *Sea $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ es matriz simétrica y definida positiva, entonces existe una única matriz triangular superior $\boldsymbol{G} \in \mathbb{R}^{p\times p}$ con elementos diagonales positivos tal que*

$$\boldsymbol{A} = \boldsymbol{G}^\top\boldsymbol{G}$$

Note que si usamos la factorización Cholesky para resolver el sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Entonces debemos resolver los sistemas triangulares

$$\boldsymbol{G}^\top\boldsymbol{z} = \boldsymbol{b}, \qquad \text{y} \qquad \boldsymbol{G}\boldsymbol{x} = \boldsymbol{z}.$$

En efecto,

$$\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{G}^\top\boldsymbol{G})\boldsymbol{x} = \boldsymbol{G}^\top(\boldsymbol{G}\boldsymbol{x}) = \boldsymbol{G}^\top\boldsymbol{z} = \boldsymbol{b}.$$

Algoritmo 1: Factorización Cholesky

Entrada: Matriz $\boldsymbol{A} \in \mathbb{R}^{p\times p}$.
Salida : Factor Cholesky $\boldsymbol{G} \in \mathbb{R}^{p\times p}$.

begin
  $g_{11} = \sqrt{a_{11}}$.
  for $j = 2$ **to** p **do**
    $g_{1j} = a_{1j}/g_{11}$.
  end
  for $i = 2$ **to** p **do**
    $g_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} g_{ki}^2}$,
    for $j = i+1$ **to** n **do**
      $g_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} g_{ki}g_{kj}\right)/g_{ii}$
    end
  end
end

Sea $\mathsf{RSS} = Q(\widehat{\boldsymbol{\beta}})$ la suma de cuadrados residuales y notando que

$$\mathsf{RSS} = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \boldsymbol{Y}^\top\boldsymbol{Y} - \boldsymbol{Y}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Tenemos,

$$\boldsymbol{Y}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{Y}^\top\boldsymbol{H}^2\boldsymbol{Y} = \widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Es decir, podemos escribir:

$$\mathsf{RSS} = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \boldsymbol{Y}^\top\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Podemos resolver (2.7) usando la descomposición Cholesky, de

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{U}^\top\boldsymbol{U},$$

con $\boldsymbol{U}$ matrix triangular superior. De este modo, debemos resolver los sistemas triangulares:

$$\boldsymbol{U}^\top\boldsymbol{z} = \boldsymbol{X}^\top\boldsymbol{Y}, \qquad \text{y} \qquad \boldsymbol{U}\widehat{\boldsymbol{\beta}} = \boldsymbol{z}.$$

Mientras que para obtener s^2, consideramos:

$$RSS = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \boldsymbol{Y}^\top\boldsymbol{Y} - \boldsymbol{z}^\top\boldsymbol{z}.$$

Adicionalmente,

$$\boldsymbol{U}^{-1}\boldsymbol{U}^{-\top} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1},$$

es proporcional a la matriz de covarianza de $\widehat{\boldsymbol{\beta}}$.

Observación 2.15. Invirtiendo $\boldsymbol{U}$ in-place, esto es, haciendo

$$\boldsymbol{U} \leftarrow \boldsymbol{U}^{-1},$$

sigue que $(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \boldsymbol{U}\boldsymbol{U}^\top$ lo que permite ahorrar espacio de almacenamiento y puede ser eficientemente calculado usando rutinas desde BLAS.

DEFINICIÓN 2.16 (Operador Sweep). Sea $\boldsymbol{A} = (a_{ij})$ matriz cuadrada $p \times p$, aplicando el *operador Sweep* sobre el k-ésimo elemento diagonal de $\boldsymbol{A}$ ($a_{kk} \neq 0$) permite obtener la matriz $\boldsymbol{B}$, definida como:

$$\begin{aligned} b_{kk} &= \frac{1}{a_{kk}}, \\ b_{ik} &= -\frac{a_{ik}}{a_{kk}}, & i \neq k, \\ b_{kj} &= \frac{a_{kj}}{a_{kk}}, & j \neq k, \\ b_{ij} &= a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}}, & i, j \neq k, \end{aligned}$$

y escribimos $\boldsymbol{B} = \mathsf{Sweep}(k)\boldsymbol{A}$.

PROPIEDAD 2.17. El operador Sweep disfruta de las siguientes propiedades:

(i) $\mathsf{Sweep}(k)\,\mathsf{Sweep}(k)\boldsymbol{A} = \boldsymbol{A}$.

(ii) $\mathsf{Sweep}(k)\,\mathsf{Sweep}(r)\boldsymbol{A} = \mathsf{Sweep}(r)\,\mathsf{Sweep}(k)\boldsymbol{A}$.

(iii) $\boldsymbol{A}^{-1} = \prod_{i=1}^{n} \mathsf{Sweep}(i)\boldsymbol{A}$.

Debemos destacar que, si $\boldsymbol{A}$ es matriz simétrica, el operador Sweep *preserva la simetría* de $\boldsymbol{A}$. Existen varias definiciones ligeramente diferentes del operador Sweep y más importantemente, es conocido que problemas de *inestabilidad* pueden ocurrir cuando algún a_{kk} es cercano a cero.

Considere $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ matriz particionada como:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix},$$

donde $\boldsymbol{A}_{11} \in \mathbb{R}^{r\times r}$ ($r < p$). Suponga que se aplica el operador Sweep sobre los elementos diagonales de $\boldsymbol{A}_{11}$. De este modo,

$$\boldsymbol{B} = \prod_{i=1}^{r} \mathsf{Sweep}(i)\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix},$$

con

$$\begin{aligned} \boldsymbol{B}_{11} &= \boldsymbol{A}_{11}^{-1}, & \boldsymbol{B}_{12} &= \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}, \\ \boldsymbol{B}_{21} &= -\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}, & \boldsymbol{B}_{22} &= \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}. \end{aligned}$$

Este resultado nos permite utilizar el operador Sweep en problemas de regresión. Considere

$$\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{n\times(p+1)},$$

luego

$$\boldsymbol{Z}^\top\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{Y} \\ \boldsymbol{Y}^\top\boldsymbol{X} & \boldsymbol{Y}^\top\boldsymbol{Y} \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}.$$

que corresponde a una matriz cuadrada de orden $(p+1) \times (p+1)$.

Aplicando el operador Sweep sobre los primeros p elementos diagonales de $\boldsymbol{Z}^\top \boldsymbol{Z}$, obtenemos:

$$\begin{aligned}\boldsymbol{B} &= \prod_{i=1}^{p} \mathsf{Sweep}(i)\boldsymbol{Z}^\top\boldsymbol{Z} \\ &= \begin{pmatrix} (\boldsymbol{X}^\top\boldsymbol{X})^{-1} & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\ -\boldsymbol{Y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} & \boldsymbol{Y}^\top\boldsymbol{Y} - \boldsymbol{Y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \end{pmatrix} \\ &= \begin{pmatrix} (\boldsymbol{X}^\top\boldsymbol{X})^{-1} & \widehat{\boldsymbol{\beta}} \\ -\widehat{\boldsymbol{\beta}}^\top & \mathsf{RSS} \end{pmatrix}.\end{aligned}$$

es decir, este procedimiento nos permite calcular todos los elementos necesarios para llevar a cabo la estimación en un modelo de regresión lineal.

En el contexto de regresión, es usual tener que $n \gg p$ y uno de los procedimientos preferidos para llevar a cabo la estimación mínimos cuadrados esta basado en la descomposición ortogonal-triangular (QR) de la matriz de diseño $\boldsymbol{X}$. Considere la siguiente definición,

Definición 2.18 (Descomposición QR). Sea $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, entonces existe $\boldsymbol{Q} \in \mathcal{O}_n$ y $\boldsymbol{R} \in \mathbb{R}^{n\times p}$, tal que

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}, \qquad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{pmatrix}$$

donde $\boldsymbol{R}_1 \in \mathbb{R}^{p\times p}$ matriz triangular superior, aquí suponemos que $n \geq p$.

En efecto, es fácil notar que si $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, entonces

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{R}_1^\top\boldsymbol{R}_1,$$

y $\boldsymbol{R}_1$ corresponde al factor Cholesky de $\boldsymbol{A}^\top\boldsymbol{A}$. Este aspecto es relevante pues no es necesario formar la matriz de productos cruzados $\boldsymbol{A}^\top\boldsymbol{A}$ para obtener el factor Cholesky $\boldsymbol{R}_1$. Antes de describir brevemente el algoritmo para obtener la descomposición QR recordamos algunas propiedades fundamentales de las matrices ortogonales:

- $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$.
- $\langle \boldsymbol{Q}\boldsymbol{x}, \boldsymbol{Q}\boldsymbol{y}\rangle = \boldsymbol{x}^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{y} = \boldsymbol{x}^\top\boldsymbol{y} = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$.
- $\|\boldsymbol{Q}\boldsymbol{x}\| = \|\boldsymbol{x}\|$.
- Si $\boldsymbol{B} = \boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q}$, entonces $\boldsymbol{A}$ y $\boldsymbol{B}$ tienen los mismos valores propios para $\boldsymbol{Q}$ matriz ortogonal.

Existen diversar variantes del algoritmo para implementar la descomposición QR. A continuación veremos una basada en *transformaciones Householder*. Primeramente, considere el siguiente problema

Problema 2.19. Para $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x} \neq \boldsymbol{0}$, hallar una matriz ortogonal $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ tal que

$$\boldsymbol{M}^\top\boldsymbol{x} = \|\boldsymbol{x}\|\,\boldsymbol{e}_1,$$

donde $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top$ denota el primer vector unidad.

DEFINICIÓN 2.20 (Reflexión). Sea $\boldsymbol{u}$ y $\boldsymbol{v}$ vectores ortonormales y $\boldsymbol{x}$ vector generado por $\boldsymbol{u}$ y $\boldsymbol{v}$. Entonces

$$\boldsymbol{x} = c_1\boldsymbol{u} + c_2\boldsymbol{v},$$

para escalares c_1, c_2. El vector

$$\widetilde{\boldsymbol{x}} = -c_1\boldsymbol{u} + c_2\boldsymbol{v},$$

el llamado una *reflexión* de $\boldsymbol{x}$ a través de la línea definida por el vector $\boldsymbol{v}$ (o $\boldsymbol{u}^{\perp}$).

DEFINICIÓN 2.21 (Transformación Householder). Sea $\boldsymbol{x} = c_1\boldsymbol{u} + c_2\boldsymbol{v}$, con $\boldsymbol{u}$ y $\boldsymbol{v}$ vectores generadores de $\boldsymbol{x}$ y considere la matriz

$$\boldsymbol{H} = \boldsymbol{I} - \lambda\boldsymbol{u}\boldsymbol{u}^{\top}, \qquad \lambda = 2/\boldsymbol{u}^{\top}\boldsymbol{u}.$$

Note que $\boldsymbol{H}\boldsymbol{x} = \widetilde{\boldsymbol{x}}$, es decir $\boldsymbol{H}$ es un reflector.

El objetivo es determinar una matriz $\boldsymbol{M}$ basado en reflexiones Householder. Debemos destacar que la *transformación Householder* satisface las siguientes propiedades:

- $\boldsymbol{H}\boldsymbol{u} = -\boldsymbol{u}$.
- $\boldsymbol{H}\boldsymbol{v} = \boldsymbol{v}$ para cualquier $\boldsymbol{v}$ ortogonal a $\boldsymbol{u}$.
- $\boldsymbol{H}^{\top} = \boldsymbol{H}$.
- $\boldsymbol{H}^{-1} = \boldsymbol{H}^{\top}$.

OBSERVACIÓN 2.22. La operación $\boldsymbol{H}\boldsymbol{x}$ puede ser obtenida usando una operación axpy.[4] En efecto,

$$\boldsymbol{H}\boldsymbol{x} = (\boldsymbol{I} - \lambda\boldsymbol{u}\boldsymbol{u}^{\top})\boldsymbol{x} = \boldsymbol{x} - \alpha\boldsymbol{u}, \qquad \alpha = \lambda\boldsymbol{u}^{\top}\boldsymbol{x}.$$

Algoritmo 2: Descomposición QR

Entrada: Matriz $\boldsymbol{A} \in \mathbb{R}^{n\times p}$.
Salida : Factores $\boldsymbol{Q}$ y $\boldsymbol{R}$, matrices ortogonal y triangular superior, respectivamente.

begin
  Hacer $\boldsymbol{Q} = \boldsymbol{I}_n$ y $\boldsymbol{R} = \boldsymbol{A}$
  for $i = 1$ **to** p **do**
    $\boldsymbol{x} = (R_{1i}, \ldots, R_{pi})^{\top}$
    $\boldsymbol{Q}_i = \begin{pmatrix} \boldsymbol{I}_{i-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}(\boldsymbol{x}) \end{pmatrix}$
    `/* M(x) obtenido usando reflexiones Householder */`
    $\boldsymbol{Q} = \boldsymbol{Q}_i\boldsymbol{Q}$
    $\boldsymbol{R} = \boldsymbol{Q}_i\boldsymbol{R}$
  end
  $\boldsymbol{Q} = \boldsymbol{Q}^{\top}$
  $\boldsymbol{R} = (R_{ij})$ para $i, j = 1, \ldots, p$.
end

[4] Corresponde a una actualización del tipo: $\boldsymbol{y} \leftarrow \alpha\boldsymbol{x} + \boldsymbol{y}$.

La descomposición QR de una matriz $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ $(n > p)$, puede ser construída a través de una secuencia de matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p$ tales que

$$\boldsymbol{Q}_p \cdots \boldsymbol{Q}_1 \boldsymbol{A} = \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix},$$

donde $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p$ son *todas* ortogonales. De este modo,

$$\boldsymbol{A} = \boldsymbol{Q}_1^\top \cdots \boldsymbol{Q}_p^\top \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{Q} \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix}.$$

El Algoritmo 2 permite obtener la descomposición QR de $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ usando transformaciones Householder.[5] En este contexto, $\boldsymbol{M}(\boldsymbol{x})$ corresponde a la matriz ortogonal desde el Problema 2.19 basada en un vector $\boldsymbol{x}$. Debemos destacar que una implementación eficiente del Algoritmo 2 solamente requiere almacenar la información mínima para formar las matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p$ y la propia matriz triangular $\boldsymbol{R}_1$ en la propia matriz $\boldsymbol{A}$, lo que permite un ahorro desde el punto de vista de almacenamiento.

Para el problema de regresión, considere la descomposición QR de $\boldsymbol{X}$, como:

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}, \qquad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{pmatrix},$$

con $\boldsymbol{R}_1 \in \mathbb{R}^{p\times p}$ matriz triangular superior $(n > p)$. Si $\text{rg}(\boldsymbol{X}) = p$, entonces $\boldsymbol{R}_1$ es no singular. Además, considere la transformación:

$$\boldsymbol{Q}^\top \boldsymbol{Y} = \boldsymbol{c}, \qquad \boldsymbol{c} = \begin{pmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{pmatrix}.$$

La descomposición QR de $\boldsymbol{X}$ permite re-escribir la función objetivo asociada al problema mínimos cuadrados, como:

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 &= \|\boldsymbol{Q}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\|^2 = \|\boldsymbol{Q}^\top\boldsymbol{Y} - \boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta}\|^2 \\ &= \|\boldsymbol{c} - \boldsymbol{R}\boldsymbol{\beta}\|^2, \end{aligned}$$

Es fácil notar que:

$$\|\boldsymbol{c} - \boldsymbol{R}\boldsymbol{\beta}\|^2 = \|\boldsymbol{c}_1 - \boldsymbol{R}_1\boldsymbol{\beta}\|^2 + \|\boldsymbol{c}_2\|^2.$$

Esto lleva a escribir el estimador de mínimos cuadrados $\widehat{\boldsymbol{\beta}}$ como solución del sistema triangular:

$$\boldsymbol{R}_1\widehat{\boldsymbol{\beta}} = \boldsymbol{c}_1.$$

El mínimo de la función objetivo está dado por $\|\boldsymbol{c}_2\|^2$, lo que permite calcular el estimador insesgado de σ^2 como

$$s^2 = \frac{1}{n-p}\|\boldsymbol{c}_2\|^2 = \frac{\mathsf{RSS}}{n-p},$$

Finalmente,

$$\boldsymbol{X}^\top\boldsymbol{X} = (\boldsymbol{R}_1^\top, \boldsymbol{0})\boldsymbol{Q}^\top\boldsymbol{Q}\begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{R}_1^\top\boldsymbol{R}_1,$$

lo que ofrece un procedimiento eficiente para obtener la matriz de covarianza de $\widehat{\boldsymbol{\beta}}$, dado por

$$\mathsf{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{R}_1^\top\boldsymbol{R}_1)^{-1} = \sigma^2\boldsymbol{R}_1^{-1}\boldsymbol{R}_1^{-\top}.$$

[5] Otro método popular para obtener la descomposición QR es usando rotaciones Givens.

DEFINICIÓN 2.23 (Descomposición Valor Singular). Sea $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ con $\text{rg}(\boldsymbol{A}) = r$, entonces existen matrices $\boldsymbol{U} \in \mathcal{O}_n$, $\boldsymbol{V} \in \mathcal{O}_p$, tal que

$$\boldsymbol{A} = \boldsymbol{U}\begin{pmatrix} \boldsymbol{D}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{V}^\top,$$

donde $\boldsymbol{D}_r = \text{diag}(\delta_1, \ldots, \delta_r)$ con $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r > 0$, que son llamados valores singulares de $\boldsymbol{A}$.

La Descomposición Valor Singular (SVD) para $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ con $\text{rg}(\boldsymbol{A}) = r$ puede ser escrita como:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

con $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ tal que $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_p$, $\boldsymbol{D} = \text{diag}(\delta_1, \ldots, \delta_r)$ y $\boldsymbol{V} \in \mathcal{O}_r$.
Para el contexto de regresión lineal, considere la SVD de $\boldsymbol{X}$,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

donde $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ tal que $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_p$, $\boldsymbol{D} = \text{diag}(\delta_1, \ldots, \delta_p)$ y $\boldsymbol{V} \in \mathcal{O}_p$. De este modo, podemos escribir el modelo de regresión lineal como:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

con $\boldsymbol{\alpha} = \boldsymbol{V}^\top\boldsymbol{\beta}$. Haciendo $\boldsymbol{Z} = \boldsymbol{U}^\top\boldsymbol{Y}$, tenemos el modelo en *forma canónica*:

$$\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\eta}, \qquad \boldsymbol{\eta} = \boldsymbol{U}^\top\boldsymbol{\epsilon},$$

donde

$$\mathsf{E}(\boldsymbol{\eta}) = \boldsymbol{0}, \qquad \mathsf{Cov}(\boldsymbol{\eta}) = \sigma^2\boldsymbol{U}^\top\boldsymbol{U} = \sigma^2\boldsymbol{I}_p.$$

Es decir, podemos el modelo canónico satisface las condiciones A1 a A4. Esto lleva al estimador LS de $\boldsymbol{\alpha}$ en el modelo canónico,

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{D}^{-1}\boldsymbol{Z}, \qquad \Rightarrow \qquad \widehat{\boldsymbol{\beta}} = \boldsymbol{V}\widehat{\boldsymbol{\alpha}}.$$

Además,

$$\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{Z} - \boldsymbol{D}\widehat{\boldsymbol{\alpha}}\|^2.$$

Finalmente,

$$\mathsf{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top)^{-1} = \sigma^2\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\top.$$

OBSERVACIÓN 2.24. Interesantemente, la SVD permite manipular problemas de rango deficiente. En efecto, cuando $\text{rg}(\boldsymbol{X}) < p$ podemos considerar

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{D}^{-}\boldsymbol{Z},$$

con $\boldsymbol{D}^{-}$ una inversa generalizada de $\boldsymbol{D}$ por ejemplo

$$\boldsymbol{D}^{-} = \text{diag}(1/\delta_1, \ldots, 1/\delta_r, 0, \ldots, 0), \qquad r = \text{rg}(\boldsymbol{X}),$$

y luego obtener $\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\widehat{\boldsymbol{\alpha}}$.

El último procedimiento de estimación que revisaremos en esta sección corresponde al método de Gradientes Conjugados (CG), el que permite optimizar la siguiente función objetivo:

$$\phi(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

El algoritmo básico produce la secuencia de estimaciones,

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \lambda_k\boldsymbol{p}_k, \qquad k = 0, 1, \ldots.$$

con el siguiente 'largo de paso'

$$\lambda_k = \frac{\boldsymbol{p}_k^\top \boldsymbol{g}_k}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k}, \qquad \boldsymbol{g}_k = \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}).$$

y actualizamos la dirección de búsqueda como:

$$\boldsymbol{p}_{k+1} = \boldsymbol{g}_{k+1} + \delta_k \boldsymbol{p}_k, \qquad \delta_k = -\frac{\boldsymbol{g}_{k+1}^\top \boldsymbol{p}_k}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k}.$$

Para el contexto de regresión se ha sugerido modificar la versión básica del algoritmo, considerando (ver McIntosh, 1982)

$$\lambda_k = \frac{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{Y}}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k},$$

y actualizar la dirección de búsqueda,

$$\boldsymbol{p}_{k+1} = \boldsymbol{g}_{k+1} + \delta_{k+1} \boldsymbol{p}_k, \qquad \delta_{k+1} = -\frac{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{g}_k}{\boldsymbol{p}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k}.$$

Para hacer el proceso más simple es recomendable calcular el vector

$$\boldsymbol{h}_k = \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}_k,$$

lo que lleva a una implementación que solo requiere operaciones matriz-vector. Además, debemos destacar que no hace falta formar la matriz $\boldsymbol{X}^\top \boldsymbol{X}$. Lo que permite notar que los requerimientos de almacenamiento del algoritmo sólo es de $4p$ ubicaciones de memoria.

Algoritmo 3: Gradientes conjugados para regresión lineal.

Entrada : Datos $\boldsymbol{X}$ y $\boldsymbol{Y}$
Parámetros: Tolerancia τ.

begin
  Hacer $\boldsymbol{\beta} = \boldsymbol{0}$, $\boldsymbol{p} = \boldsymbol{g} = -\boldsymbol{X}^\top \boldsymbol{Y}$, $\delta = 0$ y $\gamma = \|\boldsymbol{g}\|^2$
  while $\gamma > \tau$ **do**
    Calcular $\boldsymbol{h} = \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p}$ y $u = \boldsymbol{p}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{p} = \boldsymbol{p}^\top \boldsymbol{h}$
    $v = \boldsymbol{g}^\top \boldsymbol{g}$
    $\lambda = -v/u$
    $\boldsymbol{\beta} = \boldsymbol{\beta} + \lambda \boldsymbol{p}$
    $\boldsymbol{g} = \boldsymbol{g} + \lambda \boldsymbol{h}$
    $\delta = \boldsymbol{g}^\top \boldsymbol{g}/v$
    $\boldsymbol{p} = \boldsymbol{g} + \delta \boldsymbol{p}$
  end
  return $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$
end

Los cinco procedimientos descritos en esta sección han sido implementados en la función `ols` en la biblioteca fastmatrix (Osorio y Ogueda, 2021) disponible para el ambiente de cálculo estadístico R (R Core Team, 2020).

2.4. Estimación bajo restricciones lineales

El objetivo de esta sección es abordar la estimación de $\boldsymbol{\beta}$ y σ^2 sujeto a restricciones lineales del tipo:

$$\boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}, \tag{2.8}$$

donde $\boldsymbol{G} \in \mathbb{R}^{q\times p}$ con $\text{rg}(\boldsymbol{G}) = q$ y $\boldsymbol{g} \in \mathbb{R}^q$. Consideraremos dos procedimientos para obtener estimadores restrigidos, a saber:

- Método del modelo reducido.
- Método de multiplicadores de Lagrange.

Además, estudiaremos las propiedades estadísticas de los estimadores.

2.4.1. Método del modelo reducido. Para introducir este procedimiento, considere la siguiente partición $\boldsymbol{G} = (\boldsymbol{G}_r, \boldsymbol{G}_q)$ donde $\boldsymbol{G}_q \in \mathbb{R}^{q\times q}$ de rango q. De este modo, podemos escribir las restricciones en (2.8) como:

$$\boldsymbol{G}\boldsymbol{\beta} = (\boldsymbol{G}_r, \boldsymbol{G}_q)\begin{pmatrix}\boldsymbol{\beta}_r \\ \boldsymbol{\beta}_q\end{pmatrix} = \boldsymbol{G}_r\boldsymbol{\beta}_r + \boldsymbol{G}_q\boldsymbol{\beta}_q = \boldsymbol{g},$$

como $\boldsymbol{G}_q$ es no singular, tenemos

$$\boldsymbol{\beta}_q = \boldsymbol{G}_q^{-1}(\boldsymbol{g} - \boldsymbol{G}_r\boldsymbol{\beta}_r).$$

Particionando $\boldsymbol{X}$ del mismo modo que $\boldsymbol{\beta} = (\boldsymbol{\beta}_r^\top, \boldsymbol{\beta}_q^\top)^\top$, sigue que

$$\begin{aligned}\boldsymbol{X}\boldsymbol{\beta} &= (\boldsymbol{X}_r, \boldsymbol{X}_q)\begin{pmatrix}\boldsymbol{\beta}_r \\ \boldsymbol{\beta}_q\end{pmatrix} = \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{X}_q\boldsymbol{\beta}_q \\ &= \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{X}_q\boldsymbol{G}_q^{-1}(\boldsymbol{g} - \boldsymbol{G}_r\boldsymbol{\beta}_r) \\ &= (\boldsymbol{X}_r - \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{G}_r)\boldsymbol{\beta}_r + \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{g}\end{aligned}$$

Es decir, podemos escribir el modelo lineal

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

como:

$$\boldsymbol{Y} = (\boldsymbol{X}_r - \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{G}_r)\boldsymbol{\beta}_r + \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{g} + \boldsymbol{\epsilon},$$

esto lleva al *modelo reducido*, dado por

$$\boldsymbol{Y}_R = \boldsymbol{X}_R\boldsymbol{\beta}_r + \boldsymbol{\epsilon}, \tag{2.9}$$

donde

$$\boldsymbol{Y}_R = \boldsymbol{Y} - \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{g}, \qquad \boldsymbol{X}_R = \boldsymbol{X}_r - \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{G}_r,$$

corresponde al vector de respuesta y la matriz de diseño en el modelo reducido. La principal ventaja de este procedimiento es que permite obtener estimadores en el modelo (2.9) de forma simple. En efecto,

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_r &= (\boldsymbol{X}_R^\top\boldsymbol{X}_R)^{-1}\boldsymbol{X}_R^\top\boldsymbol{Y}_R, \\ s_r^2 &= \frac{1}{n-r}Q_R(\widetilde{\boldsymbol{\beta}}_r),\end{aligned}$$

con

$$Q_R(\widetilde{\boldsymbol{\beta}}_r) = \boldsymbol{Y}_R^\top(\boldsymbol{I} - \boldsymbol{X}_R(\boldsymbol{X}_R^\top\boldsymbol{X}_R)^{-1}\boldsymbol{X}_R^\top)\boldsymbol{Y}_R$$

Por otro lado, el vector de coeficientes estimados bajo las restricciones lineales en (2.8)

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} = \begin{pmatrix}\widetilde{\boldsymbol{\beta}}_r \\ \widetilde{\boldsymbol{\beta}}_q\end{pmatrix} &= \begin{pmatrix}\widetilde{\boldsymbol{\beta}}_r \\ \boldsymbol{G}_q^{-1}(\boldsymbol{g} - \boldsymbol{G}_r\widetilde{\boldsymbol{\beta}}_r)\end{pmatrix} \\ &= \begin{pmatrix}\boldsymbol{0} \\ \boldsymbol{G}_q^{-1}\boldsymbol{g}\end{pmatrix} + \begin{pmatrix}\boldsymbol{I} \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r\end{pmatrix}\widetilde{\boldsymbol{\beta}}_r,\end{aligned} \tag{2.10}$$

Mientras que el vector de 'residuos' en el modelo reducido adopta la forma,

$$\begin{aligned}\boldsymbol{Y}_R - \boldsymbol{X}_R\widetilde{\boldsymbol{\beta}}_r &= \boldsymbol{Y} - \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{g} - (\boldsymbol{X}_r - \boldsymbol{X}_q\boldsymbol{G}_q^{-1}\boldsymbol{g})\widetilde{\boldsymbol{\beta}}_r \\ &= \boldsymbol{Y} - (\boldsymbol{X}_r, \boldsymbol{X}_q)\begin{pmatrix}\widetilde{\boldsymbol{\beta}}_r \\ \widetilde{\boldsymbol{\beta}}_q\end{pmatrix} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}},\end{aligned}$$

de este modo

$$Q_R(\widetilde{\boldsymbol{\beta}}_r) = \|\boldsymbol{Y}_R - \boldsymbol{X}_R\widetilde{\boldsymbol{\beta}}_r\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\|^2 = Q(\widetilde{\boldsymbol{\beta}}). \tag{2.11}$$

Las expresiones anteriores permiten establecer el siguiente resultado.

RESULTADO 2.25. *Para el modelo lineal $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ sujeto a las restricciones $\boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}$ con $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. El MLE restringido de $\boldsymbol{\beta}$ es dado por (2.10) y tenemos que*

$$\widetilde{\boldsymbol{\beta}} \sim \mathsf{N}_p(\boldsymbol{\beta}, \mathsf{Cov}(\widetilde{\boldsymbol{\beta}})),$$

donde

$$\mathsf{Cov}(\widetilde{\boldsymbol{\beta}}) = \sigma^2\begin{pmatrix}\boldsymbol{I} \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r\end{pmatrix}(\boldsymbol{X}_R^\top\boldsymbol{X}_R)^{-1}\begin{pmatrix}\boldsymbol{I} \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r\end{pmatrix}^\top.$$

Mientras que

$$s_r^2 = \frac{1}{n-r}Q_R(\widetilde{\boldsymbol{\beta}}_r) = \frac{1}{n-r}Q(\widetilde{\boldsymbol{\beta}}),$$

donde

$$\frac{(n-r)s_r^2}{\sigma^2} \sim \chi^2(n-r),$$

y $\widetilde{\boldsymbol{\beta}}$, $Q(\widetilde{\boldsymbol{\beta}})$ son independientes.

DEMOSTRACIÓN. Sabemos que

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_r &\sim \mathsf{N}_r(\boldsymbol{\beta}_r, \sigma^2(\boldsymbol{X}_R^\top\boldsymbol{X}_R)^{-1}), \\ \frac{(n-r)s_r^2}{\sigma^2} &= \frac{Q_R(\widetilde{\boldsymbol{\beta}}_r)}{\sigma^2} = \frac{\boldsymbol{Y}_R^\top(\boldsymbol{I} - \boldsymbol{H}_R)\boldsymbol{Y}_R}{\sigma^2} \sim \chi^2(n-r).\end{aligned}$$

Así, por (2.10), tenemos

$$\begin{aligned}\mathsf{E}(\widetilde{\boldsymbol{\beta}}) &= \begin{pmatrix}\boldsymbol{0} \\ \boldsymbol{G}_q^{-1}\boldsymbol{g}\end{pmatrix} + \begin{pmatrix}\boldsymbol{I}_r \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r\end{pmatrix}\mathsf{E}(\widetilde{\boldsymbol{\beta}}_r) \\ &= \begin{pmatrix}\boldsymbol{\beta}_r \\ \boldsymbol{G}_q^{-1}(\boldsymbol{g} - \boldsymbol{G}_r\boldsymbol{\beta}_r)\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}_r \\ \boldsymbol{\beta}_q\end{pmatrix} = \boldsymbol{\beta}\end{aligned}$$

Además,

$$\begin{aligned}\mathsf{Cov}(\widetilde{\boldsymbol{\beta}}) &= \begin{pmatrix} \boldsymbol{I}_r \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r \end{pmatrix} \mathsf{Cov}(\widetilde{\boldsymbol{\beta}}_r) \begin{pmatrix} \boldsymbol{I}_r \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r \end{pmatrix}^\top \\ &= \sigma^2 \begin{pmatrix} \boldsymbol{I}_r \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r \end{pmatrix} (\boldsymbol{X}_R^\top \boldsymbol{X}_R)^{-1} \begin{pmatrix} \boldsymbol{I}_r \\ -\boldsymbol{G}_q^{-1}\boldsymbol{G}_r \end{pmatrix}^\top .\end{aligned}$$

como $\widetilde{\boldsymbol{\beta}}$ es una función lineal de $\widetilde{\boldsymbol{\beta}}$ la normalidad sigue. La independencia entre $\widetilde{\boldsymbol{\beta}}$ y $Q(\widetilde{\boldsymbol{\beta}})$ es consecuencia del Resultado 2.7. □

2.4.2. Método de multiplicadores de Lagrange. Considere

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

La función Langrangiana asociada a las restricciones lineales $\boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}$ es dada por:

$$F(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \ell(\boldsymbol{\theta}) + \frac{1}{\sigma^2}\boldsymbol{\lambda}^\top(\boldsymbol{G}\boldsymbol{\beta} - \boldsymbol{g}),$$

con $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$. De este modo,

$$\begin{aligned}\frac{\partial F(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{\sigma^2}\boldsymbol{G}^\top\boldsymbol{\lambda} \\ \frac{\partial F(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\{Q(\boldsymbol{\beta}) - 2\boldsymbol{\lambda}^\top(\boldsymbol{G}\boldsymbol{\beta} - \boldsymbol{g})\} \\ \frac{\partial F(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} &= \boldsymbol{G}\boldsymbol{\beta} - \boldsymbol{g}\end{aligned}$$

Desde la condición de primer orden, obtenemos las ecuaciones de estimación,

$$\begin{aligned}&\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{G}^\top\boldsymbol{\lambda} = \boldsymbol{0}, \\ &n\sigma^2 - \{Q(\boldsymbol{\beta}) - 2\boldsymbol{\lambda}^\top(\boldsymbol{G}\boldsymbol{\beta} - \boldsymbol{g})\} = 0, \\ &\boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g},\end{aligned}$$

es decir,

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{G}^\top\boldsymbol{\lambda}, \tag{2.12}$$

$$\sigma^2 = \frac{1}{n}\, Q(\boldsymbol{\beta}) \tag{2.13}$$

$$\boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}, \tag{2.14}$$

Resolviendo la Ecuación (2.12) con relación a $\boldsymbol{\beta}$ obtenemos

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{G}^\top\boldsymbol{\lambda})$$

Substituyendo este resultado en (2.14) y resolviendo para $\boldsymbol{\lambda}$, sigue que

$$\boldsymbol{G}\widetilde{\boldsymbol{\beta}} = \boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{G}^\top\boldsymbol{\lambda}) = \boldsymbol{g},$$

es decir,

$$\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top\boldsymbol{\lambda} = \boldsymbol{g},$$

por tanto,

$$\widetilde{\boldsymbol{\lambda}} = (\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{g} - \boldsymbol{G}\widehat{\boldsymbol{\beta}}).$$

Reemplazando este resultado en $\widetilde{\boldsymbol{\beta}}$ resulta

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y}+\boldsymbol{G}^\top\widetilde{\boldsymbol{\lambda}})\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}+(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{g}-\boldsymbol{G}\widehat{\boldsymbol{\beta}})\\ &= \widehat{\boldsymbol{\beta}}+(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{g}-\boldsymbol{G}\widehat{\boldsymbol{\beta}}).\end{aligned}$$

Que puede ser reorganizado como:

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}\widehat{\boldsymbol{\beta}}+\boldsymbol{B}\boldsymbol{g} = \widehat{\boldsymbol{\beta}}-\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g}),$$

donde $\widehat{\boldsymbol{\beta}}$ corresponde al MLE no restringido para $\boldsymbol{\beta}$, con

$$\boldsymbol{B} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1} \tag{2.15}$$

$$\boldsymbol{A} = \boldsymbol{I}-\boldsymbol{B}\boldsymbol{G} \tag{2.16}$$

y el estimador insesgado para σ^2 es dado por

$$s_r^2 = \frac{1}{n-r}\,Q(\widetilde{\boldsymbol{\beta}}).$$

Para estudiar las propiedades de este MLE restringido, considere primeramente el siguiente lema.

LEMA 2.26. *La matriz $\boldsymbol{A}$ definida en (2.16) tiene las siguientes propiedades:*

(i) $\boldsymbol{A}$ *es idempotente con* $\operatorname{rg}(\boldsymbol{A}) = r$.
(ii) $\boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ *es idempotente y simétrica con rango* r.
(iii) $\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top = \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top$.

DEMOSTRACIÓN. Para mostrar que $\boldsymbol{A} = \boldsymbol{I}-\boldsymbol{B}\boldsymbol{G}$ es idempotente, basta mostrar que $\boldsymbol{B}\boldsymbol{G}$ es idempotente. En efecto,

$$\boldsymbol{G}\boldsymbol{B}\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}\boldsymbol{G} = \boldsymbol{G},$$

de ahí que $\boldsymbol{B}\boldsymbol{G}$ es idempotente. Además,

$$\operatorname{rg}(\boldsymbol{B}\boldsymbol{G}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{G}) = \operatorname{tr}(\boldsymbol{G}\boldsymbol{B}) = q,$$

así $\operatorname{rg}(\boldsymbol{I}-\boldsymbol{B}\boldsymbol{G}) = \operatorname{tr}(\boldsymbol{A}) = p-q = r$, lo que muestra la parte *(i)*.
Para notar la parte *(ii)*, sea $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, luego

$$\boldsymbol{C}^2 = \boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{X}\boldsymbol{A}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top,$$

como $\boldsymbol{A}$ es idempotente, sigue que $\boldsymbol{C}^2 = \boldsymbol{C}$. Esto permite escribir

$$\operatorname{rg}(\boldsymbol{C}) = \operatorname{tr}(\boldsymbol{C}) = \operatorname{tr}(\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}) = \operatorname{tr}(\boldsymbol{A}) = r.$$

Por otro lado,

$$\boldsymbol{C}^\top = (\boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top.$$

Tenemos que,

$$\boldsymbol{A}^\top = \boldsymbol{I}-\boldsymbol{G}^\top\boldsymbol{B}^\top = \boldsymbol{I}-\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1},$$

luego premultiplicando por $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ y factorizando lleva a,

$$\begin{aligned}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ &= (\boldsymbol{I}-\boldsymbol{B}\boldsymbol{G})(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1},\end{aligned} \tag{2.17}$$

así

$$\boldsymbol{C}^\top = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top\boldsymbol{X} = \boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X} = \boldsymbol{C}.$$

Finalmente, Ecuación (2.17) permite notar la primera igualdad de la parte *(iii)*. Ahora, por (2.17), sigue que

$$\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top = \boldsymbol{A}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1},$$

pues $\boldsymbol{A}$ es idempotente y esto termina la prueba. □

Esto nos habilita para establecer el siguiente resultado,

RESULTADO 2.27. *Para el modelo lineal*

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

El MLE de $\boldsymbol{\beta}$ bajo las restricciones lineales $\boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}$, es dado por

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}\widehat{\boldsymbol{\beta}} + \boldsymbol{B}\boldsymbol{g},$$

con distribución

$$\widetilde{\boldsymbol{\beta}} \sim \mathsf{N}_p(\boldsymbol{\beta}, \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top).$$

Mientras que el estimador insesgado de σ^2 es

$$s_r^2 = \frac{1}{n-r}\, Q(\widetilde{\boldsymbol{\beta}}),$$

donde

$$\frac{Q(\widetilde{\boldsymbol{\beta}})}{\sigma^2} \sim \chi^2(n-r),$$

y $\widetilde{\boldsymbol{\beta}}$ es independiente de $Q(\widetilde{\boldsymbol{\beta}})$.

DEMOSTRACIÓN. La normalidad de $\widetilde{\boldsymbol{\beta}}$ sigue desde la linealidad con relación a $\widehat{\boldsymbol{\beta}}$. Ahora,

$$\mathsf{E}(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{A}\,\mathsf{E}(\widehat{\boldsymbol{\beta}}) + \boldsymbol{B}\boldsymbol{g} = (\boldsymbol{I} - \boldsymbol{B}\boldsymbol{G})\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{g} = \boldsymbol{\beta} - \boldsymbol{B}(\boldsymbol{G}\boldsymbol{\beta} - \boldsymbol{g}) = \boldsymbol{\beta},$$

y

$$\mathsf{Cov}(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{A}\,\mathsf{Cov}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^\top = \sigma^2\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top.$$

Notando que

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \boldsymbol{A}\widehat{\boldsymbol{\beta}} + \boldsymbol{B}\boldsymbol{g} = \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{B}\boldsymbol{g} \\
&= \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) + \boldsymbol{B}\boldsymbol{g} \\
&= \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon} + \boldsymbol{B}\boldsymbol{g} \\
&= \boldsymbol{\beta} + \boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}
\end{aligned}$$

lo que permite escribir

$$\begin{aligned}
\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} &= \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon} \\
&= \boldsymbol{\epsilon} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon} \\
&= (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon},
\end{aligned}$$

por la parte *(ii)* del Lema 2.26, sigue que

$$\frac{Q(\widetilde{\boldsymbol{\beta}})}{\sigma^2} = \frac{\boldsymbol{\epsilon}^\top(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon}}{\sigma^2} \sim \chi^2(n-r).$$

Para notar la independencia entre $\widetilde{\boldsymbol{\beta}}$ y $Q(\widetilde{\boldsymbol{\beta}})$ debemos tener:[6]

$$\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)=\boldsymbol{0}.$$

En efecto,

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)&\\ =\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}-\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top.&\end{aligned}$$

Notando que $\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top$, obtenemos

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top&=\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top\\ =\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top=\boldsymbol{A}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top&=\boldsymbol{A}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top,\end{aligned}$$

y esto concluye la demostración. □

2.5. Test de hipótesis lineales

El objetivo de esta sección recae en desarrollar el test de razón de verosimilitudes para probar hipótesis lineales de la forma:

$$H_0:\boldsymbol{G}\boldsymbol{\beta}=\boldsymbol{g},\qquad\text{versus}\qquad H_1:\boldsymbol{G}\boldsymbol{\beta}\neq\boldsymbol{g}\tag{2.18}$$

donde $\boldsymbol{G}$ es una matriz de contrastes de orden $q\times p$ con $\operatorname{rg}(\boldsymbol{G})=q$ y $\boldsymbol{g}\in\mathbb{R}^q$.

Observación 2.28. H_0 es expresada como un sistema de ecuaciones mientras que H_1 indica que al menos una ecuación no se satisface.

Para abordar hipótesis lineales, usaremos el principio de verosimilitud. Es decir, consideraremos el estadístico

$$\Lambda=\frac{\max\limits_{G\beta=g}L(\boldsymbol{\beta},\sigma^2)}{\max\limits_{\Theta}L(\boldsymbol{\beta},\sigma^2)}=\frac{L(\widetilde{\boldsymbol{\beta}},\widetilde{\sigma}^2)}{L(\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2)}.$$

Asumiendo $\boldsymbol{Y}\sim\mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I})$, tenemos la función de verosimilitud

$$L(\boldsymbol{\beta},\sigma^2)=(2\pi\sigma^2)^{-n/2}\exp\Big\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\Big\}.$$

Sabemos que

$$\begin{aligned}\max_{\Theta}L(\boldsymbol{\beta},\sigma^2)&=L(\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2)\\ &=(2\pi\widehat{\sigma}^2)^{-n/2}\exp\Big\{-\frac{1}{2\widehat{\sigma}^2}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\Big\}\\ &=\{2\pi Q(\widehat{\boldsymbol{\beta}})/n\}^{-n/2}\exp\Big\{-\frac{n}{2Q(\widehat{\boldsymbol{\beta}})}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\Big\}\\ &=\{2\pi Q(\widehat{\boldsymbol{\beta}})/n\}^{-n/2}\exp(-n/2).\end{aligned}$$

Mientras que bajo $H_0:\boldsymbol{G}\boldsymbol{\beta}=\boldsymbol{g}$, tenemos[7]

$$\begin{aligned}\max_{G\beta=g}L(\boldsymbol{\beta},\sigma^2)=L(\widetilde{\boldsymbol{\beta}},\widetilde{\sigma}^2)&=(2\pi\widetilde{\sigma}^2)^{-n/2}\exp\Big\{-\frac{1}{2\widetilde{\sigma}^2}\|\boldsymbol{Y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}\|^2\Big\}\\ &=\{2\pi Q(\widetilde{\boldsymbol{\beta}})/n\}^{-n/2}\exp(-n/2).\end{aligned}$$

[6]Lo que es consecuencia de escribir $\widetilde{\boldsymbol{\beta}}$ y $Q(\widetilde{\boldsymbol{\beta}})$ en términos de $\boldsymbol{\epsilon}$.

[7]Con espacio paramétrico nulo, $\Theta_0=\{\boldsymbol{\theta}=(\boldsymbol{\beta}^\top,\sigma^2)^\top:\boldsymbol{G}\boldsymbol{\beta}=\boldsymbol{g}\}$.

De este modo, el estadístico de razón de verosimilitudes

$$\begin{aligned}\Lambda &= \frac{L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)}{L(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)} = \frac{\{2\pi Q(\widetilde{\boldsymbol{\beta}})/n\}^{-n/2}\exp(-n/2)}{\{2\pi Q(\widehat{\boldsymbol{\beta}})/n\}^{-n/2}\exp(-n/2)} \\ &= \Big\{\frac{Q(\widehat{\boldsymbol{\beta}})}{Q(\widetilde{\boldsymbol{\beta}})}\Big\}^{n/2},\end{aligned}$$

y de acuerdo con el principio de verosimilitud rechazamos $H_0 : \boldsymbol{G\beta} = \boldsymbol{g}$ si Λ es pequeño.

Alternativamente, podemos considerar

$$\Lambda^{2/n} = \frac{Q(\widehat{\boldsymbol{\beta}})}{Q(\widetilde{\boldsymbol{\beta}})}.$$

Recuerde que $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})$, así

$$\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})) = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{XB}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}).$$

Además,

$$\begin{aligned}Q(\widetilde{\boldsymbol{\beta}}) &= \|\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{\beta}} - \boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})\|^2 \\ &= \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{XB}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})\|^2 \\ &+ (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top \boldsymbol{XB}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}) + (\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})^\top \boldsymbol{B}^\top \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).\end{aligned}$$

Sin embargo,

$$\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{0}$$

lo que nos lleva a:

$$\begin{aligned}Q(\widetilde{\boldsymbol{\beta}}) &= Q(\widehat{\boldsymbol{\beta}}) + (\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})^\top \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{XB}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}) \\ &\geq Q(\widehat{\boldsymbol{\beta}})\end{aligned}$$

Observación 2.29. Es decir,

$$0 \leq \frac{Q(\widehat{\boldsymbol{\beta}})}{Q(\widetilde{\boldsymbol{\beta}})} \leq 1,$$

por tanto, $\Lambda^{2/n} \in [0, 1]$.

Además, recordando que

$$\boldsymbol{B} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1},$$

obtenemos

$$\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{XB} = (\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}.$$

De este modo,

$$Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}).$$

Así

$$\frac{Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}})}{Q(\widehat{\boldsymbol{\beta}})} = \Lambda^{-2/n} - 1,$$

es decir, valores pequeños de Λ (en cuyo caso rechazamos H_0) implican valores grandes de la razón anterior.

Ahora considere

$$\begin{aligned}\mathsf{E}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g}) &= \boldsymbol{G}\,\mathsf{E}(\widehat{\boldsymbol{\beta}})-\boldsymbol{g}=\boldsymbol{G}\boldsymbol{\beta}-\boldsymbol{g}\\ \mathsf{Cov}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g}) &= \boldsymbol{G}\,\mathsf{Cov}(\widehat{\boldsymbol{\beta}})\boldsymbol{G}^\top=\sigma^2\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top.\end{aligned}$$

Note que

$$(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}=\Big\{\frac{1}{\sigma^2}\,\mathsf{Cov}(\boldsymbol{G}\widehat{\boldsymbol{\beta}})\Big\}^{-1}.$$

De esta forma

$$\frac{(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})}{\sigma^2}\sim\chi^2(q;\delta),$$

pues $\sigma^{-2}(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}\,\mathsf{Cov}(\boldsymbol{G}\widehat{\boldsymbol{\beta}})=\boldsymbol{I}$, es matriz idempotente, y

$$\delta=\frac{1}{2\sigma^2}(\boldsymbol{G}\boldsymbol{\beta}-\boldsymbol{g})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\boldsymbol{\beta}-\boldsymbol{g})$$

Por otro lado, sabemos que

$$\frac{Q(\widehat{\boldsymbol{\beta}})}{\sigma^2}\sim\chi^2(n-p;0)$$

Para notar la independencia, considere $\boldsymbol{\beta}^*$ cualquier vector que satisface la condición $\boldsymbol{G}\boldsymbol{\beta}^*=\boldsymbol{g}$. Entonces,

$$\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}=(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*),$$

pues $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}=\boldsymbol{0}$ y

$$\begin{aligned}\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g} &= \boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}-\boldsymbol{G}\boldsymbol{\beta}^*=\boldsymbol{G}\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}-\boldsymbol{\beta}^*\}\\ &= \boldsymbol{G}\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}^*\}\\ &= \boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*).\end{aligned}$$

En nuestro caso,

$$\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*\sim\mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\boldsymbol{\beta}^*,\sigma^2\boldsymbol{I})\overset{\mathsf{d}}{=}\mathsf{N}_n(\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*),\sigma^2\boldsymbol{I}).$$

De este modo,

$$Q(\widehat{\boldsymbol{\beta}})=\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2=(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*)^\top(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*),$$

mientras que

$$\begin{aligned}Q(\widetilde{\boldsymbol{\beta}})-Q(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*)^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}\\ &\quad\times\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*).\end{aligned}$$

Como

$$(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top=\boldsymbol{0},$$

sigue la independencia y permite construir la estadística

$$F=\frac{\{Q(\widetilde{\boldsymbol{\beta}})-Q(\widehat{\boldsymbol{\beta}})\}/q}{Q(\widehat{\boldsymbol{\beta}})/(n-p)}\sim F(q,n-p;\delta).$$

Esto lleva al siguiente resultado.

RESULTADO 2.30. *Para el modelo lineal* $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *con los supuestos A1 a A4**. *Un test de tamaño* α *para probar*

$$H_0 : \boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}, \qquad \textit{versus} \qquad H_1 : \boldsymbol{G}\boldsymbol{\beta} \neq \boldsymbol{g},$$

es dado por, rechazar H_0 *cuando:*

$$F = \frac{Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}})}{qs^2} \geq F_{1-\alpha}(q, n-p; 0).$$

DEMOSTRACIÓN. Bajo $H_0 : \boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}$, tenemos $\delta = 0$, de ahí que $F \sim F(q, n-p; 0)$, lo que lleva al resultado deseado. □

OBSERVACIÓN 2.31. Note que podemos escribir el estadístico F de varias formas equivalentes, a saber:

$$\begin{aligned} F &= \Big(\frac{n-p}{q}\Big)\frac{Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}})}{Q(\widehat{\boldsymbol{\beta}})} \\ &= \frac{Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}})}{qs^2} \\ &= \frac{(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})^\top (\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})}{qs^2} \sim F(q, n-p, \delta). \end{aligned}$$

Hemos notado la relación que existe entre el test de razón de verosimilitudes con el estadístico F para probar hipótesis lineales en el modelo de regresión lineal

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

A continuación exploramos la relación entre el estadístico F con los *test de Wald, score* y *gradiente* para hipótesis lineales, del tipo

$$H_0 : \boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}, \qquad \text{versus} \qquad H_1 : \boldsymbol{G}\boldsymbol{\beta} \neq \boldsymbol{g}.$$

Primeramente, note que la matriz de información de Fisher para $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$, adopta la forma:

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{n}{2\sigma^2} \end{pmatrix}.$$

El estadístico de Wald para hipótesis lineales de la forma $H_0 : \boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}$, es dado por

$$\begin{aligned} W &= n(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})^\top (\boldsymbol{G}\{\boldsymbol{\mathcal{F}}(\widehat{\boldsymbol{\beta}})\}^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}) \\ &= \frac{n(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})^\top (\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})}{\widehat{\sigma}^2}. \end{aligned}$$

Mientras que el test score es dado por

$$\begin{aligned} R &= \frac{1}{n}\boldsymbol{U}^\top(\widetilde{\boldsymbol{\beta}})\{\boldsymbol{\mathcal{F}}(\widetilde{\boldsymbol{\beta}})\}^{-1}\boldsymbol{U}(\widetilde{\boldsymbol{\beta}}) \\ &= \frac{1}{n\widetilde{\sigma}^2}(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}). \end{aligned}$$

Como,

$$\begin{aligned} \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) &= \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g})) \\ &= \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}) \\ &= \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}). \end{aligned}$$

Sigue que,

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}})^\top \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \\ &= (\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g}) \\ &= (\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g}) \\ &= (\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g}).\end{aligned}$$

Finalmente,

$$R=\frac{(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})}{n\widetilde{\sigma}^2}$$

Por otro lado, el estadístico gradiente es dado por:

$$T=\boldsymbol{U}^\top(\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})=(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})$$

Sabemos que $\widetilde{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}-\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})$, esto lleva a

$$\begin{aligned}T&=(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}+\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})) \\ &=(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g}) \\ &=(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}-\boldsymbol{g})\end{aligned}$$

Observación 2.32. Es decir, podemos escribir:

$$W=\frac{qn}{n-p}F,\qquad R=\Big\{1+\Big(\frac{n-p}{q}\Big)F^{-1}\Big\}^{-1},\qquad T=\frac{qs^2}{n-p}F.$$

Lo que permite notar que basta usar el estadístico F para probar hipótesis lineales en el modelo de regresión lineal.

2.6. Regiones de confianza

Una región de confianza del $100(1-\alpha)\,\%$ para $\boldsymbol{\gamma}$ se define como la región en el espacio paramétrico, digamos $RC(\boldsymbol{\gamma})$ con la propiedad

$$\mathsf{P}(\boldsymbol{\gamma}\in RC(\boldsymbol{\gamma}))=1-\alpha,$$

donde $\boldsymbol{\gamma}$ es el verdadero vector de parámetros.

Sea $\boldsymbol{\gamma}=\boldsymbol{G}\boldsymbol{\beta}$ un vector de parámetros q-dimensional. Sabemos que

$$\frac{(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})}{qs^2}\sim F(q,n-p),$$

luego, puede ser usada como una región de confianza. Es decir,

$$CR(\boldsymbol{\gamma})=\big\{\boldsymbol{\gamma}:(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})^\top(\boldsymbol{G}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{G}^\top)^{-1}(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}})\le qs^2F_{1-\alpha}(q,n-p)\big\},$$

donde $F_{1-\alpha}(q,n-p)$ es el valor cuantil $1-\alpha$ de la distribución F con q y $n-p$ grados de libertad. De este modo, para $\boldsymbol{G}=\boldsymbol{I}$ obtenemos el siguiente caso particular

$$CR(\boldsymbol{\beta})=\big\{\boldsymbol{\beta}:(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\le ps^2F_{1-\alpha}(p,n-p)\big\}.$$

Notando que

$$\frac{\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}} - \boldsymbol{a}^\top\boldsymbol{\beta}}{s\sqrt{\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}}} \sim t(n-p), \qquad \boldsymbol{a} \in \mathbb{R}^p,$$

corresponde a una cantidad pivotal, podemos escribir un intervalo de confianza para la combinación lineal $\tau = \boldsymbol{a}^\top\boldsymbol{\beta}$. Esto es,

$$\tau \in \left[\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}} \mp t_{1-\alpha/2}(n-p)s\sqrt{\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}}\right],$$

donde $t_{1-\alpha/2}(n-p)$ denota el valor cuantil $1-\alpha/2$ de la distribución t de Student con $n-p$ grados de libertad. También podemos escribir de forma equivalente:

$$CR(\boldsymbol{a}^\top\boldsymbol{\beta}) = \left\{\tau : (\tau - \boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}) \leq s^2\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}\, F_{1-\alpha}(1, n-p)\right\}.$$

En particular, para β_j $(j = 1, \ldots, p)$ basta escoger $\boldsymbol{a} = \boldsymbol{e}_j$ el j-ésimo vector unidad. Así,

$$CI(\beta_j) = \left[\widehat{\beta}_j \mp t_{1-\alpha/2}(n-p)s\sqrt{c_{jj}}\right],$$

donde c_{jj} representa el j-ésimo elemento de la diagonal de $\boldsymbol{C} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

Ejercicios

2.1 Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes con $Y_i \sim \mathsf{N}(\alpha + \theta z_i, \sigma^2)$, $i = 1, \ldots, n$, donde $\{z_i\}$ son constantes conocidas, tales que $\sum_{i=1}^n z_i = 0$. Obtenga el estimador ML de $\boldsymbol{\beta} = (\alpha, \theta)^\top$ y determine su matriz de covarianza. ¿Son $\widehat{\alpha}$ y $\widehat{\theta}$ independientes?

2.2 Sea Y_{ij}, para $i = 1, 2, 3$ y $j = 1, \ldots, m$ variables aleatorias independientes con distribución normal, tales que $\mathsf{E}(Y_{ij}) = \mu_{ij}$, $\mathsf{var}(Y_{ij}) = \sigma^2$, y

$$\mu_{1j} = \tau, \qquad \mu_{2j} = \tau + \theta, \qquad \mu_{3j} = \tau - \theta.$$

a) Determine la matriz de diseño $\boldsymbol{X}$.
b) Obtener el estimador ML de $(\tau, \theta)^\top$ y $\mathsf{var}(\widehat{\theta})$.
c) Derive el estadístico F para probar la hipótesis $H_0 : \theta = 0$.

2.3 Sea $\overline{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$ y $z_{ij} = x_{ij} - \overline{x}_j$, para todo $i = 1, \ldots, n$; $j = 1, \ldots, k$ y considere $\boldsymbol{Z} = (z_{ij})$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$. En cuyo caso, tenemos el *modelo centrado*:

$$\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{1}, \boldsymbol{Z})\begin{pmatrix}\alpha\\ \boldsymbol{\beta}\end{pmatrix} + \boldsymbol{\epsilon},$$

donde $\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\mathsf{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$. Muestre que los BLUE de α y $\boldsymbol{\beta}$ son independientes.

2.4 Considere las regresiones de Y sobre x para los datos a continuación, especificadas por:

$$\mathcal{M}_1 : \mathsf{E}(Y) = \beta_0 x \qquad \text{y} \qquad \mathcal{M}_2 : \mathsf{E}(Y) = \beta_1 x + \beta_2 x^2.$$

Obtenga $\widehat{\beta}_0$, $\widehat{\beta}_1$ y $\widehat{\beta}_2$. ¿Cuál de esos modelos es preferido?

Y	5	7	7	10	16	20
x	1	2	3	4	5	6

2.5 Considere las rectas de regresión:

$$\mathcal{R}_1 : Y_{1i} = \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \qquad \text{y}$$
$$\mathcal{R}_2 : Y_{2i} = \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i},$$

para $i = 1, \ldots, n$, donde los errores $\{\epsilon_{1i}\}$ y $\{\epsilon_{2i}\}$ son variables aleatorias iid con media cero y varianza común σ^2. Obtenga el estadístico F para probar la hipótesis de que las rectas de regresión $\mathcal{R}_1$ y $\mathcal{R}_2$ son paralelas.

2.6 Considere el modelo

$$Y_{ij} = \theta_i x_j + \epsilon_{ij}, \qquad i = 1, 2; j = 1, \ldots, T,$$

con $\{\epsilon_{ij}\}$ variables aleatorias independientes $\mathsf{N}(0, \sigma^2)$ y $\{x_i\}$ constantes conocidas. Obtenga el estadístico F para probar $H_0 : \theta_1 = \theta_2$.

Capítulo 3

Chequeo del Modelo y Alternativas a Mínimos Cuadrados

En este capítulo se describe la inferencia en modelos lineales. Primeramente introducimos algunas definiciones y supuestos en los que se basan los modelos de regresión.

3.1. Colinealidad

Considere el modelo

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

donde $\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\mathsf{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ con $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ tal que $\mathrm{rg}(\boldsymbol{X}) = p$. Es bien conocido que cuando $\boldsymbol{X}$ es mal condicionada, el sistema de ecuaciones

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{Y}, \tag{3.1}$$

puede ser muy inestable.

Debemos destacar que, cuando la matriz de diseño es de rango (columna) deficiente, entonces podemos considerar alguna inversa generalizada para obtener una solución del problema en (3.1). En esta sección, nos enfocaremos en el problema en que $\mathrm{rg}(\boldsymbol{X}) = p$, y sin embargo tenemos que existe $\boldsymbol{a}$ tal que $\boldsymbol{X}\boldsymbol{a} \approx \mathbf{0}$.

Observación 3.1. Este es un problema numérico que puede tener consecuencias inferenciales importantes, por ejemplo:

(a) Tipícamente los coeficientes estimados $\widehat{\boldsymbol{\beta}}$ tendrán varianzas "grandes".

(b) Test estadísticos presentarán bajo poder y los intervalos de confianza serán muy amplios.

(c) Signos de algunos coeficientes son "incorrectos"(basados en conocimiento previo).

(d) Resultados cambian bruscamente con la eliminación de una o varias columnas de $\boldsymbol{X}$.

Algunas herramientas para el diagnóstico de colinealidad, son:

(a) Examinar la *matriz de correlación* entre los regresores y la respuesta, esto es:

$$\begin{pmatrix} \boldsymbol{R}_{XX} & \boldsymbol{R}_{XY} \\ & 1 \end{pmatrix},$$

correlaciones altas entre dos variables regresoras pueden indicar un posible problema de colinealidad.

(b) *Factores de inflación de varianza*: Suponga que los datos han sido centrados y escalados, entonces

$$\boldsymbol{R}^{-1} = (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}, \qquad \widetilde{\boldsymbol{X}} = (x_{ij} - \overline{x}_j),$$

y los elementos diagonales de $\boldsymbol{R}^{-1}$ son llamados factores de inflación de varianza VIF_j. Se puede mostrar que

$$\mathsf{VIF}_j = \frac{1}{1 - R_j^2},$$

donde R_j^2 es el coeficiente de correlación multiple de $\boldsymbol{X}_j$ *"regresado"* sobre el resto de variables explicativas y de ahí que un VIF_j *"alto"* indica R_j^2 cercano a 1 y por tanto presencia de colinealidad.

(c) Examinar los valores/vectores propios (o *componentes principales*) de la matriz de correlación $\boldsymbol{R}$.

(d) *Número condición*: Desde la SVD de $\boldsymbol{X}$ podemos escribir

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

donde $\boldsymbol{U} \in \mathbb{R}^{n\times p}$, $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_p$, $\boldsymbol{D} = \operatorname{diag}(\delta_1, \ldots, \delta_p)$ y $\boldsymbol{V} \in \mathcal{O}_p$. La detección de colinealidad puede ser llevada a cabo usando

$$\kappa(\boldsymbol{X}) = \|\boldsymbol{X}\|\|\boldsymbol{X}^+\| = \frac{\delta_1}{\delta_p},$$

y $\kappa(\boldsymbol{X})$ *"grande"* ($\kappa > 30$) es un indicador de colinealidad.

Note que, el caso de *deficiencia de rango* puede ser manipulado sin problemas usando SVD. En efecto,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{U}\begin{pmatrix}\boldsymbol{D}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{V}^\top$$

donde $\boldsymbol{D}_1 \in \mathbb{R}^{r\times r}$, $\operatorname{rg}(\boldsymbol{X}) = r < p$. De este modo

$$\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\begin{pmatrix}\boldsymbol{D}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} \quad \Rightarrow \quad \boldsymbol{X}(\boldsymbol{V}_1, \boldsymbol{V}_2) = (\boldsymbol{U}_1, \boldsymbol{U}_2)\begin{pmatrix}\boldsymbol{D}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix},$$

desde donde sigue que

$$\boldsymbol{X}\boldsymbol{V}_1 = \boldsymbol{U}_1\boldsymbol{D}_1, \qquad \boldsymbol{X}\boldsymbol{V}_2 = \boldsymbol{0}.$$

Es decir, SVD permite *"detectar"* la dependencia lineal.

Una vez que hemos detectado que estamos en presencia de colinealidad, podemos sobrellevarla usando, por ejemplo, métodos de estimación sesgados. A continuación revisaremos dos procedimientos, *regresión por componentes principales* y el *estimador ridge*.

3.1.1. Regresión por componentes principales. Considere la descomposición espectral de $\boldsymbol{X}^\top\boldsymbol{X}$, dada por

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = (\boldsymbol{U}_1, \boldsymbol{U}_2)\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{U}_1^\top\\ \boldsymbol{U}_2^\top\end{pmatrix},$$

donde $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ y $\boldsymbol{\Lambda}_2 = \operatorname{diag}(\lambda_{r+1}, \ldots, \lambda_p)$, mientras que $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2)$ es matriz ortogonal. Esto lleva a la siguiente definición.

Resultado 3.2 (Estimador componentes principales). *Bajo los supuestos del modelo lineal en A1-A4⋆, el estimador componentes principales para $\boldsymbol{\beta}$ puede ser escrito como*

$$\widehat{\boldsymbol{\beta}}_r = \boldsymbol{U}_1(\boldsymbol{U}_1^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1^\top \boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top \boldsymbol{X}^\top \boldsymbol{Y}.$$

Demostración. Por la ortogonalidad de $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2)$, sigue que

$$\boldsymbol{U}_1^\top \boldsymbol{U}_1 = \boldsymbol{I}_r, \qquad \boldsymbol{U}_2^\top \boldsymbol{U}_2 = \boldsymbol{I}_{p-r}, \qquad \boldsymbol{U}_1\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{U}_2^\top = \boldsymbol{I}_p,$$

y $\boldsymbol{U}_1^\top \boldsymbol{U}_2 = \boldsymbol{0}$. Ahora,

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{U}_2^\top.$$

Usando que $\boldsymbol{U}_1^\top \boldsymbol{U}_2 = \boldsymbol{0}$ $(= \boldsymbol{U}_2^\top \boldsymbol{U}_1)$, obtenemos

$$\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2 = \boldsymbol{U}_2^\top (\boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{U}_2^\top)\boldsymbol{U}_2 = \boldsymbol{\Delta}_2^{-1}$$

De este modo, $[\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2]^{-1} = \boldsymbol{\Lambda}_2$, lo que permite escribir

$$\begin{aligned}
&(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2[\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2]^{-1}\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&\quad = (\boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{U}_2^\top)\boldsymbol{U}_2\boldsymbol{\Lambda}_2\boldsymbol{U}_2^\top (\boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{U}_2^\top) \\
&\quad = \boldsymbol{U}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{U}_2^\top.
\end{aligned}$$

Es decir,

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2[\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2]^{-1}\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top.$$

Como $\boldsymbol{U}_1^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_1 = \boldsymbol{\Lambda}_1$. Resulta

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_r &= [(\boldsymbol{X}^\top \boldsymbol{X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2[\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{U}_2]^{-1}\boldsymbol{U}_2^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}]\boldsymbol{X}^\top \boldsymbol{Y} \\
&= \boldsymbol{U}_1(\boldsymbol{U}_1^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1^\top \boldsymbol{X}^\top \boldsymbol{Y},
\end{aligned}$$

lo que concluye la prueba. □

Observación 3.3. El estimador PC es un caso particular del *estimador restringido* con respecto a:

$$\boldsymbol{U}_2^\top \boldsymbol{\beta} = \boldsymbol{0}.$$

En efecto, $\widehat{\boldsymbol{\beta}}_r$ depende del *'parámetro'* r, basta notar que

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{U}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{U}_2^\top)\boldsymbol{X}^\top \boldsymbol{Y}.$$

De este modo podemos interpretar $\widehat{\boldsymbol{\beta}}_r$ como una modificación del estimador OLS que *desconsidera* $\boldsymbol{U}_2\boldsymbol{\Lambda}_2^{-1}\boldsymbol{U}_2^\top$.

Una alternativa para seleccionar r, es utilizar el test F. Suponga r fijo y considere $H_0 : \boldsymbol{U}_2^\top \boldsymbol{\beta} = \boldsymbol{0}$. Tenemos el estadístico

$$F = \Big(\frac{n-p}{p-r}\Big)\frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_r)^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_r)}{\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}.$$

Si para un nivel α tenemos

$$F \geq F_{1-\alpha}(p-r, n-p).$$

Entonces, rechazamos H_0 y podemos seleccionar r un poco más pequeño.

OBSERVACIÓN 3.4. Sobre el estimador por componentes principales, debemos resaltar:

- No hay manera de verificar si las restricciones son satisfechas, de modo que este estimador es *sesgado*.
- Deseamos escoger r tan pequeño como posible para solucionar el problema de colinealidad y tan grande para no introducir mucho sesgo.

3.1.2. Estimador ridge. Hoerl y Kennard (1970) propusieron usar el *estimador ridge*, definido como:

$$\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}, \qquad \lambda \geq 0$$

donde k es conocido como *parámetro ridge*. Evidentemente, podemos escribir el estimador ridge en términos del estimador de mínimos cuadrados, como:

$$\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

De este modo,

$$\mathsf{E}(\widehat{\boldsymbol{\beta}}_k) = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta},$$

para $k \neq 0$, tenemos que $\widehat{\boldsymbol{\beta}}_k$ es un estimador sesgado. Mientras que su error cuadrático medio es dado por:

$$\mathsf{MSE} = \mathsf{E}\{\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}\|^2\} = \operatorname{tr}\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_k) + \|\,\mathsf{E}(\widehat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta}\|^2.$$

En efecto,

$$\begin{aligned}\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_k) &= (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\,\mathsf{Cov}(\widehat{\boldsymbol{\beta}})\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\\ &= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}.\end{aligned}$$

Además,

$$\begin{aligned}\mathsf{bias}(\widehat{\boldsymbol{\beta}}_k, \widehat{\boldsymbol{\beta}}) &= \mathsf{E}(\widehat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}\\ &= (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\big[\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})\boldsymbol{\beta}\big]\\ &= -k(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{\beta}.\end{aligned}$$

Considere la SVD de $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, de este modo $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top$, y podemos escribir

$$\begin{aligned}\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_k) &= \sigma^2\boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top \boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{V}^\top\\ &= \sigma^2\boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-2}\boldsymbol{D}^2\boldsymbol{V}^\top.\end{aligned}$$

De este modo,

$$\operatorname{tr}\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_k) = \sigma^2 \operatorname{tr}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-2}\boldsymbol{D}^2 = \sigma^2\sum_{i=1}^{p}\frac{\delta_i^2}{(\delta_i^2 + k)^2},$$

donde $\delta_1, \ldots, \delta_p$ son los valores singulares de $\boldsymbol{X}$. Finalmente,

$$\mathsf{MSE} = \sigma^2\sum_{i=1}^{p}\frac{\delta_i^2}{(\delta_i^2 + k)^2} + k^2\boldsymbol{\beta}^\top(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-2}\boldsymbol{\beta}.$$

El estimador ridge tiene varias interpretaciones interesantes, por ejemplo:

(a) Es posible caracterizar $\widehat{\boldsymbol{\beta}}_k$ como solución del problema *regularizado*:

$$\min_{\beta} Q(\boldsymbol{\beta}, k), \qquad Q(\boldsymbol{\beta}, k) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + k\,\|\boldsymbol{\beta}\|^2, \tag{3.2}$$

que puede ser expresado de forma equivalente como

$$\min_{\beta} Q(\boldsymbol{\beta}), \qquad \text{sujeto a: } \|\boldsymbol{\beta}\|^2 \leq r^2,$$

y en este contexto, k corresponde a un multiplicador de Lagrange.

(b) Considere el modelo de regresión con *datos aumentados*:

$$\boldsymbol{Y}_a = \boldsymbol{X}_a\boldsymbol{\beta} + \boldsymbol{\epsilon}_a, \qquad \boldsymbol{\epsilon}_a \sim \mathsf{N}_{n+p}(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

donde

$$\boldsymbol{Y}_a = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{X}_a = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{k}\boldsymbol{I}_p \end{pmatrix}, \qquad \boldsymbol{\epsilon}_a = \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{u} \end{pmatrix}.$$

El interés recae en escoger algún $k > 0$ tal que la matriz de diseño $\boldsymbol{X}_a$ tenga número condición $\kappa(\boldsymbol{X}_a)$ acotado.

Observación 3.5. El tipo de regularización introducida en (3.2) es conocida como *regularización de Tikhonov.*[1]

Resultado 3.6. *Suponga que los supuestos del modelo lineal en A1-A4⋆, son satisfechos. Entonces,*

$$\|\widehat{\boldsymbol{\beta}}_{k_2}\|^2 < \|\widehat{\boldsymbol{\beta}}_{k_1}\|^2,$$

siempre que $0 \leq k_1 < k_2$.

Demostración. Tenemos $\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. De este modo,

$$\|\widehat{\boldsymbol{\beta}}_k\|^2 = \widehat{\boldsymbol{\beta}}^\top\boldsymbol{M}_k\widehat{\boldsymbol{\beta}}, \qquad \boldsymbol{M}_k = \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-2}\boldsymbol{X}^\top\boldsymbol{X}.$$

Basado en la SVD de $\boldsymbol{X}$, tenemos

$$\begin{aligned} \boldsymbol{M}_k &= \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + k\boldsymbol{V}\boldsymbol{V}^\top)^{-2}\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top \\ &= \boldsymbol{V}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-2}\boldsymbol{D}^4\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Gamma}_k\boldsymbol{V}^\top, \end{aligned}$$

con

$$\boldsymbol{\Gamma}_k = \operatorname{diag}\Big(\frac{\delta_1^4}{(\delta_1^2 + k)^2}, \ldots, \frac{\delta_p^4}{(\delta_p^2 + k)^2}\Big).$$

De ahí que, si $0 \leq k_1 < k_2$, entonces

$$\boldsymbol{M}_{k_1} - \boldsymbol{M}_{k_2} \geq \boldsymbol{0},$$

lo que lleva a $\widehat{\boldsymbol{\beta}}^\top\boldsymbol{M}_{k_2}\widehat{\boldsymbol{\beta}} < \widehat{\boldsymbol{\beta}}^\top\boldsymbol{M}_{k_1}\widehat{\boldsymbol{\beta}}$, siempre que $\widehat{\boldsymbol{\beta}} \neq \boldsymbol{0}$. □

Observación 3.7. Note que $\lim_{k\to\infty} \|\widehat{\boldsymbol{\beta}}_k\|^2 = 0$ y de ahí que

$$\lim_{k\to\infty} \widehat{\boldsymbol{\beta}}_k = \boldsymbol{0}. \tag{3.3}$$

Dado que $\widehat{\boldsymbol{\beta}}_k = \boldsymbol{W}_k\widehat{\boldsymbol{\beta}}$ con $\boldsymbol{W}_k = (\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}$. La propiedad en (3.3) ha llevado a que el estimador ridge sea considerado como un *estimador shrinkage*, en cuyo caso

$$\boldsymbol{W}_k = (\boldsymbol{I}_p + k(\boldsymbol{X}^\top\boldsymbol{X})^{-1})^{-1}, \qquad k \geq 0,$$

es llamada matrix ridge-shrinking.

[1] Razón por la que k en ocasiones es llamado *parámetro de regularización.*

Se ha propuesto diversos estimadores de k, que buscan seleccionar un $\widehat{\boldsymbol{\beta}}_{k_{\mathsf{opt}}}$ que reduzca su MSE. Algunas de estas alternativas son:

(a) La propuesta de Hoerl, Kennard y Baldwin (1975), dada por

$$\widehat{k}_{\mathsf{HKB}} = \frac{ps^2}{\|\widehat{\boldsymbol{\beta}}\|^2}.$$

(b) Estimador sugerido por Lawless y Wang (1976)

$$\widehat{k}_{\mathsf{LW}} = \frac{ps^2}{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}}.$$

(c) Mientras que usando argumentos bayesianos, Lindley y Smith (1972) propusieron usar:

$$\widehat{k}_{\mathsf{LS}} = \frac{(n-p)(p+2)}{(n+2)}\frac{s^2}{\|\widehat{\boldsymbol{\beta}}\|^2}.$$

Por otro lado, Golub, Heath y Wahba (1979) han sugerido seleccionar el parámetro ridge usando *validación cruzada generalizada* (GCV), la que minimiza el criterio

$$V(k) = \frac{1}{n}\frac{\sum_{i=1}^n (Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_k)^2}{\{1 - \operatorname{tr}(\boldsymbol{H}(k))/n\}^2},$$

donde

$$\boldsymbol{H}(k) = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top.$$

Es facil notar que $\widehat{\boldsymbol{Y}}(k) = \boldsymbol{H}(k)\boldsymbol{Y}$. En este contexto se ha definido

$$\mathsf{edf} = \operatorname{tr} \boldsymbol{H}(k),$$

como el *número de parámetros efectivos*. En efecto, para $k = 0$, sigue que $\mathsf{edf} = p$.

Considere la SVD de $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ y escriba el modelo en su *forma canónica:*

$$\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{u}, \qquad \boldsymbol{u} = \boldsymbol{U}^\top \boldsymbol{\epsilon} \sim \mathsf{N}_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

donde $\boldsymbol{Z} = \boldsymbol{U}^\top \boldsymbol{Y}$, $\boldsymbol{\alpha} = \boldsymbol{V}^\top \boldsymbol{\beta}$. De este modo, es fácil notar que

$$\widehat{\boldsymbol{\alpha}}_k = (\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{Z},$$

es decir,

$$\widehat{\alpha}_{k,j} = \frac{\delta_j z_j}{\delta_j^2 + k}, \qquad j = 1, \ldots, p.$$

Por otro lado,

$$\begin{aligned}
\mathsf{edf} = \operatorname{tr} \boldsymbol{H}(k) &= \operatorname{tr} \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + k\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \\
&= \operatorname{tr} \boldsymbol{U}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}^2\boldsymbol{U}^\top = \operatorname{tr}(\boldsymbol{D}^2 + k\boldsymbol{I})^{-1}\boldsymbol{D}^2 \\
&= \sum_{j=1}^p \frac{\delta_j^2}{\delta_j^2 + k}.
\end{aligned}$$

Además,

$$\widehat{\boldsymbol{Y}}(k) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \widehat{\boldsymbol{\beta}}_k = \boldsymbol{U}\boldsymbol{D}\widehat{\boldsymbol{\alpha}}_k.$$

Lo que permite escribir

$$V(k) = \frac{1}{n}\frac{\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k\|^2}{\{\operatorname{tr}(\boldsymbol{I} - \boldsymbol{H}(k))/n\}^2} = \frac{\|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{D}\widehat{\boldsymbol{\alpha}}_k\|^2/n}{(1 - \mathsf{edf}/n)^2}.$$

Observación 3.8. Las consideraciones anteriores ofrecen un procedimiento sencillo para evaluar $V(k)$.

En términos del modelo canónico, también podemos escribir:

$$\widehat{k}_{\mathsf{HKB}} = \frac{ps^2}{\|\widehat{\boldsymbol{\alpha}}\|^2}, \qquad \widehat{k}_{\mathsf{LW}} = \frac{ps^2}{\|\boldsymbol{Z}\|^2},$$

con $\widehat{\boldsymbol{\alpha}} = \boldsymbol{D}^{-1}\boldsymbol{Z}$ $(= \widehat{\boldsymbol{\alpha}}_0)$, $\boldsymbol{Z} = \boldsymbol{U}^\top \boldsymbol{Y}$ y $s^2 = \|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{D}\widehat{\boldsymbol{\alpha}}\|^2/(n-p)$.

Ejemplo 3.9. Woods, Steinour y Starke (1932) consideraron datos desde un estudio experimental relacionando la emisión de calor durante la producción y endurecimiento de 13 muestras de *cementos Portland*. Este estudio se enfoca en los cuatro compuestos para los clinkers desde los que se produce el cemento. La respuesta (Y) es la *emisión de calor* después de 180 días de curado, medido en calorías por gramo de cemento. Los regresores son los porcentajes de los cuatro compuestos principales: *aluminato tricálcico* (X_1), *silicato tricálcico* (X_2), *ferrito aluminato tetracálcico* (X_3) y *silicato dicálcico* (X_4).

Siguiendo a Woods, Steinour y Starke (1932) consideramos un modelo lineal sin intercepto (modelo homogéneo), cuyo número condición escalado es $\kappa(\boldsymbol{X}) = 9.432$, esto es, $\boldsymbol{X}$ es bien condicionada (mientras que para las variables centradas, obtenemos $\kappa(\widetilde{\boldsymbol{X}}) = 37.106$).

Por otro lado, Hald (1952), Gorman y Toman (1966) y Daniel y Wood (1980) adoptan un modelo con intercepto (modelo no homogéneo). En cuyo caso $\kappa(\boldsymbol{X}) = 249.578$, sugiriendo la presencia de colinealidad. El aumento en el número condición se debe a que existe una relación lineal aproximada, pues

$$x_1 + x_2 + x_3 + x_4 \approx 100,$$

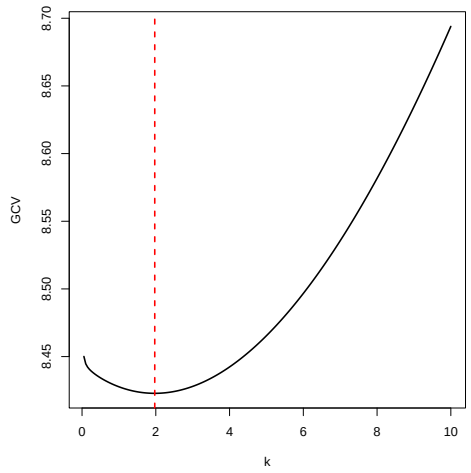
de modo que incluir el intercepto causa una colinealidad severa. Podemos usar la rutina para regresión ridge disponible en la biblioteca fastmatrix (Osorio y Ogueda, 2021)

```
# carga biblioteca 'fastmatrix' y base de datos
> library(fastmatrix)
> load("portland.rda")

# ajuste usando regresión ridge
> ridge(y ~ x1 + x2 + x3 + x4, data = portland, lambda = 10,
+       method = "grid")

Call:
ridge(formula = y ~ x1 + x2 + x3 + x4, data = portland,
      lambda = 10, method = "grid")
Coefficients:
(Intercept)          x1          x2          x3          x4
    0.08568     2.16549     1.15860     0.73845     0.48948

Optimal ridge parameter: 1.9598
Number of observations: 13
Effective number of parameters: 3.9796
Scale parameter estimate: 4.0553
```

La función `ridge` desde el paquete fastmatrix, permite también obtener los estimadores sugeridos por Hoerl, Kennard y Baldwin (1975) y Lawless y Wang (1976). Para el caso de los datos de cemento, tenemos:

$$\widehat{k}_{\mathsf{HKB}} = 0.0077, \qquad \widehat{k}_{\mathsf{LW}} = 0.0032,$$

mientras que el valor óptimo de k obtenido mediante minimizar el criterio de validación cruzada generalizada, es $\widehat{k}_{\mathsf{opt}} = 1.9598$. En la siguiente figura, se presenta evaluación de la función $V(k)$ para una grilla de valores de k. Se ha indicado el mínimo $\widehat{k}_{\mathsf{opt}}$ como una línea segmentada en color rojo,



Adicionalmente, es interesante llevar a cabo una comparativa usando el estimador mínimos cuadrados en el modelo sin intercepto. Considere la siguiente tabla de resumen de estimación:

Parámetro	Estimación OLS		Estimación ridge		
	homogéneo	no homogéneo	HKB	LW	GCV
β_0	—	62.4054	8.5870	17.1889	0.0855
β_1	2.1930	1.5511	2.1046	2.0162	2.1653
β_2	1.1533	0.5102	1.0648	0.9762	1.1586
β_3	0.7585	0.1019	0.6681	0.5776	0.7383
β_4	0.4863	-0.1441	0.3996	0.3127	0.4895
σ^2	4.0469	3.6818	4.0005	3.9478	5.0902
k	—	—	0.0077	0.0032	1.9716
edf	4.0000	5.0000	4.1369	4.2749	3.9795
κ	9.4325	249.5783	92.4131	130.8854	10.7852

Desde la tabla se aprecia que la elección del parámetro ridge usando el método de validación cruzada lleva a un número condición bastante pequeño, y en efecto, los resultados de estimación son muy cercanos al modelo homogéneo.

3.2. Errores correlacionados

El objetivo de esta sección es revisar el supuesto de *homogeneidad de varianzas*. De este modo, nos enfocaremos en el siguiente modelo:

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

donde $\mathsf{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$, y

$$\mathsf{Cov}(\boldsymbol{\epsilon}) = \mathsf{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2\boldsymbol{\Omega}, \qquad \boldsymbol{\Omega} > \boldsymbol{0}.$$

Primeramente vamos a suponer que $\boldsymbol{\Omega}$ es conocida y sea $\boldsymbol{\Omega} = \boldsymbol{B}\boldsymbol{B}^\top$, con $\boldsymbol{B}$ matriz no singular $n \times n$. Note que

$$\boldsymbol{B}^{-1}\boldsymbol{Y} = \boldsymbol{B}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{-1}\boldsymbol{\epsilon}.$$

haciendo $\boldsymbol{Y}_* = \boldsymbol{B}^{-1}\boldsymbol{Y}$, $\boldsymbol{X}_* = \boldsymbol{B}^{-1}\boldsymbol{X}$ y $\boldsymbol{\epsilon}_* = \boldsymbol{B}^{-1}\boldsymbol{\epsilon}$. Entonces $\mathsf{E}(\boldsymbol{\epsilon}_*) = \boldsymbol{0}$, y

$$\mathsf{Cov}(\boldsymbol{\epsilon}_*) = \boldsymbol{B}^{-1}\,\mathsf{Cov}(\boldsymbol{\epsilon})\boldsymbol{B}^{-\top} = \sigma^2\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{B}^{-\top} = \sigma^2\boldsymbol{I}.$$

Es decir, el modelo transformado

$$\boldsymbol{Y}_* = \boldsymbol{X}_*\boldsymbol{\beta} + \boldsymbol{\epsilon}_*,$$

satisface las condiciones A1-A4. Así,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}} &= (\boldsymbol{X}_*^\top\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^\top\boldsymbol{Y}_* \\
&= (\boldsymbol{X}^\top\boldsymbol{B}^{-\top}\boldsymbol{B}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{B}^{-\top}\boldsymbol{B}^{-1}\boldsymbol{Y} \\
&= (\boldsymbol{X}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{B}\boldsymbol{B}^\top)^{-1}\boldsymbol{Y} \\
&= (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{Y}.
\end{aligned}$$

Es fácil mostrar que

$$\begin{aligned}
\mathsf{E}(\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}) &= (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\,\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{\beta} \\
\mathsf{Cov}(\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}) &= \sigma^2(\boldsymbol{X}_*^\top\boldsymbol{X}_*)^{-1} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}.
\end{aligned}$$

Además, podemos definir el vector de residuos,

$$\boldsymbol{e}_* = \boldsymbol{Y}_* = \boldsymbol{X}_*\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}} = \boldsymbol{B}^{-1}\boldsymbol{Y} - \boldsymbol{B}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}} = \boldsymbol{B}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}),$$

lo que permite escribir la suma de cuadrados de residuos, como:

$$\begin{aligned}
Q_\Omega(\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}) = \|\boldsymbol{e}_*\|^2 &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}})^\top\boldsymbol{B}^{-\top}\boldsymbol{B}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}) \\
&= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}).
\end{aligned}$$

Adicionalmente, podemos escribir:

$$\begin{aligned}
\boldsymbol{e}_* &= \boldsymbol{B}^{-1}\boldsymbol{Y} - \boldsymbol{B}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{Y} \\
&= \boldsymbol{B}^{-1}\boldsymbol{Y} - \boldsymbol{B}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{B}^{-\top}\boldsymbol{B}^{-1}\boldsymbol{Y},
\end{aligned}$$

de ahí que

$$\boldsymbol{e}_* = (\boldsymbol{I} - \boldsymbol{H}_\Omega)\boldsymbol{Y}_*,$$

con

$$\boldsymbol{H}_\Omega = \boldsymbol{B}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{B}^{-\top}.$$

Observación 3.10. Evidentemente $\boldsymbol{H}_\Omega^\top = \boldsymbol{H}_\Omega$ y $\boldsymbol{H}_\Omega^2 = \boldsymbol{H}_\Omega$. De este modo la suma de cuadrados residual adopta la forma:

$$Q_\Omega(\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}) = \|\boldsymbol{e}_*\|^2 = \boldsymbol{Y}_*^\top(\boldsymbol{I} - \boldsymbol{H}_\Omega)\boldsymbol{Y}_*.$$

Desafortunadamente, en general la matriz $\boldsymbol{\Omega}$ no es conocida y requiere ser estimada. Si $\widehat{\boldsymbol{\Omega}}$ es un estimador de $\boldsymbol{\Omega}$, entonces

$$\widehat{\boldsymbol{\beta}}_{\mathsf{EGLS}} = (\boldsymbol{X}^\top \widehat{\boldsymbol{\Omega}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{\Omega}}^{-1} \boldsymbol{Y}.$$

Se debe notar que las propiedades de $\widehat{\boldsymbol{\beta}}_{\mathsf{EGLS}}$ son difíciles de caracterizar.

Un caso particular importante, corresponde a *mínimos cuadrados ponderados (WLS)* en cuyo caso $\mathsf{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{W}^{-1}$ donde $\boldsymbol{W} = \mathrm{diag}(\omega_1, \ldots, \omega_n)$ con $\omega_i > 0$, $\forall\, i$. De este modo,

$$\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Y}.$$

Note que este estimador es solución del problema

$$\min_{\beta} \; Q_W(\boldsymbol{\beta}), \qquad Q_W(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Bajo el supuesto $\boldsymbol{\epsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{W}^{-1})$, con $\boldsymbol{W} = \mathrm{diag}(\omega_1, \ldots, \omega_n)$. Es fácil notar que el estimador $\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}}$ minimiza la función

$$Q_W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \omega_i (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2.$$

Además, el estimador $\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}}$ resuelve las siguientes ecuaciones de estimación

$$\boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0},$$

o equivalentemente

$$\sum_{i=1}^{n} \omega_i (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i = \boldsymbol{0}.$$

Mientras que el estimador ML para σ^2 asume la forma:

$$\widehat{\sigma}^2_{\mathsf{WLS}} = \frac{1}{n} \sum_{i=1}^{n} \omega_i (Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\mathsf{WLS}})^2.$$

Observación 3.11. El problema anterior puede ser resuelto usando OLS en el siguiente 'problema modificado'

$$\boldsymbol{W}^{1/2} \boldsymbol{Y}, \qquad \boldsymbol{W}^{1/2} \boldsymbol{X}.$$

Consideraciones computacionales sobre GLS y WLS son dadas por ejemplo en el Capítulo 4 de Björck (1996).

3.2.1. Estimación de funciones de varianza. El objetivo de esta sección es considerar modelos de regresión heterocedásticos, tal que

$$\mathsf{E}(Y_i) = \mu_i, \qquad \mathsf{var}(Y_i) = \sigma^2 g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi}), \qquad i = 1, \ldots, n,$$

donde $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $\boldsymbol{x}_i$ y $\boldsymbol{z}_i$ representan vectores de covariables (que podrían ser iguales), $\boldsymbol{\beta}$ son coeficientes de regresión, g es función de varianza (que permite modelar la heterogeneidad), $\sigma^2 > 0$ y $\boldsymbol{\phi}$ son parámetros de escala desconocidos.

Ejemplo 3.12. Considere

$$\mathsf{var}(Y_i) = \sigma^2 \{g(\boldsymbol{x}_i; \boldsymbol{\beta})\}^{2\phi}, \qquad \phi > 0.$$

Si suponemos $g(\boldsymbol{x}_i; \boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ y $\phi = \frac{1}{2}$ tenemos la *estructura de varianza Poisson*, mientras que $\phi = 1$ es de *tipo-gama*.

Ejemplos habituales de funciones de varianza son los siguientes:

(a) Función de varianza cuadrática

$$\sigma g(\boldsymbol{z}_i; \mu_i, \phi) = 1 + \phi_1 z_{1i} + \phi_2 z_{2i}^2.$$

(b) Modelo potencia extendido

$$\mathsf{var}(Y_i) = \sigma^2(\phi_1 + \phi_2 \mu_i^{\phi_3}).$$

(c) Modelo exponencial

$$\mathsf{var}(Y_i) = \sigma^2 \exp(2\phi\mu_i).$$

(d) También puede depender de $\boldsymbol{\phi}$ según un predictor lineal

$$\mathsf{var}(Y_i) = \sigma^2 \exp(2\boldsymbol{z}_i^\top \boldsymbol{\phi}).$$

Primeramente, supondremos el modelo

$$\mathsf{E}(Y_i) = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \qquad \mathsf{var}(Y_i) = \sigma^2 g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi}),$$

con $\boldsymbol{\phi}$ *conocido*. El método WLS sugiere un procedimiento natural para la estimación de varianzas heterogéneas usando una estratégia de *mínimos cuadrados iterativamente ponderados (IWLS)*.

Algoritmo 4: IWLS para estimación de varianzas

begin

Considerar una estimación inicial para $\boldsymbol{\beta}$, digamos $\boldsymbol{\beta}^{(0)}$, resolviendo

$$\sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i = \mathbf{0}.$$

y hacemos $k \leftarrow 0$.

Construir *pesos*

$$\omega_i^{(k)} = 1/g^2(\boldsymbol{z}_i; \mu_i^{(k)}, \boldsymbol{\phi}), \qquad \mu_i^{(k)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)}.$$

Actualizar $\boldsymbol{\beta}^{(k+1)}$, resolviendo

$$\sum_{i=1}^{n} \omega_i^{(k)} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i = \mathbf{0},$$

hacer $k \leftarrow k + 1$ y volver a Paso 3.

end

Evidentemente el Paso 4 del Algoritmo 4, debe ser resuelta usando mínimos cuadrados ponderados.

En el algoritmo anterior σ^2 puede ser estimado por analogía a WLS. Específicamente, podemos considerar

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\omega}_i (Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})^2,$$

donde $\widehat{\omega}_i$ $(i = 1, \ldots, n)$ y $\widehat{\boldsymbol{\beta}}$ representan los valores de ω_i y $\boldsymbol{\beta}$ a la convergencia del Algoritmo 4.

Alternativamente, es posible incorporar una etapa adicional al Paso 4 del Algoritmo 4 como:

$$\sigma^{2\,(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \omega_i^{(k)} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k+1)})^2.$$

Claramente, no siempre es posible especificar valores para $\boldsymbol{\phi}$. Para simplificar la exposición considere

$$Y_i \overset{\text{ind}}{\sim} \mathsf{N}(\mu_i, \sigma^2 g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi})), \qquad i = 1, \ldots, n,$$

con $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. Defina la matriz diagonal

$$\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \text{diag}\left(g^2(\boldsymbol{z}_1; \mu_1, \boldsymbol{\phi}), \ldots, g^2(\boldsymbol{z}_1; \mu_n, \boldsymbol{\phi})\right).$$

De este modo, el modelo anterior puede ser escrito como

$$\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{G}),$$

con función de densidad conjunta

$$f(\boldsymbol{y}) = |2\pi\sigma^2 \boldsymbol{G}|^{-1/2} \exp\Big\{ -\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{G}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \Big\}.$$

La parte relevante de la función de log-verosimilitud es

$$\ell_n(\boldsymbol{\theta}) = -\frac{1}{2} \log |\sigma^2 \boldsymbol{G}| - \frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{G}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

con $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\phi}^\top)^\top$. Por la estructura diagonal de $\boldsymbol{G}$ tenemos que

$$\boldsymbol{G}^{-1} = \text{diag}\left(g^{-2}(\boldsymbol{z}_1; \mu_1, \boldsymbol{\phi}), \ldots, g^{-2}(\boldsymbol{z}_1; \mu_n, \boldsymbol{\phi})\right),$$

$$\log |\sigma^2 \boldsymbol{G}| = \sum_{i=1}^{n} \log \sigma^2 g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi}).$$

Esto permite escribir la función de log-verosimilitud como:

$$\begin{aligned} \ell_n(\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{i=1}^{n} \log \sigma^2 g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi})} \\ &= -\frac{1}{2}\sum_{i=1}^{n} \Big\{ \frac{(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{\sigma^2 g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi})} + \log \sigma^2 g^2(\boldsymbol{z}_i; \mu_i, \boldsymbol{\phi}) \Big\} \end{aligned}$$

La estimación de parámetros se puede desarrollar alternando las siguientes dos etapas:

1. Para una estimación preliminar $\boldsymbol{\beta}^{(k)}$ de $\boldsymbol{\beta}$ minimizar con relación a $\boldsymbol{\phi}$ y σ^2, la función de *log-verosimilitud perfilada*:

$$\ell_*(\boldsymbol{\beta}^{(k)}, \sigma^2, \boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n} \Big\{ \frac{(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)})^2}{\sigma^2 g^2(\boldsymbol{z}_i; \mu_i^{(k)}, \boldsymbol{\phi})} + \log \sigma^2 g^2(\boldsymbol{z}_i; \mu_i^{(k)}, \boldsymbol{\phi}) \Big\},$$

con $\mu_i^{(k)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)}$. Diferenciando con relación a $\boldsymbol{\phi}$ y σ^2 lleva a las ecuaciones de estimación:

$$\sum_{i=1}^{n} \frac{1}{g^4(\boldsymbol{z}_i; \mu_i^{(k)}, \boldsymbol{\phi})} \{ r_i^2 - \sigma^2 g^2(\boldsymbol{z}_i; \mu_i^{(k)}, \boldsymbol{\phi}) \}\, \boldsymbol{q}(\mu_i^{(k)}, \sigma, \boldsymbol{\phi}) = \boldsymbol{0}, \tag{3.4}$$

donde

$$\boldsymbol{q}(\mu_i^{(k)}, \sigma, \boldsymbol{\phi}) = g^2(\mu_i^{(k)}, \boldsymbol{\phi}) \begin{pmatrix} 1/\sigma \\ \boldsymbol{u}(\mu_i^{(k)}, \boldsymbol{\phi}) \end{pmatrix},$$

y $\boldsymbol{u}(\mu_i^{(k)}, \boldsymbol{\phi})$ representa el vector de derivadas de $\log g(\boldsymbol{z}_i, \mu_i, \boldsymbol{\phi})$ con relación a $\boldsymbol{\phi}$, mientras que $r_i = Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)}$.

2. Actualizar $\boldsymbol{\beta}^{(k+1)}$ como la solución del problema mínimos cuadrados ponderados

$$\boldsymbol{X}^\top \boldsymbol{G}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0},$$

donde $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\beta}^{(k)}, \boldsymbol{\phi}^{(k+1)})$.

Inspección de la Ecuación (3.4) permite notar que la estimación de σ y $\boldsymbol{\phi}$ corresponde a WLS con *respuesta* r_i^2, *función de regresión* $\sigma^2 g^2(\boldsymbol{z}_i; \mu_i^{(k)}, \boldsymbol{\phi})$, *parámetros de regresión* $(\sigma, \boldsymbol{\phi}^\top)^\top$, *pesos* $g^{-4}(\boldsymbol{z}_i, \mu_i^{(k)}, \boldsymbol{\phi})$ y gradiente (matriz de diseño cuyas filas están dadas por) $\boldsymbol{q}(\mu_i^{(k)}, \sigma, \boldsymbol{\phi})$.

Detalles sobre las propiedades estadísticas del estimador obtenido por este procedimiento pueden ser hallados en Davidian y Carroll (1987) (ver también Carroll y Ruppert, 1988).

3.3. Transformaciones estabilizadoras de varianza

Para introducir ideas, considere

$$\mathsf{E}(Y) = \mu, \qquad \mathsf{var}(Y) = \sigma^2 h(Y),$$

y suponga la transformación $z = g(y)$ tal que la varianza de Z es aproximadamente independiente de μ. Adicionalmente, considere una expansión de primer orden de $g(y)$ en torno de μ, esto es

$$g(y) \approx g(\mu) + g'(\mu)(y - \mu).$$

De este modo,

$$\begin{aligned} \mathsf{E}(Z) &\approx g(\mu) \\ \mathsf{var}(Z) &\approx \{g'(\mu)\}^2 \, \mathsf{var}(Y - \mu) = \{g'(\mu)\}^2 \sigma^2 h(y) \end{aligned}$$

Así, para determinar una transformación tal que $\mathsf{var}(Z) = \sigma^2$ necesitamos que

$$g'(\mu) = \frac{1}{\sqrt{h(y)}},$$

o de forma análoga,

$$g(\mu) = \int \frac{\mathsf{d}\,\mu}{\sqrt{h(\mu)}}.$$

En particular, se podría considerar la clase de transformaciones en que $h(y)$ es una potencia de μ. Algunos ejemplos son los siguientes:

$h(\mu)$	z	Descripción
μ^4	$1/y$	recíproco
μ^2	$\log y$	logarítmico
μ	$\sqrt{y}$	raíz cuadrada
$\mu(1-\mu)$	$\sin^{-1}(\sqrt{y})$	seno inverso
$(1-\mu^2)^2$	$\log(\frac{1+y}{1-y})$	correlación

Box y Cox (1964) sugirieron llevar a cabo la estimación ML para una clase general de transformaciones. El supuesto fundamental es que $Y > 0$ y que existe alguna

potencia de Y tal que su varianza es aproximadamente constante y satisface el supuesto de normalidad. En concreto, ellos consideraron la familia de transformaciones

$$Y(\lambda) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(Y), & \lambda = 0, \end{cases}$$

el objetivo es obtener una estimación de λ desde los datos observados.

Sea $U_i = Y_i(\lambda)$ para $i = 1, \ldots, n$, y asuma que $\boldsymbol{U} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ para alguna elección de λ. El Jacobiano de la transformación es dado por

$$J = \prod_{i=1}^n \frac{\partial Y_i(\lambda)}{\partial Y_i} = \prod_{i=1}^n Y_i^{\lambda-1} = \Big(\prod_{i=1}^n Y_i^{1/n}\Big)^{n(\lambda-1)} = G^{n(\lambda-1)},$$

donde G es la media geométrica de las observaciones. De este modo, la log-verosimilitud requerida adopta la forma:

$$\ell_n(\boldsymbol{\theta}, \lambda) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} Q_\lambda(\boldsymbol{\beta}) + \log J,$$

donde

$$Q_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

Para λ fijado, tenemos

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}(\lambda), \qquad \widehat{\sigma}^2(\lambda) = \frac{1}{n}\|\boldsymbol{Y}(\lambda) - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\|^2.$$

Esto lleva a la función de log-verosimilitud perfilada, dada por

$$\ell_*(\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log(\mathsf{RSS}(\lambda)/n) - \frac{n}{2} + \log J,$$

donde

$$\mathsf{RSS}(\lambda) = \boldsymbol{Y}^\top(\lambda)(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}(\lambda).$$

Considere por conveniencia,

$$\boldsymbol{Z}(\lambda) = \frac{\boldsymbol{Y}(\lambda)}{G^{\lambda-1}},$$

de este modo podemos escribir

$$\ell_*(\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log(\mathsf{RSS}_Z(\lambda)/n) - \frac{n}{2},$$

con

$$\mathsf{RSS}_Z(\lambda) = \boldsymbol{Z}^\top(\lambda)(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}(\lambda),$$

luego la maximización de $\ell_*(\lambda)$ es equivalente a la minimización de $\mathsf{RSS}_Z(\lambda)$.

Típicamente se realiza la transformación $Z(\lambda)$ para un rango de valores de λ, se realiza el ajuste del modelo lineal y luego se examina aquél valor de λ que corresponde al valor más pequeño de $\mathsf{RSS}_Z(\lambda)$[2]

Para los datos transformados la función Box-Cox asume la forma:

$$Z(\lambda) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda\, G^{\lambda-1}}, & \lambda \neq 0, \\ G\log(Y), & \lambda = 0, \end{cases}$$

obteniendo el valor $\widehat{\lambda}$ se lleva a cabo el ajuste $\mathsf{E}(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{\beta}$ con $\boldsymbol{Z} = \boldsymbol{Z}(\widehat{\lambda})$.

[2]Usualmente basta considerar $-2 \leq \lambda \leq 2$ con incrementos no muy pequeños.

OBSERVACIÓN 3.13. Note que $\lambda = 1$ implica que el modelo no debe ser transformado.

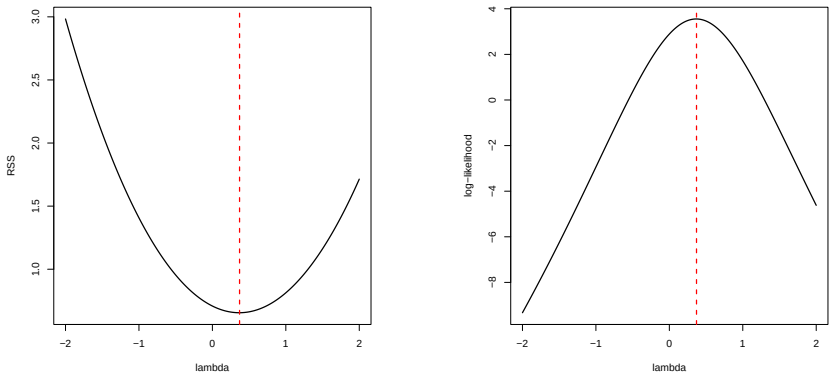
Para llevar a cabo la estimación de parámetros usando el procedimiento de Box-Cox, podemos considerar el siguiente fragmento de código en R:

```
boxcox.lm <- function(x, y, lambda) {
  boxcox <- function(y, lambda) {
    n <- length(y)
    lambda <- rep(lambda, n)
    z <- ifelse(lambda != 0., (y^lambda - 1.) / lambda, log(y))
    z
  }
  n <- nrow(x)
  p <- ncol(x)
  k <- length(lambda)
  RSS <- rep(0, k)
  logLik <- rep(0, k)
  for (i in 1:k) {
    geom <- geomean(y)
    z <- boxcox(y, lambda = lambda[i])
    z <- z / geom^(lambda[i] - 1.)
    fm <- ols.fit(x, z, method = "sweep")
    RSS[i] <- fm$RSS
    logLik[i] <- -.5 * n * log(2 * pi)
                 - .5 * n * log(RSS[i] / n) - .5 * n
  }
  idx <- order(RSS)[1]
  opt <- lambda[idx]
  obj <- list(lambda = lambda, RSS = RSS, logLik = logLik,
              opt = opt)
  obj
}
```

EJEMPLO 3.14. Para los datos de Forbes (ver Weisberg, 2005), llevamos a cabo selección de $\lambda \in [-2, 2]$ usando la transformación Box-Cox. Considere,

```
# carga datos desde biblioteca MASS
> library(MASS)
> data(forbes)

# ajuste preliminar
> library(fastmatrix)
> fm <- ols(pres ~ bp, data = forbes, x = TRUE, y = TRUE)

# extrae vector de respuestas y matriz de diseno
> x <- fm$x
> y <- fm$y

# interpreta script, crea 'grilla' e invoca método de ajuste
> source("boxcox.lm.R")
> lambda <- seq(-2, 2, by = 0.01)
> z <- boxcox.lm(x, y, lambda)
```

El valor óptimo basado en una grilla de valores para $\lambda \in \{-2.00, -1.99, \ldots, 1.99, 2.00\}$, se obtuvo $\lambda_{\mathsf{opt}} = 0.37$. Los siguientes gráficos presentan las funciones $\mathsf{RSS}_Z(\lambda)$ y la log-verosimilitud perfilada:



(a) $\mathsf{RSS}_Z(\lambda)$ (b) $\ell_*(\lambda)$

Adicionalmente, podemos realizar el análisis considerando algunos valores para λ así como eliminado la observación 12, en cuyo caso se obtuvo $\widetilde{\lambda}_{\mathsf{opt}} = 0.10$. Considere el siguiente fragmento de código en R y el resumen de estimación:

```
# gráficos de RSS y log-likelihood perfilada
> plot(lambda, z$RSS, type = "l", ylab = "RSS", lwd = 2)
> abline(v = z$opt, col = "red", lwd = 2, lty = 2)
> plot(lambda, z$logLik, type = "l", ylab = "log-likelihood")
> abline(v = z$opt, col = "red", lwd = 2, lty = 2)

# removiendo dato 12
> y12 <- y[-12]
> x12 <- x[-12,]
> z <- boxcox.lm(x12, y12, lambda)

# ajustando diversos modelos
f0 <- ols(pres ~ bp, data = forbes)
f1 <- ols(log(pres) ~ bp, data = forbes)
f2 <- ols((pres^.37 - 1) / .37 ~ bp, data = forbes)
f3 <- ols((pres^.10 - 1) / .10 ~ bp, data = forbes)
```

Parámetro	λ			
	—	0.00	0.10	0.37
β_0	-81.0637	-0.9709	-1.9885	-7.6462
β_1	0.5229	0.0206	0.0285	0.0681
σ^2	0.0542	0.0001	0.0001	0.0008
RSS	0.8131	0.0011	0.0021	0.0114
$\ell(\widehat{\boldsymbol{\theta}})$	1.7186	57.5378	52.3791	37.9802

Es decir, el mejor modelo corresponde a la transformación Box-Cox con $\lambda = 0$ (esto es la transformación logarítmica). Adicionalmente es instructivo llevar a cabo el test de razón de verosimilitudes $H_0 : \lambda = 1$.

3.4. Análisis de residuos y leverages

Suponga el modelo lineal,

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

con los Supuestos A1-A4. El vector de residuos es dado por:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y},$$

con $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, es decir $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ $(= \widehat{\mathsf{E}}(\boldsymbol{Y}))$. Bajo el supuesto de normalidad $\boldsymbol{Y} \sim \mathsf{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, tenemos

$$\boldsymbol{e} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})),$$

es decir

$$\mathsf{E}(e_i) = 0, \quad \mathsf{var}(e_i) = \sigma^2(1 - h_{ii}), \quad \mathsf{Cov}(e_i, e_j) = -\sigma^2 h_{ij}.$$

De ahí que los residuos tienen *varianzas diferentes* y *son correlacionados*. A continuación se introducirá definiciones de residuos estandarizados. Primeramente, considere que σ^2 es conocido, de este modo

$$z_i = \frac{e_i}{\sigma} \sim \mathsf{N}(0, 1).$$

De este modo, podemos definir el *residuo estandarizado* como:

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \qquad i = 1, \ldots, n.$$

Cook y Weisberg (1982) mostraron que

$$\frac{r_i^2}{n - p} \sim \mathsf{Beta}\Big(\frac{1}{2}, \frac{n - p - 1}{2}\Big).$$

Esto nos permite notar que,

$$\mathsf{E}(r_i) = 0, \quad \mathsf{var}(r_i) = 1, \quad \mathsf{Cov}(r_i, r_j) = \frac{-h_{ij}}{\sqrt{(1 - h_{ii})(1 - h_{jj})}}.$$

Por otro lado, considere el *residuo studentizado*:

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}, \qquad i = 1, \ldots, n.$$

donde

$$s^2_{(i)} = \frac{1}{n - p - 1}\sum_{j \neq i}^{n}(y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_{(i)})^2,$$

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)},$$

denotan los estimadores de σ^2 y $\boldsymbol{\beta}$ una vez que la i-ésima observación ha sido eliminada. Es decir basados en las siguientes matrices de datos,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(i)} \\ \boldsymbol{x}_i^\top \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{(i)} \\ Y_i \end{pmatrix}$$

Una interpretación interesante de t_i es que corresponde al estadístico t para probar la hipótesis $H_0 : \gamma = 0$ en el *modelo de salto en la media*, dado por:

$$Y_j = \boldsymbol{x}_j^\top \boldsymbol{\beta} + d_j\gamma + \epsilon_j, \qquad j = 1, \ldots, n,$$

donde $d_j = 1$ si $j = i$ y 0 en caso contrario.

El modelo puede ser escrito como:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{d}_i\gamma + \boldsymbol{\epsilon},$$

con $\boldsymbol{d}_i = (\boldsymbol{0}, 1, \boldsymbol{0})^\top$ con un cero en la i-ésima posición.

Lo anterior permite notar que $t_i \sim t(n-p-1)$, y de este modo,

$$\mathsf{E}(t_i) = 0, \qquad \mathsf{var}(t_i) = \frac{n-p-1}{n-p-3} \approx 1.$$

Es decir, los residuos estandarizados y estudentizados tienen propiedades similares a los errores $\{\epsilon_1, \ldots, \epsilon_n\}$.

Basado en la propiedad $\mathsf{Cov}(\boldsymbol{e}, \widehat{\boldsymbol{Y}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$ se ha sugerido el diagrama de residuos versus valores predichos. Este tipo de herramientas gráficas permite verificar desvios evidentes del modelo. Adicionalmente, se ha propuesto la construcción de gráficos cuantil-cuantil (QQ-plot) con *envelopes* para evaluar el supuesto de normalidad (consultar Atkinson, 1985, para más detalles).
Por otro lado, sabemos que

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}, \qquad \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \tag{3.5}$$

de ahí sigue que,

$$\widehat{Y}_i = \sum_{j=1}^{n} h_{ij}Y_j = h_{ii}Y_i + \sum_{j\neq i} h_{ij}Y_j, \tag{3.6}$$

es decir el valor predicho es una combinación lineal de las respuestas observadas con pesos dados por los elementos de la matriz de proyección $\boldsymbol{H}$. A continuación llamaremos a los elementos diagonales h_{ii}, $i = 1, \ldots, n$, como *leverages*. A continuación revisamos algunas de las propiedades fundamentales de la matriz $\boldsymbol{H}$:

PROPIEDAD 3.15. $\boldsymbol{H}$ es simétrica e idempotente con $\mathrm{rg}(\boldsymbol{H}) = \mathrm{tr}(\boldsymbol{H}) = p$.

PROPIEDAD 3.16. Los elementos diagonales de $\boldsymbol{H}$ están acotados, en efecto:

$$0 \leq h_{ii} \leq 1, \qquad i = 1, \ldots, n.$$

DEMOSTRACIÓN. Sabemos que

$$\widehat{\boldsymbol{Y}} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H}), \qquad \boldsymbol{e} \sim \mathsf{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})).$$

De este modo,

$$\mathsf{var}(\widehat{Y}_i) = \sigma^2 h_{ii}, \qquad \mathsf{var}(e_i) = \sigma^2(1 - h_{ii}),$$

de ahí sigue el resultado.

Para otra demostración, ver Resultado A.6 desde el Apéndice A. □

PROPIEDAD 3.17. Tenemos,

$$h_{ii} = \boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i, \qquad i = 1, \ldots, n.$$

Además,

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\,\mathrm{tr}(\boldsymbol{H}) = \frac{p}{n}.$$

PROPIEDAD 3.18. Sea $\widetilde{\boldsymbol{X}}$ la matriz de datos centrados. En este caso, los elementos diagonales de $\widetilde{\boldsymbol{H}}$ están dados por

$$\widetilde{h}_{ii} = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \qquad i = 1, \ldots, n.$$

Luego $\widetilde{h}_{ii}$ es la distancia ponderada desde $\boldsymbol{x}_i$ al *centroide* $\overline{\boldsymbol{x}}$.

Hoaglin y Welsch (1978) sugirieron que aquellas observaciones que exceden dos veces su promedio

$$h_{ii} > 2p/n \quad (= 2\overline{h})$$

indican un alto leverage. Mientras que Huber (1981) sugirió identificar observaciones tal que

$$h_{ii} > 0.5,$$

independiente de n o p. En la práctica se debe prestar atención a casos inusualmente grandes *con relación al resto* de h_{ii}'s.

PROPIEDAD 3.19. Desde Ecuación (3.5), sigue que

$$\frac{\partial \widehat{\boldsymbol{Y}}}{\partial \boldsymbol{Y}^\top} = \boldsymbol{H},$$

y en particular $\partial \widehat{Y}_i / \partial Y_i = h_{ii}$, para $i = 1, \ldots, n$.

PROPIEDAD 3.20. Si el modelo *tiene intercepto*, entonces $\boldsymbol{H}\mathbf{1} = \mathbf{1}$.

DEMOSTRACIÓN. Considere $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1)$. Sabemos que $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$, y de ahí que

$$\boldsymbol{H}(\mathbf{1}, \boldsymbol{X}_1) = (\mathbf{1}, \boldsymbol{X}_1),$$

y el resultado sigue. □

3.5. Diagnóstico de por eliminación de casos

Suponga que se desea evaluar el efecto de eliminar una observación sobre la estimación de $\boldsymbol{\beta}$. En este caso, podemos considerar el *modelo de datos eliminados*

$$\boldsymbol{Y}_{(i)} = \boldsymbol{X}_{(i)}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{(i)}, \tag{3.7}$$

de ahí que

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1} \boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)}.$$

Un aspecto interesante de la estimación de parámetros en el modelo definido en la Ecuación (3.7) es que no requiere re-ajustar n modelos (uno por cada observación que hemos removido), sino que podemos escribir el estimador $\widehat{\boldsymbol{\beta}}_{(i)}$ en términos de información calculada para el modelo con *datos completos*. Para notar esto, considere:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(i)} \\ \boldsymbol{x}_i^\top \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{(i)} \\ Y_i \end{pmatrix},$$

de ahí que

$$\boldsymbol{X}^\top \boldsymbol{X} = (\boldsymbol{X}_{(i)}^\top, \boldsymbol{x}_i) \begin{pmatrix} \boldsymbol{X}_{(i)} \\ \boldsymbol{x}_i^\top \end{pmatrix} = \boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)} + \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

$$\boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{X}_{(i)}^\top, \boldsymbol{x}_i) \begin{pmatrix} \boldsymbol{Y}_{(i)} \\ Y_i \end{pmatrix} = \boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)} + \boldsymbol{x}_i Y_i.$$

Reagrupando, podemos escribir

$$\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)} = \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{I} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^\top),$$

cuya matriz inversa es dada por

$$(\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1} = (\boldsymbol{I} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$
$$= \Big\{\boldsymbol{I} + \frac{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top}{1-\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i}\Big\}(\boldsymbol{X}^\top\boldsymbol{X})^{-1},$$

sigue que el estimador de $\boldsymbol{\beta}$ en el modelo con datos eliminados, puede ser escrito como:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(i)} &= (\boldsymbol{X}_{(i)}^\top\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^\top\boldsymbol{Y}_{(i)}\\ &= \Big\{\boldsymbol{I} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big\}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y} - \boldsymbol{x}_iY_i)\\ &= \Big\{\boldsymbol{I} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big\}\big(\widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_iY_i\big)\\ &= \widehat{\boldsymbol{\beta}} - \frac{e_i}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i.\end{aligned}$$

Este resultado permite definir, por ejemplo,

$$e_{j(i)} = Y_j - \widehat{Y}_{j(i)} = Y_j - \boldsymbol{x}_j^\top\widehat{\boldsymbol{\beta}}_{(i)},$$

como el j-ésimo residuo con la i-ésima observación eliminada. Adicionalmente,

$$\begin{aligned}e_{i(i)} &= Y_i - \widehat{Y}_{i(i)} = Y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{(i)} = Y_i - \boldsymbol{x}_i^\top\Big(\widehat{\boldsymbol{\beta}} - \frac{e_i}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\Big)\\ &= e_i + \frac{e_ih_{ii}}{1-h_{ii}} = \frac{e_i}{1-h_{ii}},\end{aligned}$$

es conocido como *residuo eliminado*. Lo que por su vez, permite evaluar el efecto de la i-ésima observación sobre el estimador de σ^2. Considerare el estimador $s^2_{(i)}$, definido mediante:

$$\mathsf{RSS}_{(i)} = \sum_{j\neq i}(Y_j - \boldsymbol{x}_j^\top\widehat{\boldsymbol{\beta}}_{(i)})^2 = \mathsf{RSS} - \frac{e_i^2}{1-h_{ii}}.$$

De ahí que, el estimador insesgado de σ^2 cuando removemos la i-ésima observación es dado por:

$$s^2_{(i)} = \frac{1}{n-p-1}\Big\{(n-p)s^2 - \frac{e_i^2}{1-h_{ii}}\Big\}$$

RESULTADO 3.21. *Considere el modelo de salto en la media:*

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{d}_i\gamma + \boldsymbol{\epsilon}, \tag{3.8}$$

con $\boldsymbol{\epsilon} \sim \mathsf{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ *y* $\boldsymbol{d}_i = (\boldsymbol{0}, 1, \boldsymbol{0})^\top$ *un vector de ceros con un 1 en la* i*-ésima posición. De este modo, el estimador ML de* $\boldsymbol{\beta}$ *en el modelo* (3.7) *y* (3.8) *coinciden.*

DEMOSTRACIÓN. El resultado sigue mediante escribir el modelo en (3.8) como

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{d}_i), \qquad \boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \gamma)^\top,$$

y $\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}$. Tenemos que

$$\boldsymbol{Z}^\top\boldsymbol{Z} = \begin{pmatrix}\boldsymbol{X}^\top\\ \boldsymbol{d}_i^\top\end{pmatrix}(\boldsymbol{X}, \boldsymbol{d}_i) = \begin{pmatrix}\boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{d}_i\\ \boldsymbol{d}_i^\top\boldsymbol{X} & \boldsymbol{d}_i^\top\boldsymbol{d}_i\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{x}_i\\ \boldsymbol{x}_i^\top & 1\end{pmatrix},$$
$$\boldsymbol{Z}^\top\boldsymbol{Y} = \begin{pmatrix}\boldsymbol{X}^\top\\ \boldsymbol{d}_i^\top\end{pmatrix}\boldsymbol{Y} = \begin{pmatrix}\boldsymbol{X}^\top\boldsymbol{Y}\\ Y_i\end{pmatrix}.$$

Sabemos que

$$(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} = \begin{pmatrix} (\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1} & -\frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i \\ -\frac{1}{1-h_{ii}}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1} & \frac{1}{1-h_{ii}} \end{pmatrix},$$

luego

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_* &= \Big\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\Big\}\boldsymbol{X}^\top\boldsymbol{Y} - \frac{Y_i}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i \\ &= \Big\{\boldsymbol{I} + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big\}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} - \frac{Y_i}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i \\ &= \widehat{\boldsymbol{\beta}} + \frac{\widehat{Y}_i}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i - \frac{Y_i}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i = \widehat{\boldsymbol{\beta}}_{(i)},\end{aligned}$$

lo que termina la prueba. □

Basado en el elipsoide de confianza del $100(1-\alpha)\,\%$ para $\boldsymbol{\beta}$,

$$\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{ps^2} \leq F_{p,n-p}(1-\alpha).$$

Cook (1977) propuso determinar la influencia de la i-ésima observación, usando

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{(i)}-\widehat{\boldsymbol{\beta}})}{ps^2} = \frac{r_i^2}{p}\Big(\frac{h_i}{1-h_i}\Big),$$

para $i = 1,\ldots,n$, y recomendó comparar D_i con algún percentil de la distribución $F_{p,n-p}$ (α) con $\alpha = 0.10$. Aunque otra alternativa que puede ser más razonable es usar $\alpha = 0.50$. Adicionalmente se ha sugerido que $D_i > 1$ es un *indicador de observaciones influyentes*.

Se han introducido diversas medidas de diagnóstico, por ejemplo, Welsch y Kuh (1977) propusieron medir el impacto en la i-ésima observación sobre el valor predicho como

$$\mathsf{DFFIT}_i = \widehat{Y}_i - \widehat{Y}_{i(i)} = \boldsymbol{x}_i^\top(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}) = \frac{h_{ii}e_i}{1-h_{ii}},$$

o bien, utilizando su versión estandarizada

$$\begin{aligned}\mathsf{DFFITS}_i &= \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}} = \Big(\frac{h_{ii}}{1-h_{ii}}\Big)^{1/2}\frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \\ &= \Big(\frac{h_{ii}}{1-h_{ii}}\Big)^{1/2}t_i.\end{aligned}$$

Sobre este tipo de medidas, Belsley et al. (1980) sugieren poner especial atención en aquellos casos donde $\mathsf{DFFITS}_i > 2\sqrt{p/n}$.

Observación 3.22. En ocasiones esta medida es conocida como *distancia Welsch-Kuh*.

Por otro lado, Atkinson (1981) sugirió usar una versión modificada de la distancia de Cook, como

$$\begin{aligned} AK_i &= \sqrt{\Big(\frac{n-p}{p}\Big)\Big(\frac{h_{ii}}{1-h_{ii}}\Big)}\Big|\frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}\Big| \\ &= \sqrt{\Big(\frac{n-p}{p}\Big)\Big(\frac{h_{ii}}{1-h_{ii}}\Big)}|t_i| \\ &= \sqrt{\frac{n-p}{p}}|\,\mathsf{DFFITS}_i\,|. \end{aligned}$$

Cuando $h_{ii} = p/n, \forall i$, tenemos $AK_i = |t_i|$ debido a esto se recomienda hacer el gráfico de AK_i vs. $|t_i|$. Además podemos identificar la i-ésima observación como influyente si $AK_i > 2$.

Observación 3.23. AK_i puede considerarse como una medida de influencia conjunta sobre $\widehat{\boldsymbol{\beta}}$ y s^2 simultáneamente.

Como un intento de fundamentar las técnicas para llevar a cabo diagnóstico por eliminación de casos, Cook y Weisberg (1980) propusieron considerar medidas generales de influencia, basadas en la función de influencia empírica, lo que los llevó a definir:

$$D_i(\boldsymbol{M}, c) = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^\top \boldsymbol{M} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{c}, \qquad i = 1, \ldots, n,$$

donde $\boldsymbol{M}$ es matriz definida positiva $p \times p$ y $c > 0$ es un factor de escala. En efecto, podemos escribir diversas medidas bajo esta perspectiva, considere la siguiente tabla:

$\boldsymbol{M}$	c	Medida	Referencia
$\boldsymbol{X}^\top\boldsymbol{X}$	ps^2	D_i	Cook (1977)
$\boldsymbol{X}^\top\boldsymbol{X}$	$ps^2_{(i)}$	$(\mathsf{DFFITS}_i)^2$	Welsch y Kuh (1977)
$\boldsymbol{X}^\top\boldsymbol{X}$	$(n-1)^2 ps^2_{(i)}/(n-p)$	AK_i	Atkinson (1981)

Es posible, centrar nuestra atención en diversos aspectos de la modelación, no solamente en el efecto sobre los estimadores. Por ejemplo, resulta interesante comparar $\mathsf{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ con la matriz de covarianza que resulta de eliminar el i-ésimo caso. Esto llevó a Belsley et al. (1980) a definir,

$$\begin{aligned} COVRATIO_i &= \frac{\det\{s^2_{(i)}(\boldsymbol{X}^\top_{(i)}\boldsymbol{X}_{(i)})^{-1}\}}{\det\{s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\}} = \Big(\frac{s^2_{(i)}}{s^2}\Big)^p \frac{\det(\boldsymbol{X}^\top_{(i)}\boldsymbol{X}_{(i)})^{-1}}{\det(\boldsymbol{X}^\top\boldsymbol{X})^{-1}} \\ &= \frac{1}{1-h_i}\Big(\frac{n-p-r_i^2}{n-p-1}\Big)^p. \end{aligned}$$

Además, se ha planteado el siguiente punto de corte $|COVRATIO_i - 1| > 3p/n$.

Debemos destacar que Belsley et al. (1980) y Velleman y Welsch (1981) han discutido *estrategias para amenizar el cálculo* de estas medidas de influencia. Para modelos de regresión lineal algunas de estas medidas han sido implementadas en software estadístico tales como SAS, SPSS o S-Plus/R.

En particular, R (o S-Plus) disponen de las funciones `lm.influence` y `ls.diag` asociadas con las funciones `lm` (o `glm`) y `lsfit`, respectivamente. Estas cantidades pueden ser obtenidas de *forma eficiente* usando la descomposición QR o SVD.

3.6. Procedimientos para estimación robusta

En esta sección presentamos dos métodos para obtener estimadores que permiten atenuar el efecto de outliers o observaciones atípicas. El primero es conocido como M-estimadores, mientras que el segundo se basa en considerar distribuciones con colas más pesadas que la normal.

3.6.1. *M*-estimadores. Para introducir ideas considere $X_1, \ldots, X_n$ variables aleatorias IID. Sabemos que la media muestral $\overline{X}$ es solución del problema,

$$\min_{\theta} \sum_{i=1}^{n} (x_i - \theta)^2,$$

o análogamente,

$$\sum_{i=1}^{n} (x_i - \theta) = 0.$$

Mientras que la mediana, denotada como $\mathsf{me}(\boldsymbol{x})$, que es *robusta contra outliers*, es solución del problema

$$\min_{\theta} \sum_{i=1}^{n} |x_i - \theta|,$$

es decir $\mathsf{me}(\boldsymbol{x})$ debe satisfacer la ecuación

$$\sum_{i=1}^{n} \big\{(-1) I_{(-\infty,0)}(x_i - \theta) + I_{(0,\infty)}(x_i - \theta)\big\} = 0.$$

Sabemos que para una muestra aleatoria $X_1, \ldots, X_n$ desde el modelo estadístico $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ con función de densidad $f(x;\theta)$, el MLE $\widehat{\theta}_{\mathsf{ML}}$ corresponde al minimizador del negativo de la log-verosimilitud

$$\min_{\theta} \Big\{ -\sum_{i=1}^{n} \log f(x_i;\theta) \Big\},$$

y en el caso de que $\ell(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$ sea diferenciable, $\widehat{\theta}_{\mathsf{ML}}$ debe ser solución de:

$$\sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial \theta} = 0,$$

lo que lleva a la siguiente definición.

Definición 3.24 (M-estimador). Para $X_1, \ldots, X_n$ variables aleatorias IID. El *M-estimador* $\widehat{\theta}_{\mathsf{M}}$ con respecto a la función $\psi : \mathbb{R} \times \Theta \to \mathbb{R}$ se define como la solución de la ecuación

$$\sum_{i=1}^{n} \psi(X_i; \widehat{\theta}_{\mathsf{M}}) = 0. \tag{3.9}$$

Usualmente el estimador definido por la Ecuación (3.9) corresponde a la solución del problema de optimización

$$\min_{\theta} \sum_{i=1}^{n} \rho(x_i; \theta),$$

si ρ es diferenciable, entonces

$$\psi(x; \theta) = c \, \frac{\partial \rho(x; \theta)}{\partial \theta},$$

para alguna constante c.

OBSERVACIÓN 3.25. Para simplificar la notación podemos hacer

$$\rho(x; \theta) = \widetilde{\rho}(x - \theta), \qquad \psi(x; \theta) = \widetilde{\psi}(x - \theta).$$

EJEMPLO 3.26 (Estimador LS). Sea $z = x - \theta$. Las funciones $\widetilde{\rho}(z) = z^2$ y $\widetilde{\psi}(z) = z$ llevan a la media muestral.

EJEMPLO 3.27 (Estimador LAD[3]). La mediana es un M-estimador con $\widetilde{\rho}(z) = |z|$, y

$$\widetilde{\psi}(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0. \end{cases}$$

EJEMPLO 3.28 (Media recortada). La primera propuesta de Huber (1981) para reducir la influencia de outliers es:

$$\widetilde{\rho}(z) = \begin{cases} z^2, & |z| \leq k, \\ k^2, & |z| > k, \end{cases}$$

donde k es una *constante de tuning*, y

$$\widetilde{\psi}(z) = \begin{cases} z, & |z| \leq k, \\ 0, & |z| > k. \end{cases}$$

EJEMPLO 3.29 (Media Winsorizada). Huber (1981) propuso un compromiso entre la media y la mediana, como:

$$\widetilde{\rho}(z) = \begin{cases} \frac{1}{2} z^2, & |z| \leq k, \\ k|z| - \frac{1}{2} k^2, & |z| > k. \end{cases}$$

El estimador $\widehat{\theta}_{\mathsf{M}}$ es solución de:

$$\sum_{i=1}^{n} \widetilde{\psi}(x_i - \theta) = 0,$$

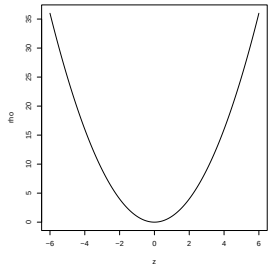
donde

$$\widetilde{\psi}(z) = \begin{cases} -k, & z < -k, \\ z, & |z| \leq k, \\ k, & z > k. \end{cases}$$

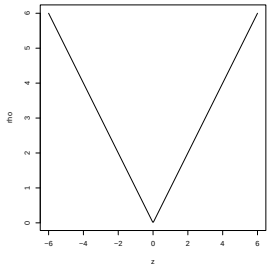
Debemos destacar que, para $k \to 0$ obtenemos la mediana, cuando $k \to \infty$ lleva a la media, mientras que $k = 1.345$ tiene 95 % de eficiencia bajo normalidad.

[3]Estimadores *mínimo desvío absoluto (LAD)* corresponden a estimadores L_1.

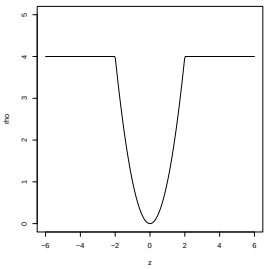
En los gráficos a continuación se presenta las funciones $\rho(\cdot)$ para cada uno de los ejemplos anteriores,



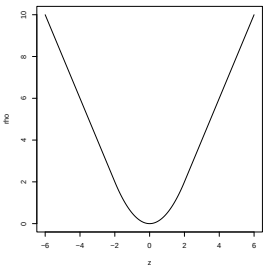
(a) media



(b) mediana



(c) media recortada



(d) media winsorizada

Sabemos que el estimador LS en regresión es solución del problema de estimación

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2, \tag{3.10}$$

y, bajo normalidad, el MLE de $\boldsymbol{\beta}$ minimiza la función

$$-\sum_{i=1}^{n} \log f(Y_i; \boldsymbol{\beta}) = \frac{n}{2} \log 2\pi\sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2. \tag{3.11}$$

Para obtener estimadores robustos Huber (1981) sugirió substituir el negativo de la función de log-verosimilitud en (3.11) por una función que permita disminuir el efecto de outliers. De este modo Huber (1981) propuso obtener estimadores tipo-ML, conocidos como M-estimadores resolviendo el problema

$$\min_{\beta} \sum_{i=1}^{n} \rho(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}). \tag{3.12}$$

Es usual incorporar el parámetro de escala σ^2 definiendo,

$$z_i = \frac{Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}, \qquad i = 1, \ldots, n,$$

lo que lleva a considerar la función objetivo:

$$Q_\rho(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho(z_i).$$

Es fácil notar que,

$$\frac{\partial Q_\rho(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \rho(z_i)}{\partial z_i}\frac{\partial z_i}{\partial \beta_j} = \sum_{i=1}^{n} W(z_i) z_i \frac{\partial z_i}{\partial \beta_j}, \qquad W(z_i) = \frac{1}{z_i}\frac{\partial \rho(z_i)}{\partial z_i}.$$

Tenemos que $\psi(z) = \partial\rho(z)/\partial z$, así $W(z) = \psi(z)/z$. Por tanto el M-estimador de $\boldsymbol{\beta}$ es solución del sistema de ecuaciones

$$\sum_{i=1}^{n} \psi\Big(\frac{Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\Big)\frac{x_{ij}}{\sigma} = 0, \qquad j = 1, \ldots, p,$$

que puede ser escrito en forma compacta como

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} W_i(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})\boldsymbol{x}_i = \mathbf{0} \tag{3.13}$$

Sea $\boldsymbol{W} = \text{diag}(W_1, \ldots, W_n)$ de este modo podemos escribir (3.13) como:

$$\boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}.$$

Sin embargo, debemos resaltar que los pesos W_i dependen de z_i, los que a su vez dependen de $\boldsymbol{\beta}$ y por tanto se requiere métodos iterativos para obtener el M-estimador, $\widehat{\boldsymbol{\beta}}_{\mathsf{M}}$.

Usando una estimación inicial $\boldsymbol{\beta}^{(0)}$ podemos considerar el siguiente esquema iterativo

$$\boldsymbol{\beta}^{(r+1)} = (\boldsymbol{X}^\top \boldsymbol{W}^{(r)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(r)} \boldsymbol{Y}, \tag{3.14}$$

con

$$\boldsymbol{W}^{(r)} = \text{diag}(W_1^{(r)}, \ldots, W_n^{(r)}), \qquad W_i^{(r)} = \psi(e_i^{(r)}/\sigma)/(e_i^{(r)}/\sigma),$$

y

$$\boldsymbol{e}^{(r)} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(r)}.$$

Desde el punto de vista computacional es preferible usar

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \boldsymbol{p}_r,$$

con

$$\boldsymbol{p}_r = (\boldsymbol{X}^\top \boldsymbol{W}^{(r)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(r)} \boldsymbol{e}^{(r)}.$$

Observación 3.30. El procedimiento delineado en (3.14) es conocido como *mínimos cuadrados iterativamente ponderados (IRLS)*. Debemos destacar que la convergencia del proceso iterativo en (3.14) sólo es garantizada para funciones ρ *convexas* así como para funciones de *redescenso*. Además, una serie de autores han propuesto métodos refinados para obtener M-estimadores basados en IRLS (ver, por ejemplo O'Leary, 1990).

Existe una gran variedad de funciones $\rho(\cdot)$ o $\psi(\cdot)$ para definir M-estimadores, por ejemplo:

- *Tukey's biweight*

$$\psi(z) = z\big[1 - (t/k)\big]_+^2, \qquad k = 4.685.$$

- *Hampel's* ψ

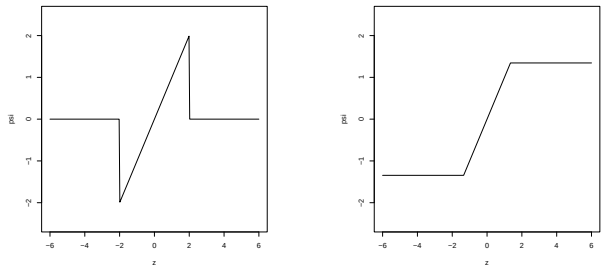
$$\psi(z) = \text{sign}(z) \begin{cases} |z|, & 0 < |z| \leq a, \\ a, & a < |z| \leq b, \\ a(c - |z|)/(c - b), & b < |z| \leq c, \\ 0, & c < |z|, \end{cases}$$

  con $a = 1.645$, $b = 3$ y $c = 6.5$.
- *Función seno de Andrews*

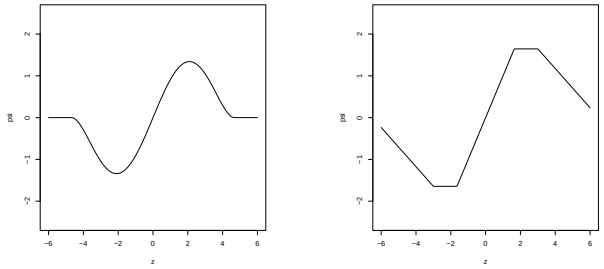
$$\psi(z) = \begin{cases} \sin(z/a), & |z| \leq \pi a, \\ 0, & |z| > \pi a, \end{cases}$$

  donde $a = 1.339$.

Resulta instructivo visualizar el comportamiento de algunas funciones $\psi(\cdot)$ seleccionadas en la figura a continuación,



(a) media recortada (b) Huber



(c) Tukey bisquare (d) Hampel

Aunque es de interés primario la estimación de los coeficientes de regresión, se ha sugerido usar el siguiente estimador robusto para σ,

$$\widehat{\sigma}_{\mathsf{rob}} = \frac{\mathsf{MAD}(\boldsymbol{e})}{0.6745},$$

con

$$\mathsf{MAD}(\boldsymbol{e}) = \mathsf{me}(|\boldsymbol{e} - \mathsf{me}(\boldsymbol{e})|),$$

que es conocido como *desviación mediana absoluta* de los residuos minimos cuadrados $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathsf{LS}}$.

Análogamente a $\boldsymbol{\beta}$ es posible obtener un M-estimador para σ resolviendo

$$\frac{1}{n-p}\sum_{i=1}^{n}\chi\Big(\frac{Y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma}\Big)=\delta,$$

para δ una constante positiva.

La fundamentación de las propiedades asintóticas de la clase de M-estimadores de $\boldsymbol{\beta}$ escapan al ámbito en el que se desarrolla la asigntura. Sin embargo, podemos proveer un estimador de la matriz de covarianza de $\widehat{\boldsymbol{\beta}}_{\mathsf{M}}$ considerando,

$$\kappa=1+\frac{p}{n}\frac{\mathsf{var}(\psi')}{\{\mathsf{E}(\psi')\}^2},$$

que es evaluado en la distribución de los errores $\boldsymbol{\epsilon}$ (en la práctica deben ser estimados desde los residuos). Entonces la matriz de covarianza asintótica de $\widehat{\boldsymbol{\beta}}_{\mathsf{M}}$ es dada por (ver Huber, 1981)

$$\kappa^2\frac{\sum_i\psi^2(e_i)/(n-p)}{[\sum_i\psi'(e_i)/n]^2}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.$$

Una subclase interesante de M-estimadores, corresponde a los problemas de estimación norma-L_p, los que son definidos como la solución del problema:

$$\min_{\beta}\Big(\sum_{i=1}^{n}|Y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}|^p\Big)^{1/p},\qquad p\geq 1. \tag{3.15}$$

En efecto, basta considerar

$$\rho(z)=|z|^p,\qquad \psi(z)=|z|^{p-2},\qquad W(z)=|z|^{p-2}.$$

Observación 3.31. Considere $Y_1,\ldots,Y_n$ variables aleatorias provenientes de la función de densidad

$$f_p(y)=c\exp(-|y|^p),\qquad p\geq 1,$$

con c una constante de normalización. Esto permite caracterizar la estimación de norma-L_p como estimación ML basado en la densidad $f_p(y)$. Además, se debe destacar que la *distribución Laplace* es obtenida para $p=1$, mientras que la *distribución normal* es recuperada para $p=2$.

Para $1<p<2$ podemos usar IRLS como un método para aproximar la solución del problema en (3.15). En efecto, Osborne (1985) reestableció el problema de estimación norma-L_p como:

$$\min_{\beta}Q_p(\boldsymbol{\beta}),\qquad Q_p(\boldsymbol{\beta})=\sum_{i=1}^{n}|\epsilon_i|^p=\sum_{i=1}^{n}|\epsilon_i|^{p-2}\epsilon_i^2,$$

que puede ser interpretado como un problema de mínimos cuadrados ponderados,

$$\min_{\beta}\|\boldsymbol{W}^{(p-2)/2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\|_2^2,\qquad \boldsymbol{W}=\mathrm{diag}(|\epsilon_1|,\ldots,|\epsilon_n|).$$

Esto lleva al siguiente algoritmo.

Algoritmo 5: IRLS para estimación L_p.

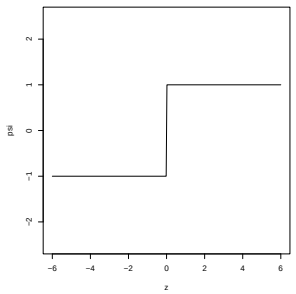
Entrada: Datos $\boldsymbol{X}$, $\boldsymbol{y}$, $p \in [1, 2)$ y estimación inicial $\boldsymbol{\beta}^{(0)}$.
Salida : Aproximación del estimador L_p, $\widehat{\boldsymbol{\beta}}$.

begin
  for $r = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{e}^{(r)} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(r)}$
    $\boldsymbol{W}^{(r)} = \text{diag}(|e_1^{(r)}|^{(p-2)/2}, \ldots, |e_n^{(r)}|^{(p-2)/2})$
    Resolver $\boldsymbol{p}_r$ desde
      $\min_{p_r} \|\boldsymbol{W}^{(r)}(\boldsymbol{e}^{(r)} - \boldsymbol{X}\boldsymbol{p}_r)\|_2^2$
    Hacer
    $\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \boldsymbol{p}_r$
  end
  return $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(\star)}$
end

Debemos destacar que, para $p = 1$ tenemos

$$\psi(z) = z/|z|.$$

con



De este modo, el procedimiento de estimación norma-L_1 es robusto, con pesos

$$W_i = 1/|e_i|, \qquad i = 1, \ldots, n.$$

Schlossmacher (1973) fue uno de los primeros autores en proponer el uso de IRLS para estimación norma-L_1 en regresión basado en la ecuación

$$\sum_{i=1}^{n} \frac{Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{|Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}|} \boldsymbol{x}_i = \boldsymbol{0}.$$

Sin embargo este procedimiento ha sido fuertemente criticado por una gran cantidad de autores (ver discusión en Capítulo 4 de Björck, 1996), por no ser capaz de identificar las *observaciones básicas* que definen el estimador.

Una perspectiva diferente fue adoptada por Charnes et al. (1955) quienes mostraron que problema de regresión L_1,

$$\min_{\beta} Q_1(\boldsymbol{\beta}), \qquad Q_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}|,$$

es equivalente a resolver el siguiente problema de *programación lineal (LP)*

$$\begin{aligned} &\text{mín} \sum_{i=1}^{n} (\epsilon_i^+ + \epsilon_i^-), \\ \text{sujeto a:}\ \ &\epsilon_i^+ \geq 0,\ \epsilon_i^- \geq 0, \\ &\epsilon_i^+ - \epsilon_i^- = \epsilon_i = Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}, \end{aligned}$$

para $i = 1, \ldots, n$, con ϵ_i^+ y ϵ_i^- variables no negativas.

OBSERVACIÓN 3.32. Barrodale y Roberts (1973, 1974) presentaron un algoritmo de propósito especial para resolver el problema de regresión L_1, modificando el método simplex y la estructura de datos requerida.

3.6.2. Estimación usando distribuciones con colas pesadas. Frecuentemente se ha sugerido substituir la distribución normal por distribuciones con colas más pesadas como un mecanismo para la *acomodación de outliers*. En efecto, en la Observación 3.31 se indicó que usar la distribución Laplace, permite obtener estimadores L_1 en regresión, lo que corresponde a un método robusto contra outliers. En esta sección sin embargo, confiaremos en la simpleza y elegancia del procedimiento de estimación por máxima verosimilitud. En esta sección nos enfocaremos en abordar la estimación de los parámetros de regresión usando la clase de distribuciones de contornos elípticos, discutida en la Sección 1.6.

Debemos resaltar que *diferentemente* al caso de la distribución normal, en el caso general de la familia elíptica, podemos tener los siguientes enfoques:

(a) *Modelo dependiente:* Supondremos $Y_1, \ldots, Y_n$ tal que su *densidad conjunta* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \sim \mathsf{EC}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n; g)$, sigue una distribución de contornos elipticos.[4]

(b) *Modelo independiente:* Considere $Y_1, \ldots, Y_n$ variables aleatorias *independientes* cada una con distribución $\mathsf{EC}_1(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2; g)$.

Debido a esta consideración a continuación nos enfocaremos en el *modelo independiente*,

$$Y_i \overset{\text{ind}}{\sim} \mathsf{EC}_1(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2; g), \qquad i = 1, \ldots, n,$$

donde

$$f(y_i; \boldsymbol{\theta}) = \frac{1}{\sigma} g\big((Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 / \sigma^2\big),$$

con $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$, y la función generadora de densidades g debe satisfacer:

$$\int_0^\infty u^{-1/2} g(u)\, \mathsf{d}\, u < +\infty.$$

OBSERVACIÓN 3.33. Como se mencionó en la Sección 1.6, la función generadora de densidades $g(\cdot)$ puede depender de *parámetros de forma*, los que controlan el grado de curtosis de la distribución (es decir, que tan *pesada* es la cola de la distribución). A continuación asumiremos que tales parámetros son conocidos, o equivalentemente que la función $g(\cdot)$ es conocida.

[4] El estimador de $\boldsymbol{\beta}$ en el modelo dependiente es equivalente al LSE y por tanto NO es robusto.

Obtenemos los estimadores máximo verosímiles de $\boldsymbol{\beta}$ y σ^2, asumiendo $g(\cdot)$ conocido, mediante maximizar la función de log-verosimilitud:

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log\sigma^2 + \sum_{i=1}^{n}\log g(u_i),$$

donde $u_i = (Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2/\sigma^2$, para $i = 1,\ldots,n$. Diferenciando $\ell(\boldsymbol{\theta})$ con relación a $\boldsymbol{\beta}$, obtenemos

$$\begin{aligned}
\mathsf{d}_\beta\,\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\mathsf{d}_\beta\log g(u_i) = \sum_{i=1}^{n}\frac{g'(u_i)}{g(u_i)}\,\mathsf{d}_\beta u_i \\
&= -\frac{2}{\sigma^2}\sum_{i=1}^{n}\frac{g'(u_i)}{g(u_i)}(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})\boldsymbol{x}_i^\top\,\mathsf{d}\,\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}W_i(\boldsymbol{\theta})(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})\boldsymbol{x}_i^\top\,\mathsf{d}\,\boldsymbol{\beta}
\end{aligned}$$

donde $W_i(\boldsymbol{\theta}) = -2g'(u_i)/g(u_i)$, para $i = 1,\ldots,n$. Análogamente, diferenciando $\ell(\boldsymbol{\theta})$ con relación a σ^2, tenemos

$$\begin{aligned}
\mathsf{d}_{\sigma^2}\,\ell(\boldsymbol{\theta}) &= -\frac{n}{2\sigma^2}\,\mathsf{d}\,\sigma^2 + \sum_{i=1}^{n}\frac{g'(u_i)}{g(u_i)}\,\mathsf{d}_{\sigma^2}\,u_i \\
&= -\frac{n}{2\sigma^2}\,\mathsf{d}\,\sigma^2 - \frac{1}{\sigma^4}\sum_{i=1}^{n}\frac{g'(u_i)}{g(u_i)}(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2\,\mathsf{d}\,\sigma^2 \\
&= -\frac{n}{2\sigma^2}\,\mathsf{d}\,\sigma^2 + \frac{1}{2\sigma^4}\sum_{i=1}^{n}W_i(\boldsymbol{\theta})(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2\,\mathsf{d}\,\sigma^2 \\
&= -\frac{n}{2\sigma^2}\,\mathsf{d}\,\sigma^2 + \frac{1}{2\sigma^4}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\,\mathsf{d}\,\sigma^2
\end{aligned}$$

donde

$$\boldsymbol{W} = \operatorname{diag}\big(W_1(\boldsymbol{\theta},\ldots,W_n(\boldsymbol{\theta}))\big).$$

La condición de primer orden, lleva al siguiente sistema de ecuaciones:

$$\begin{aligned}
\boldsymbol{X}^\top\widehat{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) &= \boldsymbol{0}, \\
\widehat{\sigma}^2 &= \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}),
\end{aligned}$$

que *no tiene solución en forma explícita* y por tanto se requiere el uso de métodos iterativos.

Por ejemplo, usando una estimación inicial $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, actualizamos las estimaciones para $\boldsymbol{\beta}$ y σ^2, como:

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= (\boldsymbol{X}^\top\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}^{(k)}\boldsymbol{Y}, \\
\sigma^{2(k+1)} &= \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)})\boldsymbol{W}^{(k)}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}),
\end{aligned}$$

a la convergencia del algoritmo, hacemos $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$.

Ejemplo 3.34. Las funciones de pesos para algunas distribuciones elípticas seleccionadas son dadas a continuación:

- *Normal:* $Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} \mathsf{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, tenemos:

$$W_i(\boldsymbol{\theta}) = 1, \qquad i = 1, \ldots, n.$$

- *t-Student:* $Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} t(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2; \nu)$, $\nu > 0$,

$$W_i(\boldsymbol{\theta}) = \frac{\nu + 1}{\nu + u_i}, \qquad i = 1, \ldots, n,$$

  la distribución $\mathsf{Cauchy}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, es obtenida para $\nu = 1$.

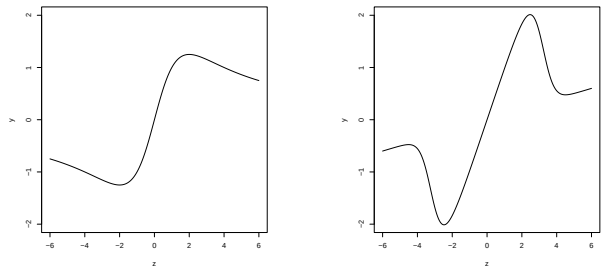
- *Normal contaminada:* $Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} \mathsf{CN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2; \epsilon, \gamma)$, con $\epsilon \in [0, 1)$, $\gamma > 0$,

$$W_i(\boldsymbol{\theta}) = \frac{(1-\epsilon)\exp(-u/2) + \epsilon\gamma^{-3/2}\exp(-u/(2\gamma))}{(1-\epsilon)\exp(-u/2) + \epsilon\gamma^{-1/2}\exp(-u/(2\gamma))}, \qquad i = 1, \ldots, n.$$

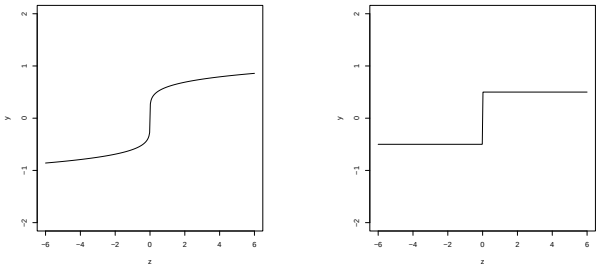
- *Exponencial Potencia:* $Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} \mathsf{PE}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2; \lambda)$, $\lambda > 0$,

$$W_i(\boldsymbol{\theta}) = \lambda u_i^{\lambda - 1}, \qquad i = 1, \ldots, n.$$

Es interesante notar el comportamiento de las funciones de influencia para cada una de estas distribuciones, considere el siguiente gráfico:



(a) t de Student, $\nu = 4$ (b) CN, $\epsilon = 0.05, \gamma = 10$



(c) PE, $\lambda = 0.6$ (d) PE, $\lambda = 0.5$

La clase de distribuciones elípticas contiene distribuciones con colas más pesadas y también más livianas que la normal. A continuación nos enfocaremos en una subclase conocida como *mezclas de escala normal* (Andrews y Mallows, 1974), clase que tiene una interesante interpretación que será explotada para la estimación de parámetros.

Sea $Y_1, \ldots, Y_n$ variable aleatorias independientes con distribución $\mathsf{SMN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2; \mathsf{H})$ cada una con densidad

$$f(y_i; \boldsymbol{\theta}) = (2\pi\sigma)^{-1/2} \int_0^\infty \omega^{1/2} \exp\big(-\omega u_i/2\big)\, \mathsf{d}\,\mathsf{H}(\omega),$$

donde $u_i = (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2/\sigma^2$, $i = 1, \ldots, n$.

Observación 3.35. Sabemos que, una variable aleatoria $Y_i \sim \mathsf{SMN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2; \mathsf{H})$, admite la representación:

$$Y_i | W = \omega \sim \mathsf{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2/\omega), \qquad W \sim \mathsf{H}(\boldsymbol{\delta}),$$

lo que permite abordar la estimación ML usando el *algoritmo EM* (Dempster, Laird y Rubin, 1977).

El algoritmo EM permite el cálculo iterativo de *estimadores ML* en modelos con *datos incompletos*. De este modo, requiere de una *formulación de datos aumentados*. Su principal ventaja es que reemplaza una optimización *"compleja"* (asociada a la estimación ML) por una serie de maximizaciones *"simples"*. A continuación damos una breve descripción de este procedimiento, para una revisión de las propiedades de este algoritmo y una serie de consideraciones prácticas consulte McLachlan y Krishnan (2008).

3.6.2.1. Algoritmo EM (Esperanza-Maximización). Sea $\boldsymbol{Y}_{\mathsf{obs}}$ vector de *datos observados* con función de densidad $f(\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta})$. El objetivo es aumentar los datos observados $\boldsymbol{Y}_{\mathsf{obs}}$ con variables latentes $\boldsymbol{Y}_{\mathsf{mis}}$ también conocidos como *datos perdidos*. Esto es, consideramos el vector de *datos completos*

$$\boldsymbol{Y}_{\mathsf{com}} = (\boldsymbol{Y}_{\mathsf{obs}}^\top, \boldsymbol{Y}_{\mathsf{mis}}^\top)^\top,$$

tal que la densidad $f(\boldsymbol{y}_{\mathsf{com}}; \boldsymbol{\theta})$ sea simple. El algoritmo EM es útil cuando la función de log-verosimilitud

$$\begin{aligned} \ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}}) &= \log f(\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta}) \\ &= \log \int f(\boldsymbol{y}_{\mathsf{com}}; \boldsymbol{\theta})\, \mathsf{d}\boldsymbol{y}_{\mathsf{mis}}, \end{aligned}$$

es difícil de maximizar directamente. El objetivo del algoritmo EM es realizar la estimación ML iterativamente basandose en la *log-verosimilitud de datos completos*:

$$\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}}) = \log f(\boldsymbol{y}_{\mathsf{com}}; \boldsymbol{\theta}).$$

El algoritmo EM permite obtener los MLE en *problemas con datos incompletos* por medio de las etapas:

Paso E: para $\boldsymbol{\theta}^{(k)}$ estimación de $\boldsymbol{\theta}$ en la k-ésima iteración, calcular la Q-función,

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) &= \mathsf{E}\{\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}}) | \boldsymbol{Y}_{\mathsf{obs}}, \boldsymbol{\theta}^{(k)}\} \\ &= \int \ell_c(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}}) f(\boldsymbol{y}_{\mathsf{mis}} | \boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta}^{(k)})\, \mathsf{d}\boldsymbol{y}_{\mathsf{mis}}. \end{aligned}$$

Paso M: determinar $\boldsymbol{\theta}^{(k+1)}$ como

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}).$$

Adicionalmente, Dempster, Laird y Rubin (1977) definieron un *algoritmo EM generalizado (GEM)*, mediante la siguiente modificación del Paso M:

Paso M*: seleccionar $\boldsymbol{\theta}^{(k+1)}$ satisfaciendo,

$$Q(\boldsymbol{\theta}^{(k+1)};\widehat{\boldsymbol{\theta}}^{(k)}) > Q(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}).$$

OBSERVACIÓN 3.36. Por ejemplo, podemos considerar *un único* paso Newton en la optimización de $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ para definir un GEM.

Algunas propiedades del algoritmo EM se presentan en los siguientes resultados

TEOREMA 3.37. *Todo algoritmo EM o GEM incrementa la log-verosimilitud de datos observados $\ell_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{Y}_{\mathsf{obs}})$ en cada iteración, esto es,*

$$\ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k+1)};\boldsymbol{Y}_{\mathsf{obs}}) \geq \ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k)};\boldsymbol{Y}_{\mathsf{obs}}).$$

DEMOSTRACIÓN. Ver Dempster, Laird y Rubin (1977). □

TEOREMA 3.38 (Convergencia del Algoritmo EM). *Bajo condiciones suaves, la secuencia $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ generada por el algoritmo EM (GEM). Converge a un punto estacionario de $\ell_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{Y}_{\mathsf{obs}})$.*

DEMOSTRACIÓN. Ver Wu (1983). □

Debemos destacar una serie de características relevantes del algoritmo EM, a saber:

- Frecuentemente el algoritmo EM es *simple*, de bajo costo computacional y numéricamente *estable*.
- Dempster, Laird y Rubin (1977) mostraron que el algoritmo EM converge con velocidad lineal, que depende de la *proporción* de información perdida.[5]
- Para modelos con datos aumentados con densidad en la familia exponencial, el algoritmo EM se reduce a *actualizar* las estadísticas suficientes.
- Errores estándar pueden ser obtenidos por cálculo directo, diferenciación numérica o usando el *Principio de Información Perdida* (Louis, 1982).

Sea $Y_1,\ldots,Y_n$ variables aleatorias independientes $\mathsf{SMN}(\boldsymbol{x}_i\boldsymbol{\beta},\sigma^2;\mathsf{H})$. Se llevará a cabo la estimación ML usando el algoritmo EM.

De este modo, tenemos el siguiente *modelo jerárquico*:

$$Y_i|W_i=\omega_i \sim \mathsf{N}(\boldsymbol{x}_i^\top\boldsymbol{\beta},\sigma^2/\omega_i), \qquad W_i \sim \mathsf{H}(\boldsymbol{\delta}), \qquad i=1,\ldots,n.$$

En este caso el vector de datos completos es $\boldsymbol{Y}_{\mathsf{com}} = (\boldsymbol{Y}^\top,\boldsymbol{\omega}^\top)^\top$, donde

$$\boldsymbol{Y} = (Y_1,\ldots,Y_n)^\top, \qquad \boldsymbol{\omega} = (\omega_1,\ldots,\omega_n)^\top.$$

En este contexto, $\boldsymbol{Y}$ corresponde a los datos observados, mientras que $\boldsymbol{\omega}$ serán asumidos como datos perdidos.

Considere una estimación para $\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}$, entonces

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = \mathsf{E}\{\ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y};\boldsymbol{\theta}^{(k)}\} = Q_1(\boldsymbol{\beta},\sigma^2;\boldsymbol{\theta}^{(k)}) + Q_2(\boldsymbol{\delta};\boldsymbol{\theta}^{(k)}),$$

donde

$$Q_1(\boldsymbol{\beta},\sigma^2;\boldsymbol{\theta}^{(k)}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\omega_i^{(k)}(Y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta})^2,$$

$$Q_2(\boldsymbol{\delta};\boldsymbol{\theta}^{(k)}) = \mathsf{E}\{\log h^{(n)}(\boldsymbol{\omega};\boldsymbol{\delta})|\boldsymbol{y};\boldsymbol{\theta}^{(k)}\},$$

[5] puede ser **extremadamente** lento.

con $\omega_i^{(k)} = \mathsf{E}(\omega_i|\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)})$ para $i = 1,\ldots,n$. En general, la forma para la esperanza condicional requerida en el Paso-E del algoritmo EM es dada por:

$$\mathsf{E}(\omega_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \frac{\int_0^\infty \omega_i^{3/2}\exp(-\omega_i u_i/2)\,\mathsf{dH}(\boldsymbol{\delta})}{\int_0^\infty \omega_i^{1/2}\exp(-\omega_i u_i/2)\,\mathsf{dH}(\boldsymbol{\delta})},$$

con $u_i = (Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2/\sigma^2$, $i = 1,\ldots,n$.

A continuación algunos ejemplos de funciones de pesos para miembros en la familia de mezclas de escala normal,

- *t-Student:* $Y_i \overset{\mathsf{ind}}{\sim} t(\boldsymbol{x}_i^\top\boldsymbol{\beta},\sigma^2,\nu)$, $i = 1,\ldots,n$, en cuyo caso
$$\mathsf{E}(\omega_i|Y_i;\boldsymbol{\theta}) = \frac{\nu+1}{\nu+u_i}.$$
- *Slash:* $Y_i \overset{\mathsf{ind}}{\sim} \mathsf{Slash}(\boldsymbol{x}_i^\top\boldsymbol{\beta},\sigma^2,\nu)$, para $i = 1,\ldots,n$. De este modo,
$$\mathsf{E}(\omega_i|Y_i;\boldsymbol{\theta}) = \Big(\frac{2\nu+1}{u_i}\Big)\frac{P_1(\nu+3/2,u_i/2)}{P_1(\nu+1/2,u_i/2)},$$
donde
$$P_z(a,b) = \frac{b^a}{\Gamma(a)}\int_0^z t^{a-1}e^{-bt}\,\mathsf{d}t,$$
es la función gama incompleta (regularizada).
- *Exponencial Potencia:* $Y_i \overset{\mathsf{ind}}{\sim} \mathsf{PE}(\boldsymbol{x}_i^\top\boldsymbol{\beta},\sigma^2,\lambda)$, $i = 1,\ldots,n$, donde
$$\mathsf{E}(\omega_i|Y_i;\boldsymbol{\theta}) = \lambda u_i^{\lambda-1}, \qquad u_i \neq 0, \lambda \in (0,1].$$

Finalmente, el algoritmo EM para obtener los estimadores ML en el modelo

$$Y_i \overset{\mathsf{ind}}{\sim} \mathsf{SMN}(\boldsymbol{x}_i^\top\boldsymbol{\beta},\sigma^2;\mathsf{H}), \qquad i = 1,\ldots,n,$$

adopta la forma:

Paso E: para $\boldsymbol{\theta}^{(k)}$, calcular:

$$\omega_i^{(k)} = \mathsf{E}(\omega_i|Y_i;\boldsymbol{\theta}^{(k)}), \qquad i = 1,\ldots,n.$$

Paso M: actualizar $\boldsymbol{\beta}^{(k+1)}$ y $\sigma^{2(k+1)}$ como:

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}^\top\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}^{(k)}\boldsymbol{Y},$$
$$\sigma^{2(k+1)} = \frac{1}{n}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^{(k)})^\top\boldsymbol{W}^{(k)}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^{(k)}),$$

donde $\boldsymbol{W}^{(k)} = \mathrm{diag}(\omega_1(\boldsymbol{\theta}^{(k)}),\ldots,\omega_n(\boldsymbol{\theta}^{(k)}))$.

Iteramos entre las etapas E y M hasta alcanzar convergencia, en cuyo caso hacemos $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ y $\sigma^2 = \widehat{\sigma}^2$.

Usando resultados asintóticos tradicionales asociados a la estimación máximo verosímil, podemos notar que

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0},\boldsymbol{\mathcal{F}}^{-1}(\boldsymbol{\beta})),$$

donde la matriz de información de Fisher asociada a los coeficientes de regresión $\boldsymbol{\beta}$ es dada por (ver por ejemplo, Lange, Little y Taylor, 1989)

$$\boldsymbol{\mathcal{F}}(\boldsymbol{\beta}) = \frac{4\alpha}{\phi}\boldsymbol{X}^\top\boldsymbol{X}, \qquad \alpha = \mathsf{E}\{W_h(U)U^2\},$$

donde $U = Z^2$, $Z \sim \mathsf{EC}(0, 1; h)$ con h la función generadora de densidad asociada a la mezcla de escala normal.

Apéndice A

Elementos de Álgebra Matricial

En este Apéndice se introduce la notación, definiciones y resultados básicos de álgebra lineal y matricial, esenciales para el estudio de modelos estadísticos multivariados y de regresión lineal. El material presentado a continuación puede ser hallado en textos como Graybill (1983), Ravishanker y Dey (2002) y Magnus y Neudecker (2007).

A.1. Vectores y matrices

Sea $\mathbb{R}^n$ el espacio Euclidiano n-dimensional, de este modo $\boldsymbol{x} \in \mathbb{R}^n$ representa la n-upla

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

de números reales. Note que $\boldsymbol{x}$ está *orientado* como un vector "columna", y por tanto la transpuesta de $\boldsymbol{x}$ es un vector fila,

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\top.$$

Una matriz $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ es un arreglo de números reales

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix},$$

y escribimos $\boldsymbol{A} = (a_{ij})$. Los números reales a_{ij} son llamados elementos de $\boldsymbol{A}$.

A.2. Definiciones básicas y propiedades

La suma de dos matrices del mismo orden es definida como

$$\boldsymbol{A} + \boldsymbol{B} = (a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij}),$$

el producto de una matriz por un escalar λ es

$$\lambda \boldsymbol{A} = \boldsymbol{A}\lambda = (\lambda a_{ij})$$

Resultado A.1 (Propiedades de la suma matricial). *Sean $\boldsymbol{A}, \boldsymbol{B}$ y $\boldsymbol{C}$ matrices del mismo orden y λ, μ escalares. Entonces:*

(a) $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$,
(b) $(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$,
(c) $(\lambda + \mu)\boldsymbol{A} = \lambda \boldsymbol{A} + \mu \boldsymbol{A}$,
(d) $\lambda(\boldsymbol{A} + \boldsymbol{B}) = \lambda \boldsymbol{A} + \lambda \boldsymbol{B}$,
(e) $\lambda\mu \boldsymbol{A} = (\lambda\mu)\boldsymbol{A}$.

Una matriz cuyos elementos son todos cero se denomina *matriz nula* y se denota por $\mathbf{0}$. Tenemos que

$$\boldsymbol{A} + (-1)\boldsymbol{A} = \mathbf{0}.$$

Si $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices $m \times n$ y $n \times p$, respectivamente, se define el producto de $\boldsymbol{A}$ y $\boldsymbol{B}$ como

$$\boldsymbol{AB} = \boldsymbol{C}, \quad \text{donde,} \quad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

para $i = 1, \ldots, m$ y $j = 1, \ldots, p$.

Resultado A.2 (Propiedades del producto de matrices). *Sean $\boldsymbol{A}, \boldsymbol{B}$ y $\boldsymbol{C}$ matrices de órdenes apropiados. Entonces:*

(a) $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$,
(b) $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$,
(c) $(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$.

Note que la existencia de $\boldsymbol{AB}$ *no* implica la existencia de $\boldsymbol{BA}$ y cuando ambos productos existen, en general no son iguales.

La transpuesta de una matriz $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{m \times n}$ es la matriz $n \times m$, $\boldsymbol{A}^\top$ cuyo elemento ij está dado por a_{ji}, esto es

$$\boldsymbol{A}^\top = (a_{ji}).$$

Resultado A.3 (Propiedades de la transpuesta). *Tenemos*

(a) $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$,
(b) $(\boldsymbol{A} + \boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$,
(c) $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$.

Definimos el *producto interno* entre dos vectores $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ como

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i.$$

asociado al producto interno tenemos la norma Euclidiana (o largo) de un vector $\boldsymbol{x}$ definida como

$$\|\boldsymbol{x}\| = \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2} = \Big(\sum_{i=1}^{n} x_i^2 \Big)^{1/2},$$

finalmente, la distancia Euclidiana entre dos vectores $\boldsymbol{a}$ y $\boldsymbol{b}$ se define como

$$d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|.$$

Resultado A.4 (Propiedades del producto interno). *Sean $\boldsymbol{a}$, $\boldsymbol{b}$ y $\boldsymbol{c}$ vectores n-dimensionales y λ un escalar, entonces*

(a) $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \boldsymbol{b}, \boldsymbol{a} \rangle$,
(b) $\langle \boldsymbol{a}, \boldsymbol{b} + \boldsymbol{c} \rangle = \langle \boldsymbol{a}, \boldsymbol{b} \rangle + \langle \boldsymbol{a}, \boldsymbol{c} \rangle$,
(c) $\lambda \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \lambda \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \boldsymbol{a}, \lambda \boldsymbol{b} \rangle$,
(d) $\langle \boldsymbol{a}, \boldsymbol{a} \rangle \geq 0$ *con la igualdad sólo si* $\boldsymbol{a} = \mathbf{0}$,
(e) $\|\boldsymbol{a} \pm \boldsymbol{b}\|^2 = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 \pm 2\langle \boldsymbol{a}, \boldsymbol{b} \rangle$,
(f) $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$.

Proposición A.5 (Desigualdad de Cauchy-Schwarz). $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \, \|\boldsymbol{y}\|$, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ *con la igualdad sólo si* $\boldsymbol{x} = \lambda \boldsymbol{y}$, *para algún* $\lambda \in \mathbb{R}$.

Demostración. Si $\boldsymbol{x} = \lambda \boldsymbol{y}$, el resultado es inmediato. Sino, note que

$$0 < \|\boldsymbol{x} - \lambda \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \lambda^2 \|\boldsymbol{y}\|^2 - 2\lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle, \qquad \forall \lambda \in \mathbb{R},$$

de este modo el discriminante del polinomio cuadrático debe satisfacer $4\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 - 4\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 < 0$. □

El ángulo θ entre dos vectores no nulos $\boldsymbol{x}, \boldsymbol{y}$ se define en términos de su producto interno como

$$\cos\theta = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|} = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{x}} \sqrt{\boldsymbol{y}^\top \boldsymbol{y}}},$$

dos vectores se dicen *ortogonales* sólo si $\boldsymbol{x}^\top \boldsymbol{y} = 0$.

El *producto externo* entre dos vectores $\boldsymbol{x} \in \mathbb{R}^m$ y $\boldsymbol{y} \in \mathbb{R}^n$ es la matriz $m \times n$

$$\boldsymbol{x} \wedge \boldsymbol{y} = \boldsymbol{x}\boldsymbol{y}^\top = (x_i y_j).$$

Una matriz se dice cuadrada si tiene el mismo número de filas que de columnas, una matriz cuadrada $\boldsymbol{A}$ es triangular inferior (superior) si $a_{ij} = 0$ para $i < j$ (si $a_{ij} = 0$ para $i > j$). Una matriz cuadrada $\boldsymbol{A} = (a_{ij})$ se dice *simétrica* si $\boldsymbol{A}^\top = \boldsymbol{A}$ y *sesgo-simétrica* si $\boldsymbol{A}^\top = -\boldsymbol{A}$. Para cualquier matriz cuadrada $\boldsymbol{A} = (a_{ij})$ se define $\operatorname{diag}(\boldsymbol{A})$ como

$$\operatorname{diag}(\boldsymbol{A}) = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}).$$

Si $\boldsymbol{A} = \operatorname{diag}(\boldsymbol{A})$, decimos que $\boldsymbol{A}$ es *matriz diagonal*. Un tipo particular de matriz diagonal es la identidad

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = (\delta_{ij}),$$

donde $\delta_{ij} = 1$ si $i = j$ y $\delta_{ij} = 0$ si $i \neq j$ (δ_{ij} se denomina *delta de Kronecker*). Tenemos que para $\boldsymbol{A} \in \mathbb{R}^{m \times n}$

$$\boldsymbol{I}_m \boldsymbol{A} = \boldsymbol{A} \boldsymbol{I}_n = \boldsymbol{A}.$$

Una matriz cuadrada se dice *ortogonal* si

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$$

y sus columnas son ortonormales. Note que, si

$$\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) \quad \text{con } \boldsymbol{a}_j \in \mathbb{R}^n,$$

entonces $\boldsymbol{A}$ tiene columnas *ortonormales* si

$$\boldsymbol{a}_i^\top \boldsymbol{a}_j = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{si } i \neq j, \end{cases} \qquad i, j = 1, \ldots, n.$$

Una matriz rectangular $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ puede tener la propiedad $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}_m$ ó $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_n$ pero no ambas, en cuyo caso tal matriz se denomina semi-ortogonal.

Una matriz $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, se dice idempotente si $\boldsymbol{A}^2 = \boldsymbol{A}$. Decimos que $\boldsymbol{A}$ es *matriz de proyección* si es simétrica e idempotente, esto es, $\boldsymbol{A}^\top = \boldsymbol{A}$ y $\boldsymbol{A}^2 = \boldsymbol{A}$. Considere el siguiente resultado

RESULTADO A.6. *Suponga $\boldsymbol{A}$ matriz $m \times m$, simétrica e idempotente. Entonces,*

(a) $a_{ii} \geq 0$, $i = 1, \ldots, n$.
(b) $a_{ii} \leq 1$, $i = 1, \ldots, n$.
(c) $a_{ij} = a_{ji} = 0$, *para todo* $j \neq i$, *si* $a_{ii} = 0$ *o* $a_{ii} = 1$.

DEMOSTRACIÓN. Como $\boldsymbol{A}$ es simétrica e idempotente, tenemos

$$\boldsymbol{A} = \boldsymbol{A}^2 = \boldsymbol{A}^\top \boldsymbol{A},$$

de ahí que

$$a_{ii} = \sum_{j=1}^{n} a_{ji}^2,$$

que claramente es no negativo. Además, podemos escribir

$$a_{ii} = a_{ii}^2 + \sum_{j\neq i} a_{ji}^2.$$

Por tanto, $a_{ii} \geq a_{ii}^2$ y de este modo (b) es satisfecha. Si $a_{ii} = 0$ o bien $a_{ii} = 1$, entonces $a_{ii} = a_{ii}^2$ y debemos tener

$$\sum_{j\neq i} a_{ji}^2 = 0,$$

lo que junto con las simetría de $\boldsymbol{A}$, establece (c). □

Cualquier matriz $\boldsymbol{B}$ satisfaciendo

$$\boldsymbol{B}^2 = \boldsymbol{A}$$

se dice *raíz cuadrada* de $\boldsymbol{A}$ y se denota como $\boldsymbol{A}^{1/2}$ tal matriz *no* necesita ser única.

A.2.1. Formas lineales y cuadráticas. Sea $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ y $\boldsymbol{B} \in \mathbb{R}^{n\times m}$. La expresión $\boldsymbol{a}^\top \boldsymbol{x}$ se dice una *forma lineal* en $\boldsymbol{x}$ y $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ una *forma cuadrática*, mientras que $\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{y}$ es una forma bilineal.

Sin pérdida de generalidad se asumirá que la matriz asociada a la forma cuadrática $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ es simétrica. Note que siempre es posible

$$\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} = \tfrac{1}{2}\boldsymbol{x}^\top (\boldsymbol{A}^\top + \boldsymbol{A})\boldsymbol{x},$$

en cuyo caso tenemos que $\boldsymbol{B}$ es matriz simétrica.

Decimos que una matriz simétrica $\boldsymbol{A}$ es definida positiva (negativa) si $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$ ($\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} < 0$) para todo $\boldsymbol{x} \neq \boldsymbol{0}$. Cuando $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$ ($\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \leq 0$) $\forall \boldsymbol{x}$ decimos que $\boldsymbol{A}$ es semidefinida positiva (negativa).

Note que las matrices $\boldsymbol{B}^\top \boldsymbol{B}$ y $\boldsymbol{B}\boldsymbol{B}^\top$ son semidefinidas positivas y que $\boldsymbol{A}$ es (semi)definida negativa sólo si $-\boldsymbol{A}$ es (semi)definida positiva.

RESULTADO A.7. *Sea $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times p}$ y $\boldsymbol{C} \in \mathbb{R}^{n\times p}$ y $\boldsymbol{x}$ vector n-dimensional. Entonces*

(a) $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$,
(b) $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{0} \Leftrightarrow \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B} = \boldsymbol{0}$,
(c) $\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B} = \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{C} \Leftrightarrow \boldsymbol{A}\boldsymbol{B} = \boldsymbol{A}\boldsymbol{C}$.

DEMOSTRACIÓN. (a) Claramente $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Rightarrow \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. Por otro lado, si $\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, entonces $\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^\top\boldsymbol{A}\boldsymbol{x} = 0$ y de ahí que $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. (b) sigue desde (a). Finalmente, (c) sigue desde (b) mediante substituir $\boldsymbol{B} - \boldsymbol{C}$ por $\boldsymbol{B}$ en (c). □

RESULTADO A.8. *Sean $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ y $\boldsymbol{B}$, $\boldsymbol{C}$ matrices $n\times n$ con $\boldsymbol{B}$ simétrica. Entonces*

(a) $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, $\forall \boldsymbol{x} \in \mathbb{R}^n$ *sólo si* $\boldsymbol{A} = \boldsymbol{0}$,
(b) $\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x} = 0$, $\forall \boldsymbol{x} \in \mathbb{R}^n$ *sólo si* $\boldsymbol{B} = \boldsymbol{0}$,
(c) $\boldsymbol{x}^\top\boldsymbol{C}\boldsymbol{x} = 0$, $\forall \boldsymbol{x} \in \mathbb{R}^n$ *sólo si* $\boldsymbol{C}^\top = -\boldsymbol{C}$.

A.2.2. Rango de una matriz. Un conjunto de vectores $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ se dice *linealmente independiente* si $\sum_i \alpha_i\boldsymbol{x}_i = \boldsymbol{0}$ implica que *todos* los $\alpha_i = 0$. Si $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *no* son linealmente independientes, ellos se dicen *linealmente dependientes*.

Sea $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, el *rango* columna (fila) de $\boldsymbol{A}$ es el número de columnas (filas) linealmente independientes. Denotamos el rango de $\boldsymbol{A}$ como

$$\text{rg}(\boldsymbol{A}),$$

note que

$$\text{rg}(\boldsymbol{A}) \leq \text{mín}(m, n).$$

Si $\text{rg}(\boldsymbol{A}) = n$ decimos que $\boldsymbol{A}$ tiene rango columna completo. Si $\text{rg}(\boldsymbol{A}) = 0$, entonces $\boldsymbol{A}$ es la matriz nula. Por otro lado, si $\boldsymbol{A} = \boldsymbol{0}$, entonces $\text{rg}(\boldsymbol{A}) = 0$.

RESULTADO A.9 (Propiedades del rango). *Sea $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ y $\boldsymbol{B}$, $\boldsymbol{C}$ matrices de órdenes apropiados, entonces*

(a) $\text{rg}(\boldsymbol{A}) = \text{rg}(\boldsymbol{A}^\top) = \text{rg}(\boldsymbol{A}^\top\boldsymbol{A}) = \text{rg}(\boldsymbol{A}\boldsymbol{A}^\top)$,
(b) $\text{rg}(\boldsymbol{A}\boldsymbol{B}) \leq \text{mín}\{\text{rg}(\boldsymbol{A}), \text{rg}(\boldsymbol{B})\}$,
(c) $\text{rg}(\boldsymbol{B}\boldsymbol{A}\boldsymbol{C}) = \text{rg}(\boldsymbol{A})$ *si* $\boldsymbol{B}$ *y* $\boldsymbol{C}$ *son matrices de rango completo,*
(d) $\text{rg}(\boldsymbol{A} + \boldsymbol{B}) \leq \text{rg}(\boldsymbol{A}) + \text{rg}(\boldsymbol{B})$,
(e) *si* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *y* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ *para algún* $\boldsymbol{x} \neq \boldsymbol{0}$, *entonces* $\text{rg}(\boldsymbol{A}) \leq n - 1$.

El *espacio columna* de $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, denotado por $\mathcal{M}(\boldsymbol{A})$, es el conjunto de vectores

$$\mathcal{M}(\boldsymbol{A}) = \{\boldsymbol{y} : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \text{ para algún } \boldsymbol{x} \in \mathbb{R}^n\}.$$

De este modo, $\mathcal{M}(\boldsymbol{A})$ es el espacio vectorial generado por las columnas de $\boldsymbol{A}$. La dimensión de este espacio es $\text{rg}(\boldsymbol{A})$. Se tiene que

$$\mathcal{M}(\boldsymbol{A}) = \mathcal{M}(\boldsymbol{A}\boldsymbol{A}^\top)$$

para cualquier matriz $\boldsymbol{A}$.

El *espacio nulo*, $\mathcal{N}(\boldsymbol{A})$, de una matriz $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ consiste de todos los vectores n-dimensionales $\boldsymbol{x}$, tal que $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, esto es,

$$\mathcal{N}(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^n \text{ tal que } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

Note que, el espacio nulo es el conjunto de todas las soluciones del sistema lineal homogéneo $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. $\mathcal{N}(\boldsymbol{A})$ es un subespacio de $\mathbb{R}^n$ y su dimensión se denomina *nulidad* de $\boldsymbol{A}$. Además $\mathcal{N}(\boldsymbol{A}) = \{\mathcal{M}(\boldsymbol{A})\}^\perp$. Finalmente, considere la siguiente proposición

RESULTADO A.10. *Para cualquier matriz $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, entonces $n = \dim(\mathcal{N}(\boldsymbol{A})) + \text{rg}(\boldsymbol{A})$.*

A.2.3. Matriz inversa. Sea $\boldsymbol{A}$ una matriz cuadrada de orden $n \times n$. Decimos que $\boldsymbol{A}$ es *no singular* si $\operatorname{rg}(\boldsymbol{A}) = n$, y que $\boldsymbol{A}$ es *singular* si $\operatorname{rg}(\boldsymbol{A}) < n$. De este modo, si $\boldsymbol{A}$ es no singular, entonces existe una matriz no singular $\boldsymbol{B}$ tal que

$$\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}_n.$$

La matriz $\boldsymbol{B}$, denotada $\boldsymbol{A}^{-1}$ es única y se denomina *inversa* de $\boldsymbol{A}$.

RESULTADO A.11 (Propiedades de la inversa). *Siempre que todas las matrices inversas involucradas existan, tenemos que*

(a) $(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$.
(b) $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
(c) $(\lambda \boldsymbol{A})^{-1} = \frac{1}{\lambda}\boldsymbol{A}^{-1}$.
(d) $\boldsymbol{P}^{-1} = \boldsymbol{P}^\top$, *si* $\boldsymbol{P}$ *es matriz ortogonal.*
(e) *Si* $\boldsymbol{A} > 0$, *entonces* $\boldsymbol{A}^{-1} > 0$.
(f) *(Teorema de Sherman-Morrison-Woodbury)*

$$(\boldsymbol{A} + \boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1},$$

donde $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ *y* $\boldsymbol{D}$ *son matrices* $m \times m$, $m \times n$, $n \times n$ *y* $n \times m$, *respectivamente.*
(g) *Si* $1 \pm \boldsymbol{v}^\top \boldsymbol{A}^{-1}\boldsymbol{u} \neq 0$, *entonces*

$$(\boldsymbol{A} \pm \boldsymbol{uv}^\top)^{-1} = \boldsymbol{A}^{-1} \mp \frac{\boldsymbol{A}^{-1}\boldsymbol{uv}^\top\boldsymbol{A}^{-1}}{1 \pm \boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}},$$

es conocida como la fórmula de Sherman-Morrison.
(h) $(\boldsymbol{I} + \lambda\boldsymbol{A})^{-1} = \boldsymbol{I} + \sum_{i=1}^{\infty}(-1)^i\lambda^i\boldsymbol{A}^i$.

A.2.4. Determinante de una matriz. El determinante de una matriz corresponde a la función $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$, denotada comúnmente como $|\boldsymbol{A}| = \det(\boldsymbol{A})$ y definida como

$$|\boldsymbol{A}| = \sum(-1)^{\sigma(j_1,\ldots,j_n)} \prod_{i=1}^{n} a_{ij_i}$$

donde la sumatoria es tomada sobre todas las permutaciones $(j_1, \ldots, j_n)$ del conjunto de enteros $(1, \ldots, n)$, y $\sigma(j_1, \ldots, j_n)$ es el número de transposiciones necesarias para cambiar $(1, \ldots, n)$ en $(j_1, \ldots, j_n)$ (una transposición consiste en intercambiar dos números).

Una *submatriz* de $\boldsymbol{A}$ es un arreglo rectangular obtenido mediante eliminar filas y columnas de $\boldsymbol{A}$. Un *menor* es el determinante de una submatriz cuadrada de $\boldsymbol{A}$. El menor asociado al elemento a_{ij} es el determinante de la submatriz de $\boldsymbol{A}$ obtenida por eliminar su i-ésima fila y j-ésima columna. El *cofactor* de a_{ij}, digamos c_{ij} es $(-1)^{i+j}$ veces el menor de a_{ij}. La matriz $\boldsymbol{C} = (c_{ij})$ se denomina matriz cofactor de $\boldsymbol{A}$. La transpuesta de $\boldsymbol{C}$ es llamada *adjunta* de $\boldsymbol{A}$ y se denota $\boldsymbol{A}^{\#}$. Tenemos que

$$|\boldsymbol{A}| = \sum_{j=1}^{n} a_{ij}c_{ij} = \sum_{j=1}^{n} a_{jk}c_{jk}, \quad \text{para } i, k = 1, \ldots, n.$$

RESULTADO A.12 (Propiedades del determinante). *Sea* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *y* λ *un escalar. Entonces*

(a) $|\boldsymbol{A}| = |\boldsymbol{A}^\top|$.
(b) $|\boldsymbol{AB}| = |\boldsymbol{A}|\,|\boldsymbol{B}|$.

(c) $|\lambda \boldsymbol{A}| = \lambda^n |\boldsymbol{A}|$.
(d) $|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$, *si* $\boldsymbol{A}$ *es no singular.*
(e) *Si* $\boldsymbol{A}$ *es matriz triangular, entonces* $|\boldsymbol{A}| = \prod_{i=1}^{n} a_{ii}$.
(f) *El resultado en (e) también es válido para* $\boldsymbol{A} = \operatorname{diag}(\boldsymbol{A})$. *Además, es evidente que* $|\boldsymbol{I}_n| = 1$.
(g) *Si* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *y* $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, *entonces* $|\boldsymbol{I}_m + \boldsymbol{AB}| = |\boldsymbol{I}_n + \boldsymbol{BA}|$.

A.2.5. La traza de una matriz. La *traza* de una matriz cuadrada $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, denotada por $\operatorname{tr}(\boldsymbol{A})$, es la suma de sus elementos diagonales:

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}.$$

Resultado A.13 (Propiedades de la traza). *Siempre que las operaciones matriciales están definidas*

(a) $\operatorname{tr}(\boldsymbol{A} + \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$,
(b) $\operatorname{tr}(\lambda \boldsymbol{A}) = \lambda \operatorname{tr}(\boldsymbol{A})$ *si* λ *es un escalar,*
(c) $\operatorname{tr}(\boldsymbol{A}^\top) = \operatorname{tr}(\boldsymbol{A})$,
(d) $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$ *(propiedad cíclica de la traza),*
(e) $\operatorname{tr}(\boldsymbol{A}) = 0$ *si* $\boldsymbol{A} = \boldsymbol{0}$.

Note en (d) que aunque ambas $\boldsymbol{AB}$ y $\boldsymbol{BA}$ son cuadradas, no necesitan ser del mismo orden.
Además, es directo que la normal vectorial (Euclidiana), satisface

$$\|\boldsymbol{x}\| = (\boldsymbol{x}^\top \boldsymbol{x})^{1/2} = (\operatorname{tr} \boldsymbol{x}\boldsymbol{x}^\top)^{1/2},$$

de este modo, podemos definir una normal matricial (Euclidiana) como

$$\|\boldsymbol{A}\| = (\operatorname{tr} \boldsymbol{A}^\top \boldsymbol{A})^{1/2}.$$

En efecto, se tiene que $\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A}) \geq 0$ con la igualdad sólo si $\boldsymbol{A} = \boldsymbol{0}$.

A.2.6. Valores y vectores propios. Si $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices reales del mismo orden, una matriz compleja $\boldsymbol{Z}$ puede ser definida como

$$\boldsymbol{Z} = \boldsymbol{A} + i\boldsymbol{B},$$

donde i denota la unidad imaginaria que satisface $i^2 = -1$. El conjugado complejo de $\boldsymbol{Z}$, denotado por $\boldsymbol{Z}^H$, se define como

$$\boldsymbol{Z}^H = \boldsymbol{A}^\top - i\boldsymbol{B}^\top.$$

Una matriz $\boldsymbol{Z} \in \mathbb{C}^{n\times n}$ se dice *Hermitiana* si $\boldsymbol{Z}^H = \boldsymbol{Z}$ (equivalente complejo de una matriz simétrica) y *unitaria* si $\boldsymbol{Z}^H \boldsymbol{Z} = \boldsymbol{I}$ (equivalente complejo de una matriz ortogonal).

Sea $\boldsymbol{A}$ una matriz cuadrada $n \times n$. Los *valores propios* de $\boldsymbol{A}$ son definidos como las raíces de la *ecuación característica*

$$|\lambda \boldsymbol{I} - \boldsymbol{A}| = 0,$$

la ecuación anterior tiene n raíces, en general complejas y posiblemente con algunas repeticiones (multiplicidad). Sea λ un valor propio de $\boldsymbol{A}$, entonces existe un vector $\boldsymbol{v} \neq \boldsymbol{0} \in \mathbb{C}^n$ tal que $(\lambda \boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} = \boldsymbol{0}$, esto es,

$$\boldsymbol{Av} = \lambda \boldsymbol{v}.$$

el vector $\boldsymbol{v}$ se denomina *vector propio* asociado al valor propio λ. Note que, si $\boldsymbol{v}$ es un vector propio, también lo es $\alpha\boldsymbol{v}$, $\forall\alpha \in \mathbb{C}$, y en particular $\boldsymbol{v}/\|\boldsymbol{v}\|$ es un vector propio normalizado.

RESULTADO A.14. *Si $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ es matriz Hermitiana, entonces todos sus valores propios son reales*

RESULTADO A.15. *Si $\boldsymbol{A}$ es matriz cuadrada $n \times n$ y $\boldsymbol{G}$ es matriz no singular $n \times n$, entonces $\boldsymbol{A}$ y $\boldsymbol{G}^{-1}\boldsymbol{A}\boldsymbol{G}$ tienen el mismo conjunto de valores propios (con las mismas multiplicidades)*

DEMOSTRACIÓN. Note que

$$|\lambda\boldsymbol{I} - \boldsymbol{G}^{-1}\boldsymbol{A}\boldsymbol{G}| = |\lambda\boldsymbol{G}^{-1}\boldsymbol{G} - \boldsymbol{G}^{-1}\boldsymbol{A}\boldsymbol{G}| = |\boldsymbol{G}^{-1}||\lambda\boldsymbol{I} - \boldsymbol{A}||\boldsymbol{G}| = |\lambda\boldsymbol{I} - \boldsymbol{A}|$$

□

RESULTADO A.16. *Una matriz singular tiene al menos un valor propio cero*

DEMOSTRACIÓN. Si $\boldsymbol{A}$ es matriz singular, entonces $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$ para algún $\boldsymbol{v} \neq \boldsymbol{0}$, luego desde $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, tenemos que $\lambda = 0$. □

RESULTADO A.17. *Una matriz simétrica es definida positiva (semidefinida positiva) sólo si todos sus valores propios son positivos (no-negativos).*

DEMOSTRACIÓN. Si $\boldsymbol{A}$ es definida positiva y $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, entonces $\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}^\top\boldsymbol{v}$. Ahora, como $\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} > 0$ y $\boldsymbol{v}^\top\boldsymbol{v} > 0$ implica $\lambda > 0$. La conversa no será probada aquí. □

RESULTADO A.18. *Una matriz idempotente sólo tiene valores propios 0 ó 1. Todos los valores propios de una matriz unitaria tienen modulo 1*

DEMOSTRACIÓN. Sea $\boldsymbol{A}$ matriz idempotente, esto es, $\boldsymbol{A}^2 = \boldsymbol{A}$. De este modo, si $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, entonces

$$\lambda\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} = \boldsymbol{A}^2\boldsymbol{v} = \lambda\boldsymbol{A}\boldsymbol{v} = \lambda^2\boldsymbol{v}$$

y de ahí que $\lambda = \lambda^2$, esto implica que $\lambda = 0$ ó $\lambda = 1$.
Por otro lado, si $\boldsymbol{A}$ es unitaria, entonces $\boldsymbol{a}^H\boldsymbol{A} = \boldsymbol{I}$. De este modo, si $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, entonces

$$\boldsymbol{v}^H\boldsymbol{A}^H = \overline{\lambda}\boldsymbol{v}^H,$$

luego

$$\boldsymbol{v}^H\boldsymbol{v} = \boldsymbol{v}^H\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{v} = \overline{\lambda}\lambda\boldsymbol{v}^H\boldsymbol{v}.$$

Como $\boldsymbol{v}^H\boldsymbol{v} \neq 0$, obtenemos que $\overline{\lambda}\lambda = 1$ y de ahí que $|\lambda| = 1$. □

RESULTADO A.19 (Propiedades de la matrices idempotentes). *Sea $\boldsymbol{A}$ matriz $n \times n$, entonces*

(a) *$\boldsymbol{A}^\top$ y $\boldsymbol{I} - \boldsymbol{A}$ son idempotentes sólo si $\boldsymbol{A}$ es idempotente,*
(b) *si $\boldsymbol{A}$ es idempotente, entonces $\mathrm{rg}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}) = r$. Si $\mathrm{rg}(\boldsymbol{A}) = n$, entonces $\boldsymbol{A} = \boldsymbol{I}$.*

RESULTADO A.20. *Si $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ es matriz Hermitiana y $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ son vectores propios asociados a λ_1 y λ_2, respectivamente, donde $\lambda_1 \neq \lambda_2$. Entonces $\boldsymbol{v}_1 \perp \boldsymbol{v}_2$.*

El resultado anterior muestra que si todos los valores propios de una matriz Hermitiana $\boldsymbol{A}$ son distintos, entonces existe una base ortonormal de vectores propios tal que $\boldsymbol{A}$ es diagonalizable.

PROPOSICIÓN A.21 (Descomposición de Schur). *Sea $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Entonces existe una matriz unitaria $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ y una matriz triangular $\boldsymbol{M}$ cuyos elementos diagonales son los valores propios de $\boldsymbol{A}$, tal que*

$$\boldsymbol{U}^H \boldsymbol{A}\boldsymbol{U} = \boldsymbol{M}.$$

PROPOSICIÓN A.22 (Descomposición espectral). *Sea $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ matriz Hermitiana. Entonces existe una matriz unitaria $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ tal que*

$$\boldsymbol{U}^H \boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda},$$

donde $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ es matriz diagonal cuyos elementos diagonales son los valores propios de $\boldsymbol{A}$.

Para aplicaciones en Estadística siempre haremos uso de la Proposición A.22 considerando $\boldsymbol{A}$ matriz simétrica, en cuyo caso todos sus valores propios serán reales y $\boldsymbol{U}$ será una matriz ortogonal. Para $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ matriz ortogonal, denotamos el grupo de matrices ortogonales como

$$\mathcal{O}_n = \{\boldsymbol{Q} \in \mathbb{R}^{n\times n} : \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}\}$$

Note que si $\boldsymbol{A}$ es matriz simétrica y definida positiva, entonces

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = (\boldsymbol{U}\boldsymbol{\Lambda}^{1/2})(\boldsymbol{U}\boldsymbol{\Lambda}^{1/2})^\top = (\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top)^2$$

donde $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ y $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\boldsymbol{\lambda}^{1/2})$. Por tanto,

$$\boldsymbol{A} = \boldsymbol{M}\boldsymbol{M}^\top, \quad \text{con} \quad \boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2},$$

o bien,

$$\boldsymbol{A} = \boldsymbol{B}^2, \quad \text{con} \quad \boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top,$$

esto es, $\boldsymbol{B}$ es una matriz raíz cuadrada de $\boldsymbol{A}$.

RESULTADO A.23. *Sea $\boldsymbol{A}$ matriz simétrica $n \times n$, con valores propios $\lambda_1, \ldots, \lambda_n$. Entonces*

(a) $\text{tr}(\boldsymbol{A}) = \sum_{i=1}^n \lambda_i$,
(b) $|\boldsymbol{A}| = \prod_{i=1}^n \lambda_i$.

DEMOSTRACIÓN. Usando que $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. Tenemos

$$\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top) = \text{tr}(\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{U}) = \text{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^n \lambda_i$$

y

$$|\boldsymbol{A}| = |\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top| = |\boldsymbol{U}||\boldsymbol{\Lambda}||\boldsymbol{U}^\top| = |\boldsymbol{\Lambda}| = \prod_{i=1}^n \lambda_i$$

□

RESULTADO A.24. *Si $\boldsymbol{A}$ es una matriz simétrica con r valores propios distintos de cero, entonces* $\text{rg}(\boldsymbol{A}) = r$.

DEMOSTRACIÓN. Tenemos que $\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda}$ y de ahí que

$$\text{rg}(\boldsymbol{A}) = \text{rg}(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top) = \text{rg}(\boldsymbol{\Lambda}) = r$$

□

A.2.7. Matrices (semi)definidas positivas.

PROPOSICIÓN A.25. *Sea $\boldsymbol{A}$ matriz definida positiva y $\boldsymbol{B}$ semidefinida positiva. Entonces*

$$|\boldsymbol{A}+\boldsymbol{B}| \geq |\boldsymbol{A}|,$$

con la igualdad sólo si $\boldsymbol{B}=\boldsymbol{0}$.

DEMOSTRACIÓN. Tenemos $\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda}$, con $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$ y $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}$. Luego,

$$\boldsymbol{A}+\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}(\boldsymbol{I}+\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2})\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top,$$

de este modo

$$\begin{aligned}|\boldsymbol{A}+\boldsymbol{B}| &= |\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}||\boldsymbol{I}+\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}||\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top| \\ &= |\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top||\boldsymbol{I}+\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}| \\ &= |\boldsymbol{A}||\boldsymbol{I}+\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}|.\end{aligned}$$

Si $\boldsymbol{B}=\boldsymbol{0}$, tenemos $|\boldsymbol{A}+\boldsymbol{B}| = |\boldsymbol{A}|$. Por otro lado, si $\boldsymbol{B}\neq\boldsymbol{0}$. Entonces la matriz $\boldsymbol{I}+\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}$ tendrá al menos un valor propio no nulo y por tanto, $|\boldsymbol{I}+\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}| > 1$, esto es $|\boldsymbol{A}+\boldsymbol{B}| > |\boldsymbol{A}|$. □

Para dos matrices simétricas $\boldsymbol{A}$ y $\boldsymbol{B}$, escribimos $\boldsymbol{A}\geq\boldsymbol{B}$ si $\boldsymbol{A}-\boldsymbol{B}$ es semidefinida positiva. Análogamente, escribimos $\boldsymbol{A}>\boldsymbol{B}$ si $\boldsymbol{A}-\boldsymbol{B}$ es definida positiva.

RESULTADO A.26. *Sean $\boldsymbol{A}$, $\boldsymbol{B}$ matrices definidas positivas $n\times n$. Entonces $\boldsymbol{A}>\boldsymbol{B}$ sólo si $\boldsymbol{B}^{-1}>\boldsymbol{A}^{-1}$.*

PROPOSICIÓN A.27. *Sean $\boldsymbol{A}$ y $\boldsymbol{B}$ matrices definidas positivas y $\boldsymbol{A}-\boldsymbol{B}\geq 0$. Entonces $|\boldsymbol{A}|\geq|\boldsymbol{B}|$ con la igualdad sólo si $\boldsymbol{A}=\boldsymbol{B}$.*

DEMOSTRACIÓN. Sea $\boldsymbol{C}=\boldsymbol{A}-\boldsymbol{B}$. Como $\boldsymbol{B}$ es definida positiva y $\boldsymbol{C}$ es semidefinida positiva, tenemos por la Proposición A.25 que $|\boldsymbol{B}+\boldsymbol{C}|\geq|\boldsymbol{B}|$, con la igualdad sólo si $\boldsymbol{C}=\boldsymbol{0}$. □

A.2.8. Descomposiciones matriciales.

PROPOSICIÓN A.28 (Descomposición LDL). *Si $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ es matriz simétrica y no singular, entonces existe $\boldsymbol{L}$ matriz triangular inferior y $\boldsymbol{D}=\operatorname{diag}(d_1,\ldots,d_n)$, tal que*

$$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\top.$$

PROPOSICIÓN A.29 (Descomposición Cholesky). *Si $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ es simétrica y definida positiva, entonces existe una única matriz triangular inferior $\boldsymbol{G}\in\mathbb{R}^{n\times n}$ (factor Cholesky) con elementos diagonales positivos, tal que*

$$\boldsymbol{A}=\boldsymbol{G}\boldsymbol{G}^\top.$$

PROPOSICIÓN A.30 (Descomposición ortogonal-triangular). *Sea $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, entonces existe $\boldsymbol{Q}\in\mathcal{O}_m$ y $\boldsymbol{R}\in\mathbb{R}^{m\times n}$, tal que*

$$\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{R},$$

donde

$$\boldsymbol{R}=\begin{pmatrix}\boldsymbol{R}_1\\ \boldsymbol{0}\end{pmatrix}$$

con $\boldsymbol{R}_1\in\mathbb{R}^{n\times n}$ matriz triangular superior, aquí suponemos que $m\geq n$. Si $\operatorname{rg}(\boldsymbol{A})=r$, entonces las primeras n columnas de $\boldsymbol{Q}$ forman una base ortonormal para $\mathcal{M}(\boldsymbol{A})$.

Note que, si $\boldsymbol{A} = \boldsymbol{QR}$ entonces

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{QR} = \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{R}_1^\top \boldsymbol{R}_1,$$

y $\boldsymbol{R}_1$ corresponde al factor Cholesky de $\boldsymbol{A}^\top \boldsymbol{A}$.

PROPOSICIÓN A.31 (Descomposición valor singular). *Sea $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ con $\operatorname{rg}(\boldsymbol{A}) = r$, entonces existen matrices $\boldsymbol{U} \in \mathcal{O}_m$, $\boldsymbol{V} \in \mathcal{O}_n$, tal que*

$$\boldsymbol{A} = \boldsymbol{U} \begin{pmatrix} \boldsymbol{D}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{V}^\top,$$

*donde $\boldsymbol{D}_r = \operatorname{diag}(\delta_1, \ldots, \delta_r)$ con $\delta_i > 0$ para $i = 1, \ldots, r$, llamados **valores singulares** de $\boldsymbol{A}$.*

A.2.9. Matrices particionadas. Sea $\boldsymbol{A}$ una matriz $m \times n$. Considere particionar $\boldsymbol{A}$ como sigue

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}, \tag{A.1}$$

donde $\boldsymbol{A}_{11} \in \mathbb{R}^{m_1\times n_1}$, $\boldsymbol{A}_{12} \in \mathbb{R}^{m_1\times n_2}$, $\boldsymbol{A}_{21} \in \mathbb{R}^{m_2\times n_1}$, $\boldsymbol{A}_{22} \in \mathbb{R}^{m_2\times n_2}$, y $m_1 + m_2 = m$, $n_1 + n_2 = n$.

Sea $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ particionada de manera análoga a $\boldsymbol{A}$, entonces

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & \boldsymbol{A}_{12} + \boldsymbol{B}_{12} \\ \boldsymbol{A}_{21} + \boldsymbol{B}_{21} & \boldsymbol{A}_{22} + \boldsymbol{B}_{22} \end{pmatrix}.$$

Ahora, considere $\boldsymbol{C} \in \mathbb{R}^{n\times p}$ particionada en submatrices $\boldsymbol{C}_{ij}$, para $i, j = 1, 2$ con dimensiones adecuadas, entonces

$$\boldsymbol{AC} = \begin{pmatrix} \boldsymbol{A}_{11}\boldsymbol{C}_{11} + \boldsymbol{A}_{12}\boldsymbol{C}_{21} & \boldsymbol{A}_{11}\boldsymbol{C}_{12} + \boldsymbol{A}_{12}\boldsymbol{C}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{C}_{11} + \boldsymbol{A}_{22}\boldsymbol{C}_{21} & \boldsymbol{A}_{21}\boldsymbol{C}_{12} + \boldsymbol{A}_{22}\boldsymbol{C}_{22} \end{pmatrix}.$$

La transpuesta de $\boldsymbol{A}$ está dada por

$$\boldsymbol{A}^\top = \begin{pmatrix} \boldsymbol{A}_{11}^\top & \boldsymbol{A}_{21}^\top \\ \boldsymbol{A}_{12}^\top & \boldsymbol{A}_{22}^\top \end{pmatrix}.$$

Si $\boldsymbol{A}_{12}$ y $\boldsymbol{A}_{21}$ son matrices nulas y si ambas $\boldsymbol{A}_{11}$ y $\boldsymbol{A}_{22}$ son matrices no singulares, entonces la inversa de $\boldsymbol{A}$ es

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}_{11}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{pmatrix}.$$

En general, si $\boldsymbol{A}$ es matriz no singular particionada como en (A.1) y $\boldsymbol{D} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$ también es no singular, entonces

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{D}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{D}^{-1} \end{pmatrix}.$$

Por otro lado, si $\boldsymbol{A}$ es no singular y $\boldsymbol{E} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ es no singular, entonces

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{E}^{-1} & -\boldsymbol{E}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{E}^{-1} & \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{E}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \end{pmatrix}.$$

Considere el determinante

$$\begin{vmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{vmatrix} = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22}| = \begin{vmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{vmatrix},$$

si $\boldsymbol{A}_{11}$ y $\boldsymbol{A}_{22}$ son matrices cuadradas.

Ahora, para una matriz particionada como en (A.1) con $m_1 = n_1$ y $m_2 = n_2$, tenemos

$$|\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}| = |\boldsymbol{A}_{22}||\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}|,$$

si $\boldsymbol{A}_{11}$ y $\boldsymbol{A}_{22}$ son matrices no singulares.

A.3. Inversa generalizada y sistemas de ecuaciones lineales

En esta sección se generaliza el concepto de invertibilidad para matrices singulares así como para matrices rectangulares. En particular, introducimos la inversa Moore-Penrose (MP), generalización que permite resolver de forma explícita un sistema de ecuaciones lineales.

A.3.1. Inversa Moore-Penrose. Sea $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, la inversa Moore-Penrose, $\boldsymbol{G} \in \mathbb{R}^{n\times m}$ debe satisfacer las siguientes condiciones

$$\boldsymbol{AGA} = \boldsymbol{A}, \tag{A.2}$$

$$\boldsymbol{GAG} = \boldsymbol{G}, \tag{A.3}$$

$$(\boldsymbol{AG})^\top = \boldsymbol{AG}, \tag{A.4}$$

$$(\boldsymbol{GA})^\top = \boldsymbol{GA}. \tag{A.5}$$

La inversa MP de $\boldsymbol{A}$ se denota comunmente como $\boldsymbol{A}^+$. Si $\boldsymbol{G}$ satisface sólo la condición en (A.2) entonces decimos que $\boldsymbol{G}$ es *una* inversa generalizada y la denotamos por $\boldsymbol{A}^-$.

Proposición A.32 (Unicidad de la inversa MP). *Para cada $\boldsymbol{A}$, existe una única $\boldsymbol{A}^+$.*

Resultado A.33 (Propiedades de la inversa MP).

(a) $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$ *para* $\boldsymbol{A}$ *matriz no singular,*
(b) $(\boldsymbol{A}^+)^+ = \boldsymbol{A}$,
(c) $(\boldsymbol{A}^\top)^+ = (\boldsymbol{A}^+)^\top$,
(d) $\boldsymbol{A}^+ = \boldsymbol{A}$ *si* $\boldsymbol{A}$ *es simétrica e idempotente,*
(e) $\boldsymbol{AA}^+$ *y* $\boldsymbol{A}^+\boldsymbol{A}$ *son idempotentes,*
(f) $\mathrm{rg}(\boldsymbol{A}) = \mathrm{rg}(\boldsymbol{A}^+) = \mathrm{rg}(\boldsymbol{AA}^+) = \mathrm{rg}(\boldsymbol{A}^+\boldsymbol{A})$,
(g) $\boldsymbol{A}^\top\boldsymbol{AA}^+ = \boldsymbol{A} = \boldsymbol{A}^+\boldsymbol{AA}^\top$,
(h) $\boldsymbol{A}^\top\boldsymbol{A}^{+\top}\boldsymbol{A}^+ = \boldsymbol{A}^+ = \boldsymbol{A}^+\boldsymbol{A}^{+\top}\boldsymbol{A}^\top$,
(i) $\boldsymbol{A}^+ = (\boldsymbol{A}^\top\boldsymbol{A})^+\boldsymbol{A}^\top = \boldsymbol{A}^\top(\boldsymbol{AA}^\top)^+$,
(j) $\boldsymbol{A}^+ = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$, *si* $\boldsymbol{A}$ *tiene rango columna completo,*
(k) $\boldsymbol{A}^+ = \boldsymbol{A}^\top(\boldsymbol{AA}^\top)^{-1}$, *si* $\boldsymbol{A}$ *tiene rango fila completo.*

A.3.2. Solución de sistemas de ecuaciones lineales. La solución general de un sistema de ecuaciones homegéneo $\boldsymbol{Ax} = \boldsymbol{0}$ es

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})\boldsymbol{q},$$

con $\boldsymbol{q}$ un vector arbitrário. La solución de $\boldsymbol{Ax} = \boldsymbol{0}$ es única sólo si $\boldsymbol{A}$ tiene rango columna completo, esto es, $\boldsymbol{A}^\top\boldsymbol{A}$ es no singular. El sistema homogéneo $\boldsymbol{Ax} = \boldsymbol{0}$ siempre tiene al menos una solución, digamos $\boldsymbol{x} = \boldsymbol{0}$.

El sistema no homogéneo

$$\boldsymbol{Ax} = \boldsymbol{b},$$

tendrá al menos una solución si es *consistente*.

PROPOSICIÓN A.34. *Sea* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ *y* $\boldsymbol{b}$ *vector* $m \times 1$. *Entonces son equivalentes:*

(a) *la ecuación* $\boldsymbol{Ax} = \boldsymbol{b}$ *tiene una solución para* $\boldsymbol{x}$,
(b) $\boldsymbol{b} \in \mathcal{M}(\boldsymbol{A})$,
(c) $\mathrm{rg}(\boldsymbol{A} : \boldsymbol{b}) = \mathrm{rg}(\boldsymbol{A})$,
(d) $\boldsymbol{AA}^{+}\boldsymbol{b} = \boldsymbol{b}$.

PROPOSICIÓN A.35. *Una condición necesaria y suficiente para que la ecuación* $\boldsymbol{Ax} = \boldsymbol{b}$ *tenga una solución es que*

$$\boldsymbol{AA}^{+}\boldsymbol{b} = \boldsymbol{b},$$

en cuyo caso la solución general está dada por

$$\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{b} + (\boldsymbol{I} - \boldsymbol{A}^{+}\boldsymbol{A})\boldsymbol{q},$$

donde $\boldsymbol{q}$ *es un vector arbitrário.*

Si el sistema $\boldsymbol{Ax} = \boldsymbol{b}$ es consistente, entonces tendrá *solución única* sólo si $\boldsymbol{A}$ es de rango completo, en cuyo caso la solución está dada por $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$.

PROPOSICIÓN A.36. *Una condición necesaria y suficiente para que la ecuación matricial* $\boldsymbol{AXB} = \boldsymbol{C}$ *tenga una solución es que*

$$\boldsymbol{AA}^{+}\boldsymbol{CB}^{+}\boldsymbol{B} = \boldsymbol{C},$$

en cuyo caso la solución general es

$$\boldsymbol{X} = \boldsymbol{A}^{+}\boldsymbol{CB}^{+} + \boldsymbol{Q} - \boldsymbol{A}^{+}\boldsymbol{AQBB}^{+},$$

donde $\boldsymbol{Q}$ *es una matriz arbitrária de órdenes apropiados.*

Apéndice B

Diferenciación matricial

En esta sección haremos uso de la siguiente notación. ϕ, $\boldsymbol{f}$ y $\boldsymbol{F}$ representan funciones escalar, vectorial y matricial, respectivamente mientras que ζ, $\boldsymbol{x}$ y $\boldsymbol{X}$ argumentos escalar, vectorial y matricial, respectivamente.

A partir de esta convención es directo que podemos escribir los siguientes casos particulares:

$$\begin{aligned}
&\phi(\zeta) = \zeta^2, \quad \phi(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x}, \quad \phi(\boldsymbol{X}) = \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{X}),\\
&\boldsymbol{f}(\zeta) = (\zeta, \zeta^2)^\top, \quad \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{a},\\
&\boldsymbol{F}(\zeta) = \zeta^2 \boldsymbol{I}_n, \quad \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^\top, \quad \boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{X}^\top.
\end{aligned}$$

Existen varias definiciones para la derivada de una función matricial $\boldsymbol{F}(\boldsymbol{X})$ con relación a su argumento (matricial) $\boldsymbol{X}$. En este apéndice nos enfocamos en el cálculo diferencial propuesto por Magnus y Neudecker (1985).

Considere $\phi : S \to \mathbb{R}$ con $S \subset \mathbb{R}^n$, se define la derivada de ϕ con relación a $\boldsymbol{x} \in S$ como

$$\frac{\partial \phi(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial \phi}{\partial x_1}, \ldots, \frac{\partial \phi}{\partial x_n}\right)^\top = \left(\frac{\partial \phi}{\partial x_i}\right) \in \mathbb{R}^n$$

de este modo, introducimos la notación

$$\mathsf{D}\phi(\boldsymbol{x}) = \frac{\partial \phi(\boldsymbol{x})}{\partial \boldsymbol{x}^\top} \in \mathbb{R}^{1\times n}.$$

Ahora, si $\boldsymbol{f} : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$. Entonces la matriz $m \times n$,

$$\mathsf{D}\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \mathsf{D}f_1(\boldsymbol{x}) \\ \vdots \\ \mathsf{D}f_m(\boldsymbol{x}) \end{pmatrix} = \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}^\top},$$

es la *derivada* o *matriz Jacobiana* de $\boldsymbol{f}$. La transpuesta de la matriz Jacobiana $\mathsf{D}\boldsymbol{f}(\boldsymbol{x})$ se denomina *gradiente* de $\boldsymbol{f}(\boldsymbol{x})$.

B.1. Aproximación de primer orden

Considere la fórmula de Taylor de primer orden,

$$\phi(c+u) = \phi(c) + u\phi'(c) + r_c(u),$$

donde el *resto*

$$\lim_{u\to 0} \frac{r_c(u)}{u} = 0.$$

es de orden más pequeño que u conforme $u \to 0$. Note también que

$$\lim_{u\to 0} \frac{\phi(c+u) - \phi(c)}{u} = \phi'(c).$$

De este modo, se define

$$\mathsf{d}\,\phi(c;u) = u\phi'(c),$$

como el *(primer) diferencial* de ϕ en c con incremento u. Esto motiva la siguiente definición.

DEFINICIÓN B.1 (Diferencial de una función vectorial). Sea $\boldsymbol{f} : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$, si existe una matriz $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, tal que

$$\boldsymbol{f}(\boldsymbol{c}+\boldsymbol{u}) = \boldsymbol{f}(\boldsymbol{c}) + \boldsymbol{A}(\boldsymbol{c})\boldsymbol{u} + \boldsymbol{r}_c(\boldsymbol{u}),$$

para todo $\boldsymbol{u} \in \mathbb{R}^n$ con $||\boldsymbol{u}|| < \delta$, y

$$\lim_{u\to 0} \frac{\boldsymbol{r}_c(\boldsymbol{u})}{||\boldsymbol{u}||} = \boldsymbol{0},$$

entonces la función $\boldsymbol{f}$ se dice diferenciable en $\boldsymbol{c}$. El vector $m \times 1$

$$\mathsf{d}\,\boldsymbol{f}(\boldsymbol{c};\boldsymbol{u}) = \boldsymbol{A}(\boldsymbol{c})\boldsymbol{u},$$

se denomina primer diferencial de $\boldsymbol{f}$ en $\boldsymbol{c}$ con incremento $\boldsymbol{u}$.

Magnus y Neudecker (1985) mostraron la existencia y unicidad del diferencial $\mathsf{d}\,\boldsymbol{f}(\boldsymbol{c};\boldsymbol{u})$ de una función $\boldsymbol{f} : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$ ($\boldsymbol{c} \in S$), dado por

$$\mathsf{d}\,\boldsymbol{f}(\boldsymbol{c};\boldsymbol{u}) = \boldsymbol{A}(\boldsymbol{c})\boldsymbol{u}$$

también mostraron la regla de la cadena e invarianza de Cauchy para el diferencial y enunciaron su primer teorema de identificación.

TEOREMA B.2 (Primer teorema de identificación). *Sea $\boldsymbol{f} : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$ función diferenciable, $\boldsymbol{c} \in S$ y $\boldsymbol{u}$ un vector n-dimensional. Entonces*

$$\mathsf{d}\,\boldsymbol{f}(\boldsymbol{c};\boldsymbol{u}) = (\mathsf{D}\boldsymbol{f}(\boldsymbol{c}))\boldsymbol{u}.$$

La matriz $\mathsf{D}\boldsymbol{f}(\boldsymbol{c}) \in \mathbb{R}^{m\times n}$ se denomina matriz Jacobiana. Tenemos también que

$$\nabla\boldsymbol{f}(\boldsymbol{c}) = (\mathsf{D}\boldsymbol{f}(\boldsymbol{c}))^\top$$

es la matriz gradiente de $\boldsymbol{f}$.

Sea $\boldsymbol{f} : S \to \mathbb{R}^m$, $S \subset \mathbb{R}^n$ y $f_i : S \to \mathbb{R}$ el i-ésimo componente de $\boldsymbol{f}$ ($i = 1, \ldots, m$). Sea $\boldsymbol{e}_j$ un vector n-dimensional cuyo j-ésimo elemento es uno y los restantes son cero, y considere

$$\lim_{t\to 0} \frac{f_i(\boldsymbol{c}+t\boldsymbol{e}_j) - f_i(\boldsymbol{c})}{t}$$

si el límite existe, se denomina la j-ésima *derivada parcial* de f_i en $\boldsymbol{c}$ y es denotada por $\mathsf{D}_j f_i(\boldsymbol{c})$. Note que el elemento ij de $\mathsf{D}\boldsymbol{f}(\boldsymbol{c})$ es $\mathsf{D}_j f_i(\boldsymbol{c})$.

B.2. Funciones matriciales

Considere algunos ejemplos de funciones matriciales

$$\boldsymbol{F}(\zeta) = \begin{pmatrix} \cos(\zeta) & \sin(\zeta) \\ -\sin(\zeta) & \cos(\zeta) \end{pmatrix}, \quad \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^\top, \quad \boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{X}^\top, \quad \boldsymbol{X} \in \mathbb{R}^{n\times q}.$$

Antes de considerar el diferencial de una función matricial $\boldsymbol{F} : S \to \mathbb{R}^{m\times p}$, $S \subset \mathbb{R}^{n\times q}$ introducimos dos conceptos preliminares: la vectorización de una matriz y el producto Kronecker.

DEFINICIÓN B.3 (Operador de vectorización). Sea $\boldsymbol{A} \in \mathbb{R}^{n\times q}$ particionada como

$$\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q),$$

donde $\boldsymbol{a}_k \in \mathbb{R}^n$ es la k-ésima columna de $\boldsymbol{A}$. Entonces

$$\operatorname{vec}(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_q \end{pmatrix}.$$

DEFINICIÓN B.4 (Producto Kronecker). Sea $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ y $\boldsymbol{B} \in \mathbb{R}^{p\times q}$, entonces el producto Kronecker entre $\boldsymbol{A}$ y $\boldsymbol{B}$ denotado por $\boldsymbol{A}\otimes\boldsymbol{B}$ es la matriz $mp\times nq$ definida como

$$\boldsymbol{A}\otimes\boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \ldots & a_{1n}\boldsymbol{B} \\ \vdots & & \vdots \\ a_{m1}\boldsymbol{B} & \ldots & a_{mn}\boldsymbol{B} \end{pmatrix}$$

RESULTADO B.5. *Sean $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ y $\boldsymbol{D}$ matrices de órdenes apropiados y λ escalar. Entonces*

(a) $\boldsymbol{A}\otimes\boldsymbol{B}\otimes\boldsymbol{C} = (\boldsymbol{A}\otimes\boldsymbol{B})\otimes\boldsymbol{C} = \boldsymbol{A}\otimes(\boldsymbol{B}\otimes\boldsymbol{C})$,
(b) $(\boldsymbol{A}+\boldsymbol{B})\otimes(\boldsymbol{C}+\boldsymbol{D}) = \boldsymbol{A}\otimes\boldsymbol{C} + \boldsymbol{B}\otimes\boldsymbol{C} + \boldsymbol{A}\otimes\boldsymbol{D} + \boldsymbol{B}\otimes\boldsymbol{D}$,
(c) $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = \boldsymbol{A}\boldsymbol{C}\otimes\boldsymbol{B}\boldsymbol{D}$,
(d) $\lambda\otimes\boldsymbol{A} = \lambda\boldsymbol{A} = \boldsymbol{A}\otimes\lambda$,
(e) $(\boldsymbol{A}\otimes\boldsymbol{B})^\top = \boldsymbol{A}^\top\otimes\boldsymbol{B}^\top$,
(f) $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$,
(g) $(\boldsymbol{A}\otimes\boldsymbol{B})^{-} = \boldsymbol{A}^{-}\otimes\boldsymbol{B}^{-}$.

RESULTADO B.6. *Sean $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ y $\boldsymbol{B} \in \mathbb{R}^{p\times p}$. Entonces*

(a) $\operatorname{tr}(\boldsymbol{A}\otimes\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})$,
(b) $|\boldsymbol{A}\otimes\boldsymbol{B}| = |\boldsymbol{A}|^p|\boldsymbol{B}|^n$,
(c) $\operatorname{rg}(\boldsymbol{A}\otimes\boldsymbol{B}) = \operatorname{rg}(\boldsymbol{A})\operatorname{rg}(\boldsymbol{B})$.

Observe que, si $\boldsymbol{a} \in \mathbb{R}^n$ y $\boldsymbol{b} \in \mathbb{R}^p$, entonces

$$\boldsymbol{a}\boldsymbol{b}^\top = \boldsymbol{a}\otimes\boldsymbol{b}^\top = \boldsymbol{b}^\top\otimes\boldsymbol{a},$$

por otro lado, tenemos que

$$\operatorname{vec}(\boldsymbol{a}\boldsymbol{b}^\top) = \operatorname{vec}(\boldsymbol{a}\otimes\boldsymbol{b}^\top) = \operatorname{vec}(\boldsymbol{b}^\top\otimes\boldsymbol{a}) = \boldsymbol{b}\otimes\boldsymbol{a}.$$

Estos resultados sugieren una conexión entre el operador de vectorización, el producto Kronecker y la traza. Considere el siguiente resultado

RESULTADO B.7.

(a) *Si $\boldsymbol{A}$ y $\boldsymbol{B}$ son ámbas matrices de orden $m\times n$, entonces*

$$\operatorname{tr}\boldsymbol{A}^\top\boldsymbol{B} = \operatorname{vec}^\top\boldsymbol{A}\operatorname{vec}\boldsymbol{B},$$

(b) *Si $\boldsymbol{A}, \boldsymbol{B}$ y $\boldsymbol{C}$ son de órdenes adecuados, entonces*

$$\operatorname{vec}\boldsymbol{A}\boldsymbol{B}\boldsymbol{C} = (\boldsymbol{C}^\top\otimes\boldsymbol{A})\operatorname{vec}\boldsymbol{B},$$

donde $\operatorname{vec}^\top\boldsymbol{A} = (\operatorname{vec}\boldsymbol{A})^\top$.

Finalmente, tenemos el siguiente resultado

RESULTADO B.8. *Sean $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ y $\boldsymbol{D}$ matrices, tal que, el producto $\boldsymbol{ABCD}$ está definido y es cuadrado, entonces*

$$\operatorname{tr} \boldsymbol{ABCD} = \operatorname{vec}^\top \boldsymbol{D}^\top (\boldsymbol{C}^\top \otimes \boldsymbol{A}) \operatorname{vec} \boldsymbol{B} = \operatorname{vec}^\top \boldsymbol{D} (\boldsymbol{A} \otimes \boldsymbol{C}^\top) \operatorname{vec} \boldsymbol{B}^\top.$$

Sea $\boldsymbol{F} : S \to \mathbb{R}^{m \times p}$, $S \subset \mathbb{R}^{n \times q}$ una función matricial, podemos notar que

$$\operatorname{vec} \boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{f}(\operatorname{vec} \boldsymbol{X})$$

esto permite obtener el diferencial de una función matricial considerando la relación

$$\operatorname{vec} \mathsf{d}\, \boldsymbol{F}(\boldsymbol{C}; \boldsymbol{U}) = \mathsf{d}\, \boldsymbol{f}(\operatorname{vec} \boldsymbol{C}; \operatorname{vec} \boldsymbol{U})$$

en cuyo caso $\boldsymbol{F}$ tiene matriz Jacobiana

$$\mathsf{D}\boldsymbol{F}(\boldsymbol{C}) = \mathsf{D}\boldsymbol{f}(\operatorname{vec} \boldsymbol{C})$$

Las consideraciones anteriores motivan el primer teorema de indentificación para funciones matriciales (Magnus y Neudecker, 1985)

TEOREMA B.9 (Primer teorema de identificación para funciones matriciales). *Sea $\boldsymbol{F} : S \to \mathbb{R}^{m \times p}$, $S \subset \mathbb{R}^{n \times q}$ función diferenciable, $\boldsymbol{C} \in S$ y $\boldsymbol{U}$ matriz $n \times q$. Entonces*

$$\operatorname{vec} \mathsf{d}\, \boldsymbol{F}(\boldsymbol{C}; \boldsymbol{U}) = (\mathsf{D}\boldsymbol{F}(\boldsymbol{C})) \operatorname{vec} \boldsymbol{U}.$$

con $(\mathsf{D}\boldsymbol{F}(\boldsymbol{C}))^\top$ la matriz gradiente de $\boldsymbol{F}$.

B.3. Matriz Hessiana

Considere $\phi : S \to \mathbb{R}$ con $S \subset \mathbb{R}^n$, entonces se define la *matriz Hessiana* como la matriz de segundas derivadas, dada por

$$\mathsf{H}\phi(\boldsymbol{x}) = \frac{\partial^2 \phi(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\top} = \frac{\partial}{\partial \boldsymbol{x}^\top} \left(\frac{\partial \phi(\boldsymbol{x})}{\partial \boldsymbol{x}^\top} \right)^\top = \mathsf{D}(\mathsf{D}\phi(\boldsymbol{x}))^\top.$$

Es posible definir el diferencial de funciones vectoriales y matriciales de manera análoga a la delineada anteriormente. Sin embargo, en este apéndice nos enfocaremos solamente en el cálculo de diferenciales de funciones escalares. El segundo diferencial de una función escalar está dado por

$$\mathsf{d}^2\, \phi = \mathsf{d}(\mathsf{d}\, \phi).$$

Magnus y Neudecker (1985) enunciaron el siguiente teorema de identificación para matrices Hessianas de funciones escalares

TEOREMA B.10 (Segundo teorema de identificación). *Sea $\phi : S \to \mathbb{R}$, $S \subset \mathbb{R}^n$ dos veces diferenciable, $\boldsymbol{c} \in S$ y $\boldsymbol{u}$ vector n-dimensional. Entonces*

$$\mathsf{d}^2\, \phi(\boldsymbol{c}; \boldsymbol{u}) = \boldsymbol{u}^\top (\mathsf{H}\phi(\boldsymbol{c}))\boldsymbol{u}.$$

donde $\mathsf{H}\phi(\boldsymbol{c}) \in \mathbb{R}^{n \times n}$ es la matriz Hessiana de ϕ.

Algunas ventajas (prácticas) importantes del cálculo de diferenciales son:

- Sea $\boldsymbol{f}(\boldsymbol{x})$ función vectorial $m \times 1$ con argumento $\boldsymbol{x}$, vector n-dimensional, entonces

$$\mathsf{D}\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^{m \times n} \qquad \text{sin embargo,} \qquad \mathsf{d}\, \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$$

- Para funciones matriciales, $\mathsf{d}\, \boldsymbol{F}(\boldsymbol{X})$ tiene la *misma* dimensión que $\boldsymbol{F}$ *sin importar* la dimensión de $\boldsymbol{X}$.

B.4. Reglas fundamentales

A continuación se presentan algunas reglas fundamentales para el cálculo de diferenciales
Considere u y v funciones escalares y α una constante, entonces:

$$\begin{aligned}
&\mathsf{d}\,\alpha = 0, \qquad \mathsf{d}(\alpha u) = \alpha\,\mathsf{d}\,u, \quad \mathsf{d}(u+v) = \mathsf{d}\,u + \mathsf{d}\,v,\\
&\mathsf{d}(uv) = (\mathsf{d}\,u)v + u(\mathsf{d}\,v) \qquad \mathsf{d}(u/v) = \frac{(\mathsf{d}\,u)v - u(\mathsf{d}\,v)}{v^2}, (v \neq 0),\\
&\mathsf{d}\,u^\alpha = \alpha u^{\alpha-1}\,\mathsf{d}\,u, \qquad\qquad \mathsf{d}\,e^u = e^u\,\mathsf{d}\,u,\\
&\mathsf{d}\log u = u^{-1}\,\mathsf{d}\,u, (u>0) \qquad \mathsf{d}\,\alpha^u = \alpha^u \log\alpha\,\mathsf{d}\,u, (\alpha>0),
\end{aligned}$$

aquí por ejemplo,

$$\phi(x) = u(x) + v(x).$$

Análogamente para $\boldsymbol{U}$, $\boldsymbol{V}$ funciones matriciales, α un escalar (constante) y $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ constante, tenemos

$$\begin{aligned}
\mathsf{d}\,\boldsymbol{A} = \boldsymbol{0}, \qquad \mathsf{d}(\alpha\boldsymbol{U}) = \alpha\,\mathsf{d}\,\boldsymbol{U},\\
\mathsf{d}(\boldsymbol{U}+\boldsymbol{V}) = \mathsf{d}\,\boldsymbol{U} + \mathsf{d}\,\boldsymbol{V}, \quad \mathsf{d}(\boldsymbol{U}\boldsymbol{V}) = (\mathsf{d}\,\boldsymbol{U})\boldsymbol{V} + \boldsymbol{U}\,\mathsf{d}\,\boldsymbol{V},\\
\mathsf{d}(\boldsymbol{U}\otimes\boldsymbol{V}) = \mathsf{d}\,\boldsymbol{U}\otimes\mathsf{d}\,\boldsymbol{V}, \qquad \mathsf{d}(\boldsymbol{U}\odot\boldsymbol{V}) = \mathsf{d}\,\boldsymbol{U}\odot\mathsf{d}\,\boldsymbol{V},\\
\mathsf{d}\,\boldsymbol{U}^\top = (\mathsf{d}\,\boldsymbol{U})^\top, \quad \mathsf{d}\,\mathrm{vec}\,\boldsymbol{U} = \mathrm{vec}\,\mathsf{d}\,\boldsymbol{U}, \quad \mathsf{d}\,\mathrm{tr}\,\boldsymbol{U} = \mathrm{tr}\,\mathsf{d}\,\boldsymbol{U}.
\end{aligned}$$

Otros diferenciales de uso frecuente en Estadística son:

$$\begin{aligned}
\mathsf{d}\,|\boldsymbol{F}| = |\boldsymbol{F}|\,\mathrm{tr}\,\boldsymbol{F}^{-1}\,\mathsf{d}\,\boldsymbol{F}, \qquad \mathsf{d}\log|\boldsymbol{F}| = \mathrm{tr}\,\boldsymbol{F}^{-1}\,\mathsf{d}\,\boldsymbol{F},\\
\mathsf{d}\,\boldsymbol{F}^{-1} = -\boldsymbol{F}^{-1}(\mathsf{d}\,\boldsymbol{F})\boldsymbol{F}^{-1}.
\end{aligned}$$

Bibliografía

Andrews, D.F., Mallows, C.L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B* **36**, 99-102.

Arellano, R. (1994). *Distribuições Elípticas: Propriedades, Inferência e Aplicações a Modelos de Regressão*. (Unpublished doctoral dissertation). Department of Statistics, University of São Paulo, Brazil.

Atkinson, A.C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13-20.

Atkinson, A.C. (1985). *Plots, Transformations and Regressions*. Oxford University Press, Oxford.

Barlow, J.L. (1993). Numerical aspects of solving linear least squares problems. En *Handbook of Statistics, Vol. 9*, C.R. Rao (Ed.). Elsevier, Amsterdam, pp. 303-376.

Barrodale, I., Roberts, F.D.K. (1973). An improved algorithm for discrete L1 linear approximations. *SIAM Journal of Numerical Analysis* **10**, 839-848.

Barrodale, I., Roberts, F.D.K. (1974). Solution of an overdetermined system of equations in the L1 norm. *Communications of the ACM* **17**, 319-320.

Belsley, D.A., Kuh, E., Welsh, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Björck, A. (1996). *Numerical Methods for Least Squares Problems*. Society for Insdustrial and Applied Mathematics, Philadelphia.

Box, G.E.P., Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211-252.

Carroll, R.J., Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

Charnes, A., Cooper, W.W., Ferguson, R.O. (1955). Optimal estimation of executive compensation by linear programming. *Management Science* **1**, 138-151.

Chatterjee, S., Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*. Wiley, New York.

Christensen, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models*, 4th Ed. Springer, New York.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics* **19**, 15-18.

Cook, R.D., Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, 495-508.

Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, New York.

Daniel, C., Wood, F.S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd Ed. Wiley, New York.

Davidian, M. Carroll, R.J. (1987). Variance function estimation. *Journal of the American Statistical Association* **82**, 1079-1091.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Díaz-García, J.A., Gutiérrez-Jáimez, R. (1999). *Cálculo Diferencial Matricial y Momentos de Matrices Aleatorias Elípticas*. Universidad de Granada.

Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*, 2nd Ed. Chapman & Hall, Boca Raton.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer, Berlin.

Fang, K.T., Kotz, S., Ng, K.W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London.

Galea, M. (1990). Técnicas de diagnóstico en regresión lineal. *Revista de la Sociedad Chilena de Estadística* **7**, 23-44.

Gentle, J.E. (2007). *Matrix Algebra: Theory, Computation and Applications in Statistics*. Springer, New York.

Gómez, E., Gómez-Villegas, M.A., Marín, J.M. (1988). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods* **27**, 589-600.

Golub, G.H., Heath, M., Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-223.

Goodnight, J.H. (1979). A tutorial on the SWEEP operator. *The American Statistician* **33**, 149-158.

Gorman, J.W., Toman, R.J. (1966). Selection of variables for fitting equations to data. *Technometrics* **8**, 27-51.

Graybill, F.A. (1961). *An Introduction to Linear Statistical Models*. McGraw-Hill, New York.

Graybill, F.A. (1976). *Theory and Application of the Linear Model*. Wadsworth & Brooks, Pacific Grove, CA.

Graybill, F.A. (1983). *Matrices with Applications in Statistics*, 2nd Ed. Wadsworth, Belmont, CA.

Groß, J. (2003). *Linear Regression*. Springer, Berlin.

Gruber, M.H.J. (1998). *Improving Efficiency by Shrinkage*. Marcel Dekker, New York.

Hald, A. (1952). *Statistical Theory with Engineering Applications*. Wiley, New York.

Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer, New York.

Hoaglin, D.C., Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician* **32**, 17-22.

Hocking, R. (1996). *Methods and Applications of Linear Models*. Wiley, New York.

Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**, 69-82.

Hoerl, A.E., Kennard, R.W., Baldwin, K.F. (1975). Ridge regression: Some simulations. *Communications in Statistics: Theory and Methods* **4**, 105-123.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Kariya, T., Kurata, H. (2004). *Generalized Least Squares*. Wiley, Chichester.

Lange, K. (1999). *Numerical Analysis for Statisticians*. Springer, New York.

Lange, K.L., Little, R.J.A., Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881-896.

Lange, K., Sinsheimer, J.S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* **2**, 175-198.

Lawless, J.F., Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics: Theory and Methods* **5**, 307-323.

Lindley, D.V., Smith, A.F.M. (1972). Bayes estimates for the linear model *Journal of the Royal Statistical Society, Series B* **34**, 1-41.

Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics* **37**, 23-38.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226-233.

Magnus, J.R., Neudecker, H. (1985). Matrix differential calculus with applications to simple, Hadamard and Kronecker products. *Journal of Mathematical Psychology* **29**, 474-492.

Magnus, J.R., Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd Ed. Wiley, New York.

Magnus, J.R. (2010). On the concept of matrix derivative. *Journal of Multivariate Analysis* **101**, 2200-2206.

McIntosh, A. (1982). *Fitting Linear Models: An Application of Conjugate Gradients Algorithms*. Springer, New York.

McLachlan. G.J., Krishnan, T. (2008). *The EM Algorithm and Extensions*, 2nd Ed. Wiley, New York.

O'Leary, D.P. (1990). Robust regression computation using iteratively reweighted least squares. *SIAM Journal on Matrix Analysis and Applications* **11**, 466-480.

Osborne, M.R. (1985). *Finite Algorithms in Optimization and Data Analysis*. Wiley, New York.

Osorio, F., Ogueda, A. (2021). fastmatrix: *Fast computation of some matrices useful in statistics*. R package version 0.3-819. URL: faosorios.github.io/fastmatrix

Paula, G.A. (2013). *Modelos de Regressão, com Apoio Computacional*. Instituto de Matemática e Estatística, Univeridade de São Paulo, Brasil.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: www.R-project.org

Rao, C.R., Toutenburg, H., Shalabh, Heumann, C. (2008). *Linear Models and Generalizations: Least Squares and Alternatives*. Springer, New York.

Ravishanker, N., Dey, D.K. (2002). *A First Course in Linear Model Theory*. Chapman & Hall, London.

Ruppert, D., Wand, M.P., Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Schlossmacher, E.J. (1973). An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association* **68**, 857-859.

Searle, S.R. (1971). *Linear Models*. Wiley, New York.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.

Seber, G.A.F., Lee, A.J. (2003). *Linear Regression Analysis*, 2nd Ed. Wiley, New York.

Sen, A., Srivastava, M. (1990). *Regression Analysis: Theory, Methods and Applications*. Springer, New York.

Tong, Y.L. (1990). *The Multivariate Normal Distribution*. Springer, New York.

Velleman, P.F., Welsch, R.E. (1981). Efficient computing of regression diagnostics. *The American Statistician* **35**, 234-242.

Venables, W.N., Ripley, B.D. (2002). *Modern Applied Statistics with S*, 4th Ed. Springer, New York.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd Ed. Wiley, New York

Welsch, R.E., Kuh, E. (1977). Linear regression diagnostics. *Massachusetts Institute of Technology and National Bureau of Economics, Cambridge*. Working paper No. 173.

Woods, H., Steinour, H.H., Starke, H.R. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Industrial and Engineering Chemistry* **24**, 1207-1214.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**, 95-103.