UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACÍON EN ESTADÍSTICA APLICADA

"ANÁLISIS DE REGRESIÓN CON DISTRIBUCIONES DIFERENTES A LA NORMAL"

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRA EN CIENCIAS MATEMÁTICAS

PRESENTA:
BRENDA CORINA CEREZO SILVA

DIRECTORA:
DRA. SILVIA RUÍZ VELASCO ACOSTA
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS (IIMAS)

CIUDAD DE MÉXICO, JUNIO 2022.

Dedico esta tesis a todos aquellos que, como yo, aman las matemáticas pero también en más de una ocasión las han sufrido y, llenos de frustración, se han sentido incapaces de comprender un concepto, un teorema, una demostración... Esta tesis es la prueba de que con perseverancia, paciencia y ayuda siempre se puede.

Agradecimientos

La lista de aquellos que directa o indirectamente contribuyeron a la realización de esta tesis es afortunadamente larga, lo que me permite darme cuenta de que vivo rodeada de personas que me aman y apoyan y de instituciones que dan oportunidades.

Agradezco a CONACYT por el apoyo económico, lo que me permitió dedicarme de tiempo completo a la maestría en Ciencias Matemáticas que tanto anhelé estudiar y hoy es un sueño alcanzado.

Agradezco a la UNAM que, con sus profesores y alumnos, me ha ayudado a crecer y ha guiado mi camino hacia las matemáticas. Se quedan en mi corazón mis amigos y profesores de maestría, y en mi mente las enseñanzas matemáticas que considero invaluables.

Agradezco a mi asesora la Dra. Silvia Ruíz Velazco Acosta por ser mi mentora. Su compromiso e instrucción se ven reflejados en este trabajo que me permite obtener el grado de maestra.

Agradezco a mis padres y hermanos, por su consejo y apoyo incondicional con todas mis decisiones.

Y agradezco a Gerardo, porque en más de una ocasión me hizo ver fácil aquello que no lo parecía, y por su tiempo, apoyo y motivación constantes.

Con orgullo afirmo que *soy la suma de las cosas que he aprendido, las decisiones que he tomado y las personas que he encontrado en mi camino.*

"¡MÉXICO, PUMAS, UNIVERSIDAD!
¡GOYA! ¡GOYA!
¡CACHUN, CACHUN, RA, RA!
¡CACHUN, CACHUN, RA, RA!
¡GOYA!
¡¡UNIVERSIDAD!!"

Resumen

En este trabajo de investigación se presenta la teoría alrededor del modelo lineal $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$ donde $\beta_0,\ \beta_1, \ldots, \beta_k$ son parámetros desconocidos, $x_{i1},\ x_{i2}, \ldots, x_{ik}$ son k valores conocidos comúnmente llamados variables explicativas y ε_i es una variable aleatoria con media 0 y varianza constante, el subíndice $i \in \{1, 2, \ldots, n\}$ se refiere a la observación de cada individuo. Este modelo es importante pues permite entender la relación entre la variable Y_i y las variables explicativas. Ahora bien, el principal interés es estimar a través de una muestra los parámetros desconocidos y la varianza de los errores, y después, mediante distintas pruebas, hacer inferencia estadística acerca del modelo y de los resultados obtenidos.

En el presente trabajo se muestran tres formas diferentes en las que se puede encontrar el estimador para los parámetros β_i, la primera es a través de una técnica llamada mínimos cuadrados que no ocupa saber la distribución de los errores del modelo, luego se obtienen dos por método de máxima verosimilitud, uno definiendo la distribución de los errores como una normal y otro definiéndola como una perteneciente a la familia de distribuciones elípticas. Para estos últimos dos casos, además del estimador de los coeficientes del modelo, se obtiene también el estimador de la varianza de los errores y se proporcionan las estadísticas para las pruebas de hipótesis.

La información está organizada de la siguiente forma: en el Capítulo 1, se define detalladamente el modelo, su notación y los supuestos sobre los cuales se basan las estimaciones y se define la distribución normal multivariada junto con sus propiedades. En el Capítulo 2, se muestran los estimadores por mínimos cuadrados. En el Capítulo 3, se define la distribución de los errores como una normal, y se obtiene los estimadores máximo verosímiles para los parámetros desconocidos y para la varianza de los errores, así como las estadísticas de pruebas de hipótesis. En el Capítulo 4, se define la familia de distribuciones elípticas y algunas de sus propiedades. En el Capítulo 5, se define la distribución de los errores como una perteneciente a la familia elíptica y se obtiene los estimadores máximo verosímiles para los parámetros desconocidos y para la varianza de los errores, así como las estadísticas de pruebas de hipótesis. En el Capítulo 6 se realizan varias simulaciones que permiten comparar los resultados. Y en el Capítulo 7 se ofrecen algunas conclusiones. Adicional, al final en el Apéndice, a manera de resumen, se demuestran algunas propiedades útiles del álgebra matricial y se muestra una tabla de la familia de distribuciones elípticas.

Abstract

In this thesis, we show the theory around the linear model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$ where $\beta_0,\ \beta_1, \ldots, \beta_k$ are unknown parameters, $x_{i1},\ x_{i2}, \ldots, x_{ik}$ are k known values usually called explanatory variables and ε_i is a random variable with mean 0 and constant variance, the i subscript refers to the data of an individual obsterved. The main interest in this model is to estimate the unknown parameters and the variance of the errors, and then through different tests make statistical inference about the model and the results obtained.

In the present thesis three different ways to estimate the parameters β_i are shown, the first one is through a technique called least squares that does not require knowing the distribution of model errors, the next two are obtained by the maximum likelihood method, the first one with normal errors and the other one with elliptically contoured errors. The estimator of the variance of the errors is also obtained and the statistics for the hypothesis tests are provided.

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1. El problema de regresión

El análisis de regresión es una técnica estadística que se utiliza para investigar y modelar la relación entre variables (cuantitativas o cualitativas), esto es, describir en cierto sentido el comportamiento de una variable de interés, conocida como variable respuesta o variable dependiente, en función de otras, llamadas covariables, variables explicativas o variables independientes. En particular, cuando la relación es lineal se llama Regresión Lineal.

Las aplicaciones de esta técnica estadística se dan en casi todas las áreas, entre ellas, economía, administración, biología, física, química, etc. Un problema que puede ser analizado a través de una regresión lineal es, por ejemplo, el monto que una institución financiera presta a cada uno de sus clientes, esta cantidad sería la variable respuesta (Y), y puede modelarse o explicarse a través de otros datos del cliente como su edad, el sueldo promedio de los últimos 3 años, su género, la cantidad de personas que dependen económicamente de él/ella, entre otras, cada una de éstas sería una variable explicativa ($z_1, z_2, z_3, ...$). Otro ejemplo es en el área de la agricultura, supongamos que se desea conocer el total de cosecha útil, en kilogramos o en piezas, la cuál sería la variable respuesta (Y), en función de la cantidad de fertilizante, la frecuencia de riego, la distancia entre plantas, entre otras. Cada una de éstas sería una variable explicativa ($z_1, z_2, z_3, ...$). Resulta evidente, entonces, la utilidad del análisis de regresión lineal.

El análisis de regresión no solo es interesante por su practicidad sino por el fundamento matemático, es decir, la elegancia de las matemáticas en la teoría estadística que surge en su desarrollo. El uso exitoso de la regresión requiere una apreciación tanto de la teoría como de los problemas prácticos que surgen cuando la técnica se emplea con datos del mundo real (Montgomery, 2012)[11] .

Ahora bien, nos enfrentamos al problema de tener k variables explicativas $z_1, z_2, \ldots, z_k$

con $k \in \mathbb{N}$ y una variable respuesta Y. Entonces se plantea ajustar un modelo lineal de la forma:

$$Y_i = \beta_0 + \beta_1 s_1(\underline{z}_i) + \beta_2 s_2(\underline{z}_i) + \cdots + \beta_k s_k(\underline{z}_i) + \varepsilon_i; \;\; \forall\, i \in \{1, 2, \ldots, n\},$$

donde

- el tamaño de muestra será denotado por n,
- el número de covariables o variables explicativas se denota por k,
- Y_i es la variable respuesta del individuo i,
- β_j son los parámetros desconocidos en el modelo, hay $p = k + 1$,
- $\underline{z}_i$ es el vector de covariables para el individuo i, esto es $\underline{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ik})^T$,
- s_j son funciones de las covariables con $j \in \{1, 2, \ldots, k\}$, y
- ε_i es una variable aleatoria que representa el error en la relación. Usualmente con $\mathbb{E}(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$.

Se definen

$$x_{ij} = s_j(\underline{z}_i); \;\; \forall i \in \{1, 2, \ldots, n\}, \; j \in \{1, 2, \ldots, k\}.$$

Nótese que las k funciones de las covariables $\underline{z}$ dan origen a k variables que se denotarán por x, de ahí que se dice que se ajusta un modelo con k variables. Entonces, el modelo se transforma al modelo aditivo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i; \;\; \forall\, i \in \{1, 2, \ldots, n\}. \tag{1.1}$$

Este cambio permite entender que las variables explicativas pueden ser función de otras variables.

La ecuación (1.1) es la que se conoce como modelo de regresión lineal múltiple.

1.1.1. Notación matricial del modelo de regresión lineal múltiple

Los datos se tienen presentes de la siguiente forma

La variable respuesta para la i-ésima observación es Y_i (en mayúscula dado que es variable aleatoria[1]), sin embargo, una vez que el valor ha sido observado se denota en minúscula, es decir, y_i.

[1]Los supuestos del modelo se detallan en la sección 1.1.2.

		Variables explicativas			
Obs.	Variable respuesta	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Tabla 1.1: Datos de una regresión lineal múltiple.

Con base en los datos de la Tabla 1.1 y en la ec. (1.1), las ecuaciones de regresión lineal múltiple para cada observación son:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_1 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_1 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_1 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n. \end{aligned}$$

Se definen

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, X = \begin{bmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1k} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{x}_{n1} & \mathrm{x}_{n2} & \cdots & \mathrm{x}_{nk} \end{bmatrix}_{n\times p}, \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{p\times 1}, \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1},$$

donde:

- $p = k + 1$,
- $\underline{y}$: vector de $n \times 1$ de valores observados de la variable respuesta Y,
- X: matriz de $n \times p$ cuya primera columna es un vector de unos y la j-ésima columna corresponde a los valores observados de la $(j-1)$-ésima variable explicativa,
- $\underline{\beta}$: vector de $p \times 1$ de los parámetros a ser estimados,
- $\underline{\varepsilon}$: vector de $n \times 1$ de los errores aleatorios.

Entonces, el modelo con los datos observados es

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}. \tag{1.2}$$

Por lo tanto, el modelo de regresión lineal múltiple (1.1) expresado en notación matricial es

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}. \tag{1.3}$$

1.1.2. Supuestos

Para el modelo de regresión lineal múltiple dado por la ecuación (1.1) se asumen los siguientes supuestos:

(I) Los valores que toma la variable Y_i están definidos por la ec. (1.1), es decir, existe una relación lineal (en términos de los coeficientes). En otras palabras, el valor de Y_i oscila alrededor de $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ con una *fluctuación* o *diferencia* de ε_i.

Es importante mencionar que las variables $x_{i1}, x_{i2}, \ldots, x_{ik}$ son constantes conocidas y que los coeficientes $\beta_0, \beta_1, \ldots, \beta_k$ son parámetros desconocidos.

(II) $\varepsilon_i \sim f(0, \sigma^2)\ \forall\ i \in \{1, \ldots, n\}$, es decir, los errores son variables aleatorias idénticamente distribuidas con alguna función de densidad, digamos f, $\mathbb{E}(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2 < \infty$. A esto último, que la varianza sea la misma para todos los errores, se le conoce como supuesto de homocedasticidad.

(III) $Cov(\varepsilon_i, \varepsilon_j) = 0\ \forall\ i \neq j$. Esto implica que los errores sean no correlacionados.

(IV) A consecuencia del supuesto (II), dado que $Y_i = C_i + \varepsilon_i$ con $C_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ y, por propiedades de la esperanza y de la varianza, se tiene que $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ y $Var(Y_i) = \sigma^2$.

(V) Todas las variables explicativas son linealmente independientes, es decir, ninguna puede ser expresada como una combinación lineal de las otras. Además, se tienen más observaciones que variables explicativas, esto es $p < n$.

Dos consecuencias del supuesto (V) son las que se enuncian en el teorema A.0.4. En otras palabras, considerando la notación matricial (obsérvese sección 1.1.1), el Teorema A.0.4 dice que la inversa de $X^T X$ existe siempre que las variables explicativas sean linealmente independientes, es decir, que no haya ninguna columna de la matriz X que sea combinación lineal de otras.

1.1.3. Observaciones sobre los parámetros a estimar

Conviene comentar la diferencia entre Y_i y $\mathbb{E}(Y_i)$.

En el supuesto (I) se establece que la variable Y_i tiene la relación dada por la ecuación (1.1) con las variables explicativas $x_{i1}, x_{i2}, \ldots, x_{ik}$. Pero esta relación no es determinista pues, aunque las variables explicativas son constantes conocidas, se agrega un valor aleatorio ε_i.

Por otro lado, el supuesto (IV) establece que

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \tag{1.4}$$

donde el lado derecho de la igualdad no contiene variables aleatorias.

Ahora bien, no es lo mismo Y_i que $\mathbb{E}(Y_i)$. Esto es evidente, en primera instancia, por el término ε_i, sin embargo, la diferencia puede comentarse más allá de solo la expresión matemática.

Para ilustrarlo mejor, supóngase que Y_i representa el ingreso mensual de la persona i económicamente activa y supóngase que este valor está linealmente relacionado con el total de años que ha estudiado y con su edad, digamos x_{i1} y x_{i2} respectivamente, es decir, que se cumple el supuesto (I). Entonces el salario no necesariamente es el mismo para todos los individuos con la misma escolaridad y con la misma edad, es decir, el salario no es $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ para todos, pero el salario **esperado** sí.

Siguiendo con el ejemplo, se toman dos empleados de la compañía Datos S.A. de C.V., los empleados 1 y 2, y ambos tienen maestría, esto es x_1 años de estudio, entonces $x_{11} = x_{21} = x_1$, y ambos tienen x_2 años, esto es $x_{12} = x_{22} = x_2$, ¿diríamos que ambos ganan lo mismo? No necesariamente. Sean y_1 y y_2 los salarios reales, es decir, los valores observados, entonces **se espera** que tengan el mismo ingreso:

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Con esto se entiende que no es lo mismo el valor observado que el valor esperado y la diferencia entre ellos es:

$$\varepsilon_1 = y_1 - \mathbb{E}(Y_1)$$
$$\varepsilon_2 = y_2 - \mathbb{E}(Y_2) = y_2 - \mathbb{E}(Y_1),$$

de forma general

$$\varepsilon_i = y_i - \mathbb{E}(Y_i) = y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}). \tag{1.5}$$

1.2. Distribución normal multivariada

Una gran cantidad de técnicas usadas en la estadística aplicada se basan en la distribución de probabilidad normal, sobretodo en el caso multivariado o multivariante, que constituye una forma general de la distribución normal univariada. El estudio del caso de más de dos variables conviene en notación matricial, de lo contrario los cálculos se vuelven demasiado complicados.

En esta sección se presenta esta famosa distribución y sus principales propiedades útiles para el análisis de regresión[13].

1.2.1. Función de densidad

Antes de definir la función de densidad conviene enunciar el siguiente teorema:

Teorema 1.2.1 *Integral de Gauss.*

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}.$$

Demostración: Se define I como

$$I = \int_{-\infty}^{\infty} e^{-x^2}\, dx,$$

entonces, se define I^2 como el producto de I por sí misma y por teorema de Foubini se puede expresar como

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2}\, dx\right)^2 \\ &= \left(\int_{-\infty}^{\infty} e^{-x^2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2}\, dy\right) \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{-x^2}\, dx\right) e^{-y^2}\, dy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dx\right)\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dx\, dy \\ &= \int_{\mathbb{R}} e^{-(x^2+y^2)}\, dx\, dy. \end{aligned}$$

Realizando un cambio de variable a coordenadas polares donde:

$$\begin{aligned} x &= r cos\ \theta \\ y &= r sen\ \theta, \end{aligned}$$

el jacobiano es

$$J(r,\theta) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial x}{\partial \theta} \end{pmatrix} = \begin{pmatrix} cos\ \theta & -r sen\ \theta \\ sen\ \theta & r cos\ \theta \end{pmatrix},$$

y la diferencial

$$d(x,y) = |J(r,\theta)| d(r,\theta) = (r cos^2\ \theta + r sen^2\ \theta) d(r,\theta) = r d(r,\theta).$$

Entonces,

$$I^2 = \int_{\mathbb{R}} e^{-(x^2+y^2)}\, dx\, dy$$

$$
\begin{aligned}
&= \int_0^{2\pi} \int_0^{\infty} re^{-r^2} dr\, d\theta \\
&= \int_0^{2\pi} d\theta \int_0^{\infty} re^{-r^2} dr \\
&= \theta\Big|_0^{2\pi} \cdot \; -\frac{1}{2} e^{-r^2}\Big|_0^{\infty} \\
&= (2\pi - 0)\left(-\frac{1}{2}\right)\left(\lim_{r\to\infty} e^{-r^2} - e^{-0}\right) \\
&= \pi.
\end{aligned}
$$

Por lo tanto,

$$I = (I^2)^{1/2} = \int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}.$$

■

Sean Σ una matriz simétrica d.p. de $n \times n$ (Definición A.0.1), $\underline{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, con $\underline{y} \neq \underline{0}$, $\underline{\mu} \in \mathbb{R}^n$ y $k > 0$. Considérese la función

$$f(\underline{y}) = \left[(2\pi)^{n/2} det(\Sigma)^{1/2}\right]^{-1} exp\Big\{ -\frac{1}{2}(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\Big\}, \tag{1.6}$$

donde $det(\Sigma)$ denota el determinante de la matriz Σ.

Obsérvese que, dado que Σ^{-1} es también definida positiva (Teorema A.0.1), sucede que $(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu}) > 0$, así que f está acotada y toma su máximo en $\left[(2\pi)^{n/2} det(\Sigma)^{1/2}\right]^{-1}$ cuando $\underline{y} = \underline{\mu}$.

Teorema 1.2.2 *La función (1.6) es una función de densidad de probabilidad.*

Demostración:

Claramente, $f(\underline{y})$ no es negativa. Además, por el Teorema A.0.2 existe $\Sigma^{1/2}$ invertible y, por lo tanto, puede definirse

$$\underline{z} = (\Sigma^{1/2})^{-1}(\underline{y} - \underline{\mu}), \tag{1.7}$$

de tal forma que $\underline{y} = \Sigma^{1/2}\underline{z} + \underline{\mu}$. El jacobiano de esta transformación es:

$$|J| = det\left(\frac{\partial y_i}{\partial z_j}\right) = det(\Sigma^{1/2}) = \left[det(\Sigma)\right]^{1/2},$$

esta última igualdad por propiedades del determinante. Y defínase también $k = (2\pi)^{n/2} det(\Sigma)^{1/2}$.

Integrando $f(\underline{y})$ y haciendo el cambio de variable se tiene

$$\begin{aligned}
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\underline{y})\, dy_1\cdots dy_n &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} k^{-1}exp\Big\{-\frac{1}{2}(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\Big\}\, dy_1\cdots dy_n \\
&= k^{-1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} exp\Big\{-\frac{1}{2}\underline{z}^T\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}\underline{z}\Big\}\,|J|dz_1\cdots dz_n \\
&= k^{-1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} exp\Big\{-\frac{1}{2}\underline{z}^T\Sigma^{1/2}(\Sigma^{1/2}\Sigma^{1/2})^{-1}\Sigma^{1/2}\underline{z}\Big\}\,|J|dz_1\cdots dz_n \\
&= k^{-1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} exp\Big\{-\frac{1}{2}\underline{z}^T\Sigma^{1/2}(\Sigma^{1/2})^{-1}(\Sigma^{1/2})^{-1}\Sigma^{1/2}\underline{z}\Big\}\,|J|dz_1\cdots dz_n \\
&= k^{-1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} exp\Big\{-\frac{1}{2}\underline{z}^T\underline{z}\Big\}\,|J|dz_1\cdots dz_n \\
&= k^{-1}|J|\prod_{i=1}^{n}\int_{-\infty}^{\infty} exp\Big\{-\frac{1}{2}z_i^2\Big\}\, dz_i \\
&= k^{-1}det(\Sigma)^{1/2}\prod_{i=1}^{n}\sqrt{2\pi} \quad \textit{por Teorema } 1.2.1 \\
&= k^{-1}k \\
&= 1.
\end{aligned}$$

■

Definición 1.2.1 *La distribución correspondiente a la función de densidad (1.6) se llama distribución normal multivariada o multivariante.*

Teorema 1.2.3 *Si un vector aleatorio $\underline{Y}$ tiene la distribución 1.2.1, entonces $\mathbb{E}[\underline{Y}] = \underline{\mu}$ y $Var(\underline{Y}) = \Sigma$.*

Demostración:

Sea $\underline{Z} = (\Sigma^{1/2})^{-1}(\underline{Y}-\underline{\mu})$, aplicando el cambio de variable, resulta que $\underline{Z}$ tiene densidad:

$$\begin{aligned}
g(z_1,\ldots,_n) &= f(\underline{y}(\underline{z}))|J| \\
&= (2\pi)^{-n/2}det(\Sigma)^{-1/2}exp\Big\{-\frac{1}{2}\underline{z}^T\underline{z}\Big\}det(\Sigma)^{1/2} \\
&= (2\pi)^{-n/2}\prod_{i=1}^{n} exp\Big\{-\frac{1}{2}z^2\Big\}. \qquad (1.8)
\end{aligned}$$

La factorización de la función de densidad conjunta en (1.8) implica que Z_i son variables aleatorias normal estándar mutuamente independientes, esto es $Z_i \sim N(0,1)$ y

$Z_i \perp\!\!\!\perp Z_j \; \forall \, i \neq j$ con $i, j \in \{1, 2 \ldots, n\}$. Entonces, $\mathbb{E}(\underline{Z}) = \underline{0}$ y $Var(\underline{Z}) = I_n$, donde $\underline{0}$ es el vector de ceros de dimensiones $n \times 1$. Así que

$$\mathbb{E}(\underline{Y}) = \mathbb{E}(\Sigma^{1/2}\underline{Z} + \underline{\mu}) = \Sigma^{1/2}\,\mathbb{E}(\underline{Z}) + \underline{\mu} = \underline{\mu},$$
$$Var(\underline{Y}) = Var(\Sigma^{1/2}\underline{Z} + \underline{\mu}) = \Sigma^{1/2} I_n \Sigma^{1/2} = \Sigma.$$

■

Se denota $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$ para indicar que $\underline{Y}$ tiene distribución normal n-variada con función de densidad (1.6).

1.2.2. Función generadora de momentos

Sea $\underline{Z} \sim N_n(\underline{0}, I_n)$, entonces por independencia de las Z_i's, la función generadora de momentos (f.g.m.) de $\underline{Z}$ es:

$$\begin{aligned} M_{\underline{Z}}(\underline{t}) &= \mathbb{E}\left[exp(\underline{t}^T\underline{Z})\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{n} exp(t_i Z_i)\right] \\ &= \prod_{i=1}^{n} \mathbb{E}\left[exp(t_i Z_i)\right] \\ &= \prod_{i=1}^{n} \exp\left(\frac{1}{2}t_i^2\right) \\ &= \exp\left(\frac{1}{2}\underline{t}^T\underline{t}\right), \end{aligned} \tag{1.9}$$

pues se sabe que la f.g.m. de una variable aleatoria normal estándar es $\mathbb{E}[exp(tZ)] = exp(\frac{1}{2}t^2)$.

Ahora bien, si $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, se puede reescribir $\underline{Y} = \Sigma^{1/2}\underline{Z} + \underline{\mu}$, donde $\underline{Z} \sim N_n(\underline{0}, I_n)$. Entonces, usando la f.g.m. dada por (1.9) se tiene:

$$\begin{aligned} M_{\underline{Y}}(\underline{t}) &= \mathbb{E}\left[exp\Big\{\underline{t}^T(\Sigma^{1/2}\underline{Z} + \underline{\mu})\Big\}\right] \\ &= \mathbb{E}\left[exp(\underline{t}^T\Sigma^{1/2}\underline{Z})exp(\underline{t}^T\underline{\mu})\right] \\ &= exp(\underline{t}^T\underline{\mu})\,\mathbb{E}\left[exp\Big\{(\Sigma^{1/2}\underline{t})^T\underline{Z}\Big\}\right] \\ &= exp(\underline{t}^T\underline{\mu})exp\Big\{\frac{1}{2}(\Sigma^{1/2}\underline{t})^T(\Sigma^{1/2}\underline{t})\Big\} \\ &= exp(\underline{t}^T\underline{\mu})exp\Big\{\frac{1}{2}\underline{t}^T\Sigma^{1/2}\Sigma^{1/2}\underline{t}\Big\} \\ &= exp\Big\{\underline{t}^T\underline{\mu} + \frac{1}{2}\underline{t}^T\Sigma\underline{t}\Big\}. \end{aligned} \tag{1.10}$$

■

Un resultado útil que puede probarse ocupando la f.g.m. es el siguiente:

Teorema 1.2.4 *Sea* $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$*, C una matriz de $m \times n$ de rango m, y $\underline{d}$ un vector de $n \times 1$. Entonces,*

$$C\underline{Y} + \underline{d} \sim N_m(C\underline{\mu} + \underline{d}, C\Sigma C^T).$$

Demostración:

$$\begin{aligned}
\mathbb{E}\left[exp\Big\{\underline{t}^T(C\underline{Y} + \underline{d})\Big\}\right] &= \mathbb{E}\left[exp\Big\{\underline{t}^T C\underline{Y} + \underline{t}^T\underline{d}\Big\}\right] \\
&= \mathbb{E}\left[exp\Big\{\underline{t}^T C\underline{Y}\Big\} exp\Big\{\underline{t}^T\underline{d}\Big\}\right] \\
&= exp\Big\{\underline{t}^T\underline{d}\Big\}\,\mathbb{E}\left[exp\Big\{(C^T\underline{t})^T\underline{Y}\Big\}\right] \\
&= exp\Big\{\underline{t}^T\underline{d}\Big\} exp\Big\{(C^T\underline{t})^T\underline{\mu} + \frac{1}{2}(C^T\underline{t})^T\Sigma(C^T\underline{t})\Big\} \\
&= exp\Big\{\underline{t}^T\underline{d} + \underline{t}^T C\underline{\mu} + \frac{1}{2}\underline{t}^T C\Sigma C^T\underline{t})\Big\} \\
&= exp\Big\{\underline{t}^T(C\underline{\mu} + \underline{d}) + \frac{1}{2}\underline{t}^T(C\Sigma C^T)\underline{t})\Big\}.
\end{aligned}$$

Ahora bien, como C es de rango m, entonces C^T tiene sus columnas linealmente independientes y, por lo tanto, $C^T\underline{x} = \underline{0} \Leftrightarrow \underline{x} = \underline{0}$. Si $\underline{w} = C^T\underline{x}$, entonces se tiene que $\underline{x}^T C\Sigma C^T\underline{x} = \underline{w}^T\Sigma\underline{w} > 0 \;\forall\; \underline{w} \neq \underline{0}$ pues Σ es d.p. y $\underline{w} = C^T\underline{x} \neq \underline{0} \;\forall\; \underline{x} \neq \underline{0} \;\therefore\; C\Sigma C^T$ es, también, d.p. C debe ser de rango m para garantizar que $C\Sigma C^T$ sea d.p., ya que la función de densidad normal multivariada (1.6) está definida solo para matriz de varianzas y covarianzas d.p.

■

1.2.3. Distribuciones marginales

Sean $\underline{Y}$ un vector aleatorio y $\underline{\mu}$ un vector, de dimensiones $n \times 1$, y Σ una matriz de $n \times n$ d.p. Se define una partición de la siguiente manera:

$$\underline{Y} = \begin{bmatrix}\underline{Y}_1 \\ \underline{Y}_2\end{bmatrix}, \quad \underline{\mu} = \begin{bmatrix}\underline{\mu}_1 \\ \underline{\mu}_2\end{bmatrix}, \quad y \quad \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}, \tag{1.11}$$

donde $\underline{Y}_1$ y $\underline{\mu}_1$ son vectores de dimensiones $n_1 \times 1$; $\underline{Y}_2$ y $\underline{\mu}_2$ son vectores de dimensiones $n_2 \times 1$, con $n_1, n_2 \in \mathbb{N}$ y $n_1 + n_2 = n$; Σ_{11}, Σ_{12}, Σ_{21} y Σ_{22} son submatrices de Σ de dimensiones $n_1 \times n_1$, $n_1 \times n_2$, $n_2 \times n_1$ y $n_2 \times n_2$, respectivamente. Nótese que $\Sigma_{12}^T = \Sigma_{21}$ pues Σ es simétrica.

Teorema 1.2.5 *Sea* $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$ *y considérese la partición definida en (1.11). Entonces,* $\underline{Y_1} \sim N_{n_1}(\underline{\mu}_1, \Sigma_{11})$.

Demostración:

Conviene definir $\underline{Y}_1 = B\underline{Y}$ con $B = (I_{n_1}, 0_{n_1 \times n_2})$ donde $0_{n_1 \times n_2}$ es la matriz de ceros de dimensiones $n_1 \times n_2$. Entonces, $B\underline{\mu} = \underline{\mu}_1$ y $B\Sigma B^T = \Sigma_{11}$ y aplicando el Teorema 1.2.4 queda demostrado.

Cabe aclarar que Σ_{11} es d.p. ya que Σ lo es por definición.

■

En otras palabras, el Teorema 1.2.5 señala que la distribución marginal de las primeras n_1 variables del vector $\underline{Y}$ siguen una distribución normal n_1-variada con media $\underline{\mu}_1$, esto es los primeros n_1 valores del vector $\underline{\mu}$, y matriz de varianzas y covarianzas Σ_{11}, esto es los primeros n_1 renglones y columnas de la matriz Σ.

Es importante mencionar que $\underline{Y}_1$ puede ser cualquier subconjunto de $\underline{Y}$, basta con reescribir de forma apropiada la matriz B.

En conclusión, la distribución marginal de una normal multivariada es, también, una normal multivariada.

1.2.4. Distribuciones condicionales

La función (1.6) define la densidad del vector aleatorio $\underline{Y}$. Si este vector se parte de la forma (1.11), la probabilidad condicional de las primeras n_1 componentes dadas las siguientes n_2 se denota como $f_{\underline{Y}_1|\underline{Y}_2}(\underline{y}_1|\underline{y}_2)$ y se define:

$$f_{\underline{Y}_1|\underline{Y}_2}(\underline{y}_1|\underline{y}_2) = \frac{f_{\underline{Y}_1,\underline{Y}_2}(\underline{y}_1, \underline{y}_2)}{f_{\underline{Y}_2}(\underline{y}_2)} = \frac{f_{\underline{Y}}(\underline{y})}{f_{\underline{Y}_2}(\underline{y}_2)}, \tag{1.12}$$

donde $f_{\underline{Y}_1,\underline{Y}_2}(\underline{y}_1, \underline{y}_2)$ es la densidad conjunta y esto es, de hecho, $f_{\underline{Y}}(\underline{y})$; $f_{\underline{Y}_2}(\underline{y}_2)$ es la función de probabilidad marginal (Teorema 1.2.5). Entonces:

$$\begin{aligned} f_{\underline{Y}_1|\underline{Y}_2}(\underline{y}_1|\underline{y}_2) &= \frac{\left[(2\pi)^{n/2} det(\Sigma)^{1/2}\right]^{-1} exp\Big\{ -\frac{1}{2}(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) \Big\}}{\left[(2\pi)^{n/2} det(\Sigma_{22})^{1/2}\right]^{-1} exp\Big\{ -\frac{1}{2}(\underline{y}_2 - \underline{\mu}_2)^T \Sigma_{22}^{-1} (\underline{y}_2 - \underline{\mu}_2) \Big\}} \\ &= \frac{exp\Big\{ -\frac{1}{2}\left[(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) - (\underline{y}_2 - \underline{\mu}_2)^T \Sigma_{22}^{-1} (\underline{y}_2 - \underline{\mu}_2)\right] \Big\}}{(2\pi)^{n_1/2} \frac{det(\Sigma)^{1/2}}{det(\Sigma_{22})^{1/2}}}. \end{aligned} \tag{1.13}$$

Ahora bien, para simplificar la expresión anterior conviene manipular algebraicamente el exponente. Para ello se define

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{11}^* & \Sigma_{12}^* \\ \Sigma_{21}^* & \Sigma_{22}^* \end{bmatrix}, \tag{1.14}$$

considerando las particiones (1.11).

Sabiendo que $\Sigma^{-1}\Sigma = I_n = \Sigma\Sigma^{-1}$ se derivan los siguientes dos sistemas de ecuaciones.

El primero, de $\Sigma^{-1}\Sigma = I_n$:

$$\begin{aligned} \Sigma_{11}^*\Sigma_{11} + \Sigma_{12}^*\Sigma_{21} &= I_{n_1} \\ \Sigma_{11}^*\Sigma_{12} + \Sigma_{12}^*\Sigma_{22} &= 0_{n_1\times n_2} \\ \Sigma_{21}^*\Sigma_{11} + \Sigma_{22}^*\Sigma_{21} &= 0_{n_2\times n_1} \\ \Sigma_{21}^*\Sigma_{12} + \Sigma_{22}^*\Sigma_{22} &= I_{n_2}. \end{aligned} \tag{1.15}$$

Y el segundo, de $\Sigma\Sigma^{-1} = I_n$:

$$\begin{aligned} \Sigma_{11}\Sigma_{11}^* + \Sigma_{12}\Sigma_{21}^* &= I_{n_1} \\ \Sigma_{11}\Sigma_{12}^* + \Sigma_{12}\Sigma_{22}^* &= 0_{n_1\times n_2} \\ \Sigma_{21}\Sigma_{11}^* + \Sigma_{22}\Sigma_{21}^* &= 0_{n_2\times n_1} \\ \Sigma_{21}\Sigma_{12}^* + \Sigma_{22}\Sigma_{22}^* &= I_{n_2}. \end{aligned} \tag{1.16}$$

De cada uno de ellos se obtienen expresiones para Σ_{11}^*, Σ_{12}^*, Σ_{21}^* y Σ_{22}^* que facilitan la simplificación del exponente en (1.13). Las expresiones para Σ_{11}^* y Σ_{12}^* se obtienen a partir del primer sistema (1.15) y para Σ_{21}^* y Σ_{22}^* del segundo (1.16). A continuación, se muestra detalladamente cómo se concluyen.

Por una parte, la forma matricial para resolver el sistema (1.15) es

$$\left(\begin{array}{cccc|c} \Sigma_{11} & \Sigma_{21} & 0 & 0 & I_{n_1} \\ \Sigma_{12} & \Sigma_{22} & 0 & 0 & 0_{n_1\times n_2} \\ 0 & 0 & \Sigma_{11} & \Sigma_{21} & 0_{n_2\times n_1} \\ 0 & 0 & \Sigma_{12} & \Sigma_{22} & I_{n_1} \end{array}\right),$$

multiplicando el segundo renglón por Σ_{22}^{-1} ($R_2 \to R_2\Sigma_{22}^{-1}$) y el tercer renglón por Σ_{11}^{-1} ($R_3 \to R_3\Sigma_{11}^{-1}$), ambas multiplicaciones por la derecha, se tiene [2]

$$\left(\begin{array}{cccc|c} \Sigma_{11} & \Sigma_{21} & 0 & 0 & I_{n_1} \\ \Sigma_{12}\Sigma_{22}^{-1} & I_{n_2} & 0 & 0 & 0_{n_1\times n_2} \\ 0 & 0 & I_{n_1} & \Sigma_{21}\Sigma_{11}^{-1} & 0_{n_2\times n_1} \\ 0 & 0 & \Sigma_{12} & \Sigma_{22} & I_{n_1} \end{array}\right),$$

[2]Recuérdese que Σ es d.p. por lo que Σ_{11} y Σ_{22} también son d.p. y, por lo tanto, son invertibles.

restando al primer renglón el producto del segundo por Σ_{21} ($R_1 \to R_1 - R_2\Sigma_{21}$) y al cuarto el producto del tercero por Σ_{12} ($R_4 \to R_4 - R_3 2\Sigma_{12}$), ambas multiplicaciones por la derecha, se tiene

$$\left(\begin{array}{cccc|c} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 & 0 & 0 & I_{n_1} \\ \Sigma_{12}\Sigma_{22}^{-1} & I_{n_2} & 0 & 0 & 0_{n_1\times n_2} \\ 0 & 0 & I_{n_1} & \Sigma_{21}\Sigma_{11}^{-1} & 0_{n_2\times n_1} \\ 0 & 0 & 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} & I_{n_1} \end{array}\right).$$

Tomando los primeros dos renglones de la matriz anterior se tienen las siguientes ecuaciones:

$$\begin{aligned} \Sigma_{11}^*(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) &= I_{n_1} \\ \Sigma_{11}^*\Sigma_{12}\Sigma_{22}^{-1} + \Sigma_{12}^* &= 0_{n_1\times n_2}, \end{aligned}$$

por lo tanto,

$$\Sigma_{11}^* = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} = \Sigma_*^{-1} \tag{1.17}$$

$$\Sigma_{12}^* = -\Sigma_{11}^*\Sigma_{12}\Sigma_{22}^{-1} = -\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1}, \tag{1.18}$$

donde $\Sigma_*^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ también es conocido como el complemento de Schur.

Por otra parte, la forma matricial para resolver el sistema (1.16) es:

$$\left(\begin{array}{cccc|c} \Sigma_{11}^* & \Sigma_{21}^* & 0 & 0 & I_{n_1} \\ \Sigma_{12}^* & \Sigma_{22}^* & 0 & 0 & 0_{n_1\times n_2} \\ 0 & 0 & \Sigma_{11}^* & \Sigma_{21}^* & 0_{n_2\times n_1} \\ 0 & 0 & \Sigma_{12}^* & \Sigma_{22}^* & I_{n_1} \end{array}\right),$$

multiplicando el segundo renglón por $(\Sigma_{22}^*)^{-1}$ ($R_2 \to R_2(\Sigma_{22}^*)^{-1}$) y el tercer renglón por $(\Sigma_{11}^*)^{-1}$ ($R_3 \to R_3\Sigma_{11}^{-1}$), ambas multiplicaciones por la derecha, se tiene [3]

$$\left(\begin{array}{cccc|c} \Sigma_{11}^* & \Sigma_{21}^* & 0 & 0 & I_{n_1} \\ \Sigma_{12}^*(\Sigma_{22}^*)^{-1} & I_{n_2} & 0 & 0 & 0_{n_1\times n_2} \\ 0 & 0 & I_{n_1} & \Sigma_{21}^*(\Sigma_{11}^*)^{-1} & 0_{n_2\times n_1} \\ 0 & 0 & \Sigma_{12}^* & \Sigma_{22}^* & I_{n_1} \end{array}\right),$$

restando al primer renglón el producto del segundo por Σ_{21}^* ($R_1 \to R_1 - R_2\Sigma_{21}^*$) y al cuarto el producto del tercero por Σ_{12}^* ($R_4 \to R_4 - R_3 2\Sigma_{12}^*$), ambas multiplicaciones por la derecha, se tiene

$$\left(\begin{array}{cccc|c} \Sigma_{11}^* - \Sigma_{12}^*(\Sigma_{22}^*)^{-1}\Sigma_{21}^* & 0 & 0 & 0 & I_{n_1} \\ \Sigma_{12}^*(\Sigma_{22}^*)^{-1} & I_{n_2} & 0 & 0 & 0_{n_1\times n_2} \\ 0 & 0 & I_{n_1} & \Sigma_{21}^*(\Sigma_{11}^*)^{-1} & 0_{n_2\times n_1} \\ 0 & 0 & 0 & \Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^* & I_{n_1} \end{array}\right).$$

[3] Recuérdese que Σ^{-1} es también d.p. (Teorema A.0.1) por lo que Σ_{11}^* y Σ_{22}^* también son d.p. y, por lo tanto, son invertibles.

Tomando los últimos dos renglones de la matriz anterior se tienen las siguientes ecuaciones:

$$\Sigma_{21} + \Sigma_{22}\Sigma_{21}^*(\Sigma_{11}^*)^{-1} = 0_{n_2 \times n_1} \tag{1.19}$$
$$\Sigma_{22}(\Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^*) = I_{n_2}. \tag{1.20}$$

Despejando Σ_{21}^* de (1.19) y sustituyendo (1.17) resulta:

$$\Sigma_{21}^* = -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1}. \tag{1.21}$$

Despejando Σ_{22}^* de (1.20) resulta:

$$\begin{aligned}\Sigma_{22}(\Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^*) &= I_{n_2}\\ \Sigma_{22}^* - \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^* &= \Sigma_{22}^{-1}\\ \Sigma_{22}^* &= \Sigma_{22}^{-1} + \Sigma_{21}^*(\Sigma_{11}^*)^{-1}\Sigma_{12}^*,\end{aligned}$$

sustituyendo (1.17), (1.18) y (1.21) queda

$$\begin{aligned}\Sigma_{22}^* &= \Sigma_{22}^{-1} + (-\Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1})(\Sigma_*)(-\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1})\\ &= \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_*^{-1}\Sigma_*)\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ &= \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1}.\end{aligned} \tag{1.22}$$

En resumen, las expresiones (1.17), (1.18), (1.21) y (1.22), se ocuparán en la partición de Σ^{-1} definida en (1.14).

Entonces,

$$\begin{aligned}(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu}) &= \begin{bmatrix}\underline{y}_1-\underline{\mu}_1\\ \underline{y}_2-\underline{\mu}_2\end{bmatrix}^T \begin{bmatrix}\Sigma_{11}^* & \Sigma_{12}^*\\ \Sigma_{21}^* & \Sigma_{22}^*\end{bmatrix} \begin{bmatrix}\underline{y}_1-\underline{\mu}_1\\ \underline{y}_2-\underline{\mu}_2\end{bmatrix}\\
&= (\underline{y}_1-\underline{\mu}_1)^T\Sigma_{11}^*(\underline{y}_1-\underline{\mu}_1) + (\underline{y}_1-\underline{\mu}_1)^T\Sigma_{12}^*(\underline{y}_2-\underline{\mu}_2)\\
&\quad + (\underline{y}_2-\underline{\mu}_2)^T\Sigma_{21}^*(\underline{y}_1-\underline{\mu}_1) + (\underline{y}_2-\underline{\mu}_2)^T\Sigma_{22}^*(\underline{y}_2-\underline{\mu}_2)\\
&= (\underline{y}_1-\underline{\mu}_1)^T\Sigma_*^{-1}(\underline{y}_1-\underline{\mu}_1)\\
&\quad - (\underline{y}_1-\underline{\mu}_1)^T(\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1})(\underline{y}_2-\underline{\mu}_2)\\
&\quad - (\underline{y}_2-\underline{\mu}_2)^T(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1})(\underline{y}_1-\underline{\mu}_1)\\
&\quad + (\underline{y}_2-\underline{\mu}_2)^T\Sigma_{22}^{-1}(\underline{y}_2-\underline{\mu}_2)\\
&\quad + (\underline{y}_2-\underline{\mu}_2)^T(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1})(\underline{y}_2-\underline{\mu}_2)\\
&= (\underline{y}_1-(\underline{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2-\underline{\mu}_2)))^T\Sigma_*^{-1}(\underline{y}_1-(\underline{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2-\underline{\mu}_2)))\\
&\quad + (\underline{y}_2-\underline{\mu}_2)^T\Sigma_{22}^{-1}(\underline{y}_2-\underline{\mu}_2)\\
&= (\underline{y}_1-\underline{\mu}_*)^T\Sigma_*^{-1}(\underline{y}_1-\underline{\mu}_*) + (\underline{y}_2-\underline{\mu}_2)^T\Sigma_{22}^{-1}(\underline{y}_2-\underline{\mu}_2),\end{aligned} \tag{1.23}$$

donde $\underline{\mu}_* = \underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2)$.

En cuanto al cociente de determinantes de la expresión (1.13), se puede simplificar usando el determinante de una matriz particionada

$$\begin{aligned} det(\Sigma) &= det(\Sigma_{22})det(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) = det(\Sigma_{22})det(\Sigma_*), \\ \therefore \quad \frac{det(\Sigma)}{det(\Sigma_{22})} &= det(\Sigma_*). \end{aligned} \tag{1.24}$$

Por último, sustituyendo (1.23) y (1.24) en (1.13) resulta:

$$f_{\underline{Y}_1|\underline{Y}_2}(\underline{y}_1|\underline{y}_2) = \left[(2\pi)^{n_1/2}det(\Sigma_*)^{1/2}\right]^{-1} exp\Big\{ -\frac{1}{2}(\underline{y}_1 - \underline{\mu}_*)^T\Sigma_*^{-1}(\underline{y}_1 - \underline{\mu}_*)\Big\}, \tag{1.25}$$

y por la Definición (1.2.1) se tiene que

$$f_{\underline{Y}_1|\underline{Y}_2}(\underline{y}_1|\underline{y}_2) \sim N_{n_1}(\underline{\mu}_*, \Sigma_*), \tag{1.26}$$

donde $\underline{\mu}_* = \underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{y}_2 - \underline{\mu}_2)$ y $\Sigma_* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ son el vector de medias y la matriz de varianzas y covarianzas condicionales.

1.2.5. Independencia

Si dos variables aleatorias son independientes, entonces la f.g.m. de la suma es igual al producto de las f.g.m de cada una.

Teorema 1.2.6 *Sea $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$ y sean las particiones como se definieron en (1.11). Entonces, $\underline{Y}_1$ y $\underline{Y}_2$ son independientes si y solo si $\Sigma_{12} = 0_{n_1 \times n_2}$.*

Demostración:

En (1.10) se obtuvo que la f.g.m. de $\underline{Y}$ es $exp\big\{\underline{t}^T\underline{\mu} + \frac{1}{2}\underline{t}^T\Sigma\underline{t}\big\}$. Haciendo una partición a $\underline{t}$ del mismo modo que para $\underline{Y}$ la f.g.m. queda:

$$\begin{aligned} M_{\underline{Y}}(\underline{t}) &= exp\left(\begin{bmatrix}\underline{t}_1 \\ \underline{t}_2\end{bmatrix}^T \begin{bmatrix}\underline{\mu}_1 \\ \underline{\mu}_2\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\underline{t}_1 \\ \underline{t}_2\end{bmatrix}^T \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} \begin{bmatrix}\underline{t}_1 \\ \underline{t}_2\end{bmatrix}\right) \\ &= exp\left(\underline{t}_1^T\underline{\mu}_1 + \underline{t}_2^T\underline{\mu}_2 + \frac{1}{2}\underline{t}_1^T\Sigma_{11}\underline{t}_1 + \frac{1}{2}\underline{t}_2^T\Sigma_{21}\underline{t}_1 + \frac{1}{2}\underline{t}_1^T\Sigma_{12}\underline{t}_2 + \frac{1}{2}\underline{t}_2^T\Sigma_{22}\underline{t}_2\right). \end{aligned}$$

Ahora bien, dado que $\underline{t}_2^T\Sigma_{21}\underline{t}_1$ es un escalar, en consecuencia, es igual a su transpuesta, y dado que Σ es, por definición, d.p. entonces es simétrica y, por lo tanto, $\Sigma_{21} = \Sigma_{12}^T$. Así que $\underline{t}_2^T\Sigma_{21}\underline{t}_1 = (\underline{t}_2^T\Sigma_{21}\underline{t}_1)^T = \underline{t}_1^T\Sigma_{21}^T\underline{t}_2 = \underline{t}_1^T\Sigma_{12}\underline{t}_2$. Entonces la f.g.m. resulta

$$M_{\underline{Y}}(\underline{t}) = exp\left(\underline{t}_1^T\underline{\mu}_1 + \underline{t}_2^T\underline{\mu}_2 + \frac{1}{2}\underline{t}_1^T\Sigma_{11}\underline{t}_1 + \frac{1}{2}\underline{t}_2^T\Sigma_{22}\underline{t}_2 + \underline{t}_1^T\Sigma_{12}\underline{t}_2\right). \tag{1.27}$$

Si $\Sigma_{12} = 0_{n_1 \times n_2}$ el exponente puede ser escrito como

$$M_{\underline{Y}}(\underline{t}) = exp\left(\underline{t}_1^T \underline{\mu}_1 + \frac{1}{2}\underline{t}_1^T \Sigma_{11} \underline{t}_1\right) exp\left(\underline{t}_2^T \underline{\mu}_2 + \frac{1}{2}\underline{t}_2^T \Sigma_{22} \underline{t}_2\right) = M_{\underline{Y}_1}(\underline{t}_1) M_{\underline{Y}_2}(\underline{t}_2),$$

lo que implica que $\underline{Y}_1$ y $\underline{Y}_2$ son independientes.

Por el contrario, si $\underline{Y}_1$ y $\underline{Y}_2$ son independientes, entonces

$$M_{\underline{Y}}(\underline{t}_1, 0) M_{\underline{Y}}(\underline{t}_2, 0) = M_{\underline{Y}}(\underline{t}_1, \underline{t}_2),$$

lo que implica que en (1.27) $\underline{t}_1^T \Sigma_{12} \underline{t}_2 = 0 \ \forall \ \underline{t}_1$ y $\underline{t}_2$, por lo que $\Sigma_{12} = 0_{n_1 \times n_2}$.

■

El Teorema (1.2.6) establece entonces una condición necesaria y suficiente para que dos vectores con distribución normal sean mutuamente independientes, ésta es que en particiones del tipo $\Sigma = \{\Sigma_{ij}\}$ con $i, j = 1, 2, \ldots, n$ sea $\Sigma_{ij} = 0_{n_i \times n_j}$ para $i \neq j$.

Un resultado interesante y útil para el análisis de regresión es el siguiente:

Teorema 1.2.7 *Sea $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$ y se define $U = A\underline{Y}$ y $V = B\underline{Y}$. Entonces, U y V son independientes si y solo si $cov(U, V) = A\Sigma B^T = 0$.*

Demostración: Sea

$$W = \begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} \underline{Y},$$

por el Teorema (1.2.4) el vector aleatorio W es normal multivariado con matriz de varianzas y covarianzas

$$Var(W) = \begin{bmatrix} A \\ B \end{bmatrix} Var(\underline{Y}) \begin{bmatrix} A^T & B^T \end{bmatrix} = \begin{bmatrix} A\Sigma A^T & A\Sigma B^T \\ B\Sigma A^T & B\Sigma B^T \end{bmatrix}.$$

Así que por el Teorema (1.2.6), U y V son independientes si y solo si $A\Sigma B^T = 0$.

■

1.2.6. Forma cuadrática

Teorema 1.2.8 *Sean $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$ y A una matriz simétrica real de $n \times n$. Si [4] $A\Sigma$ es idempotente y de rango m, entonces $\underline{Y}^T A \underline{Y}$ (conocida como la forma cuadrática) se*

[4] El teorema se extiende en la otra dirección también, es decir, es *si y solo si*, pero para desarrollar la teoría inferencial en el presente trabajo solo se necesita la suficiencia, por lo que solo se prueba esto. La demostración de necesidad puede encontrarse en Searle (1997)[12] o Driscoll (1999)[6].

distribuye χ^2 con m grados de libertad (g.l.) y $\lambda = \underline{\mu}^T A \underline{\mu}$ parámetro de no centralidad, es decir, $\underline{Y}^T A \underline{Y} \sim \chi^2_m(\lambda)$.

Demostración:

Primero, se sabe que:

(i) Si A^{-1} existe, es simétrica si y solo si A es simétrica.

(ii) Si M es matriz de $n \times n$ de rango r e idempotente, esto es $M^2 = M$, entonces exactamente r de sus valores propios son 1 y $n - r$ son 0.

(iii) Si M es matriz de $n \times n$, entonces $(I_n - M)^{-1} = \sum_{k=0}^{\infty} M^k$ si y solo si todos los valores propios de M tienen valor absoluto menor a 1.

(iv) Si Q es v.a. χ^2 con v g.l. y λ parámetro de no centralidad, esto es, si $Q \sim \chi^2_v(\lambda)$, su f.g.m. es

$$M_Q(t) = (1-2t)^{-\frac{v}{2}} exp\left(\frac{\lambda t}{1-2t}\right). \tag{1.28}$$

Luego, la f.g.m. de $\underline{y}^T A \underline{y}$, por definición, es:

$$\begin{aligned} M_{\underline{Y}^T A \underline{Y}}(t) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} exp\{t \underline{y}^T A \underline{y}\} f(\underline{y})\, dy_1 \cdots dy_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} exp\{t \underline{y}^T A \underline{y}\} \left[(2\pi)^{n/2} det(\Sigma)^{1/2}\right]^{-1} \\ &\qquad exp\Big\{ -\frac{1}{2}(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) \Big\}\, dy_1 \cdots dy_n \\ &= \left[(2\pi)^{n/2} det(\Sigma)^{1/2}\right]^{-1} \\ &\qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} exp\Big\{ t \underline{y}^T A \underline{y} - \frac{1}{2}(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) \Big\}\, dy_1 \cdots dy_n. \end{aligned} \tag{1.29}$$

Trabajando el exponente de (1.29) se tiene

$$\begin{aligned} & t \underline{y}^T A \underline{y} - \frac{1}{2}(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) \\ &\quad = t \underline{y}^T A \underline{y} - \frac{1}{2}(\underline{y}^T \Sigma^{-1} - \underline{\mu}^T \Sigma^{-1})(\underline{y} - \underline{\mu}) \\ &\quad = t \underline{y}^T A \underline{y} - \frac{1}{2}(\underline{y}^T \Sigma^{-1} \underline{y} - \underline{y}^T \Sigma^{-1} \underline{\mu} - \underline{\mu}^T \Sigma^{-1} \underline{y} + \underline{\mu}^T \Sigma^{-1} \underline{\mu}) \\ &\quad = t \underline{y}^T A \underline{y} - \frac{1}{2} \underline{y}^T \Sigma^{-1} \underline{y} + \frac{1}{2} \underline{y}^T \Sigma^{-1} \underline{\mu} + \frac{1}{2} \underline{\mu}^T \Sigma^{-1} \underline{y} - \frac{1}{2} \underline{\mu}^T \Sigma^{-1} \underline{\mu}, \end{aligned}$$

dado que $\underline{\mu}^T\Sigma^{-1}\underline{y}$ es una matriz de 1×1 o un escalar, entonces es igual a su transpuesta $\underline{\mu}^T\Sigma^{-1}\underline{y} = (\underline{\mu}^T\Sigma^{-1}\underline{y})^T = \underline{y}^T\Sigma^{-1}\underline{\mu}$ (por hipótesis Σ es d.p. y por el Teorema A.0.1 Σ^{-1} también lo es y, por lo tanto, es simétrica). Además dado que t es una variable y no un vector se tiene que $t\underline{y}^T A\underline{y} = \underline{y}^T\, tA\underline{y}$. Por todo lo anterior el exponente queda:

$$\begin{aligned} & t\underline{y}^T A\underline{y} - \frac{1}{2}(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu}) \\ &\quad = \underline{y}^T\, tA\underline{y} - \frac{1}{2}\underline{y}^T\Sigma^{-1}\underline{y} + \underline{\mu}^T\Sigma^{-1}\underline{y} - \frac{1}{2}\underline{\mu}^T\Sigma^{-1}\underline{\mu} \\ &\quad = -\frac{1}{2}\underline{y}^T(-2tA+\Sigma^{-1})\underline{y} + \underline{\mu}^T\Sigma^{-1}\underline{y} - \frac{1}{2}\underline{\mu}^T\Sigma^{-1}\underline{\mu} \\ &\quad = -\frac{1}{2}\underline{y}^T(I_n - 2tA\Sigma)\Sigma^{-1}\underline{y} + \underline{\mu}^T\Sigma^{-1}\underline{y} - \frac{1}{2}\underline{\mu}^T\Sigma^{-1}\underline{\mu}, \end{aligned}$$

se definen $W = [(I_n - 2tA\Sigma)\Sigma^{-1}]^{-1} = \Sigma(I_n - 2tA\Sigma)^{-1}$, $\underline{g}^T = \underline{\mu}^T\Sigma^{-1}W$ y $k = -\frac{1}{2}\underline{\mu}^T\Sigma^{-1}\underline{\mu}$. Obsérvese que $\underline{g}^T W^{-1} = \underline{\mu}^T\Sigma^{-1}WW^{-1} = \underline{\mu}^T\Sigma^{-1}$. Entonces se tiene

$$\begin{aligned} & t\underline{y}^T A\underline{y} - \frac{1}{2}(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu}) \\ &\quad = -\frac{1}{2}\underline{y}^T W^{-1}\underline{y} + \underline{g}^T W^{-1}\underline{y} + k \\ &\quad = -\frac{1}{2}\underline{y}^T W^{-1}\underline{y} + \underline{g}^T W^{-1}\underline{y} - \frac{1}{2}\underline{g}^T W^{-1}\underline{g} + \frac{1}{2}\underline{g}^T W^{-1}\underline{g} + k \\ &\quad = -\frac{1}{2}\left[\underline{y}^T W^{-1}\underline{y} - 2\underline{g}^T W^{-1}\underline{y} + \underline{g}^T W^{-1}\underline{g}\right] + \frac{1}{2}\underline{g}^T W^{-1}\underline{g} + k \\ &\quad = -\frac{1}{2}\left[\underline{y}^T W^{-1}\underline{y} - \underline{g}^T W^{-1}\underline{y} - \underline{g}^T W^{-1}\underline{y} + \underline{g}^T W^{-1}\underline{g}\right] + j + k \\ &\quad = -\frac{1}{2}\left[(\underline{y}^T W^{-1} - \underline{g}^T W^{-1})\underline{y} - (\underline{y}^T W^{-1} + \underline{g}^T W^{-1})\underline{g}\right] + j + k \\ &\quad = -\frac{1}{2}\left[(\underline{y}^T W^{-1} - \underline{g}^T W^{-1})(\underline{y}-\underline{g})\right] + j + k \\ &\quad = -\frac{1}{2}\left[(\underline{y}-\underline{g})^T W^{-1}(\underline{y}-\underline{g})\right] + j + k, \end{aligned} \tag{1.30}$$

con $j = \frac{1}{2}\underline{g}^T W^{-1}\underline{g}$. Sustituyendo (1.30) en (1.29) queda

$$\begin{aligned} M_{\underline{Y}^T A\underline{Y}}(t) &= \left[(2\pi)^{n/2} det(\Sigma)^{1/2}\right]^{-1} \\ &\quad \times \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} exp\Big\{ -\frac{1}{2}\left[(\underline{y}-\underline{g})^T W^{-1}(\underline{y}-\underline{g})\right] + j + k\Big\}\, dy_1\cdots dy_n \\ &= \frac{exp\Big\{j+k\Big\}}{(2\pi)^{n/2}det(\Sigma)^{1/2}}\cdot\frac{(2\pi)^{n/2}det(W)^{1/2}}{(2\pi)^{n/2}det(W)^{1/2}} \\ &\quad \times \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} exp\Big\{ -\frac{1}{2}\left[(\underline{y}-\underline{g})^T W^{-1}(\underline{y}-\underline{g})\right]\Big\}\, dy_1\cdots dy_n \\ &= \frac{exp\Big\{j+k\Big\}det(W)^{1/2}}{det(\Sigma)^{1/2}} \end{aligned}$$

$$\times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[(2\pi)^{n/2} det(W)^{1/2}\right]^{-1} exp\Big\{ -\frac{1}{2}\left[(\underline{y}-\underline{g})^T W^{-1}(\underline{y}-\underline{g})\right]\Big\}\, dy_1 \cdots dy_n. \tag{1.31}$$

Es importante mencionar que se puede multiplicar por $\frac{(2\pi)^{n/2}det(W)^{1/2}}{(2\pi)^{n/2}det(W)^{1/2}}$ ya que el determinante de W es diferente de cero pues W^{-1} existe. Además, obsérvese que

$$\begin{aligned} W^{-1} &= (I_n - 2tA\Sigma)\Sigma^{-1} \\ &= \Sigma^{-1} - 2tA\Sigma\Sigma^{-1} \\ &= \Sigma^{-1} - 2tA, \end{aligned}$$

y dado que tanto Σ^{-1} como A son simétricas, entonces W^{-1} lo es, y por (i) se concluye que W también es simétrica.

Por otro lado, la función dentro de las integrales resulta ser la función de probabilidad normal con media $\underline{g}$ y matriz de varianzas y covarianzas W, por lo tanto, integra 1 (obsérvese Teorema 1.2.2).

Entonces, la expresión (1.31) queda:

$$\begin{aligned} M_{\underline{Y}^T A \underline{Y}}(t) &= \frac{exp\Big\{\frac{1}{2}\underline{g}^T W^{-1}\underline{g} - \frac{1}{2}\underline{\mu}^T\Sigma^{-1}\underline{\mu}\Big\} det(\Sigma(I_n - 2tA\Sigma)^{-1})^{1/2}}{det(\Sigma)^{1/2}} \cdot 1 \\ &= \frac{exp\Big\{\frac{1}{2}\underline{\mu}^T\Sigma^{-1}WW^{-1}W\Sigma^{-1}\underline{\mu} - \frac{1}{2}\underline{\mu}^T\Sigma^{-1}\underline{\mu}\Big\}}{det(I_n - 2tA\Sigma)^{1/2}} \\ &= \frac{exp\Big\{-\frac{1}{2}\underline{\mu}^T\Sigma^{-1}\Sigma(I_n - 2tA\Sigma)^{-1}\Sigma^{-1}\underline{\mu} - \frac{1}{2}\underline{\mu}^T\Sigma^{-1}\underline{\mu}\Big\}}{det(I_n - 2tA\Sigma)^{1/2}} \\ &= \frac{exp\Big\{-\frac{1}{2}\underline{\mu}^T[-(I_n - 2tA\Sigma)^{-1} + I_n]\Sigma^{-1}\underline{\mu}\Big\}}{det(I_n - 2tA\Sigma)^{1/2}} \\ &= \frac{exp\Big\{-\frac{1}{2}\underline{\mu}^T[I_n - (I_n - 2tA\Sigma)^{-1}]\Sigma^{-1}\underline{\mu}\Big\}}{det(I_n - 2tA\Sigma)^{1/2}}. \end{aligned} \tag{1.32}$$

Sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los valores propios de $A\Sigma$, entonces los de $-2tA\Sigma$ son $-2t\lambda_1, -2t\lambda_2, \ldots, -2t\lambda_n$ y, por lo tanto, los de $I_n - 2tA\Sigma$ son $1-2t\lambda_1, 1-2t\lambda_2, \ldots, 1-2t\lambda_n$. Entonces

$$det(I_n - 2tA\Sigma) = \prod_{i=1}^{n}(1 - 2t\lambda_i).$$

Ahora, dado que $A\Sigma$ es idempotente y de rango m por hipótesis del teorema, entonces por (ii) m de sus valores propios son 1 y $n-m$ son 0. Así que

$$det(I_n - 2tA\Sigma) = (1-2t)^m. \tag{1.33}$$

Luego, si $t < 1/2$, por (iii) se tiene la serie infinita

$$\begin{aligned}(I_n - 2tA\Sigma)^{-1} &= \sum_{k=0}^{\infty}(2tA\Sigma)^k \\ &= I_n + \sum_{k=1}^{\infty}(2tA\Sigma)^k \\ &= I_n + \sum_{k=1}^{\infty}(2t)^k(A\Sigma)^k \\ &= I_n + \sum_{k=1}^{\infty}(2t)^k(A\Sigma) \\ &= I_n + [(1-2t)^{-1} - 1](A\Sigma), \end{aligned} \tag{1.34}$$

pues $A\Sigma$ es idempotente y suponiendo $|t| < 1$ la serie geométrica converge.

Sustituyendo (1.33) y (1.34) en (1.32) se tiene

$$\begin{aligned}M_{\underline{Y}^T A\underline{Y}}(t) &= (1-2t)^{-m/2}exp\Big\{ -\frac{1}{2}\underline{\mu}^T \left[-[(1-2t)^{-1} - 1]A\Sigma\right]\Sigma^{-1}\underline{\mu}\Big\} \\ &= (1-2t)^{-m/2}exp\left\{ -\frac{1}{2}\underline{\mu}^T \left[1 - \frac{1}{1-2t}\right] A\underline{\mu}\right\} \\ &= (1-2t)^{-m/2}exp\left\{ \frac{t}{1-2t}\underline{\mu}^T A\underline{\mu}\right\},\end{aligned}$$

que, comparando con (1.28) es la f.g.m. de una v.a. con distribución χ^2 con m g.l. y $\lambda = \underline{\mu}^T A\underline{\mu}$ parámetro de no centralidad.

■

Teorema 1.2.9 *Sea $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, las formas cuadráticas $\underline{Y}^T A\underline{Y}$ y $\underline{Y}^T B\underline{Y}$ con A y B simétricas, se distribuyen de forma independiente si*[5] *$A\Sigma B = 0$ ó $B\Sigma A = 0$ (siendo estas últimas expresiones equivalentes).*

Demostración:

La condición $A\Sigma B = 0$ es equivalente a $B\Sigma A = 0$, dado que A, B y Σ son simétricas. Por lo tanto, cada condición implica la otra.

[5]El teorema se extiende a la otra dirección también, es decir, es *si y solo si*, pero para el desarrollo de la teoría inferencial en el presente trabajo solo se necesita la suficiencia, por lo que solo se prueba esto. La demostración de necesidad puede encontrarse en Searle (1997)[12].

Se sabe que para cualquier matriz simétrica A de $n \times n$ con $rango(A) = r \; \exists \; L$ tal que $A = LL^T$ con L de $n \times r$, es decir, de columnas linealmente independientes, de igual forma $\exists \; M$ tal que $B = MM^t$.

Luego, si $A\Sigma B = 0 \;\Rightarrow\; LL^T\Sigma MM^t = 0$ y, dado que $(L^TL)^{-1}$ y $(M^TM)^{-1}$ existen pues son de rango completo, se tiene $L^T\Sigma M = 0$. Así que

$$Cov(L^T\underline{Y}, \underline{Y}^Y M) = L^T\Sigma M = 0,$$

y, dado que $\underline{Y}$ tiene distribución normal, $L^T\underline{Y}$ y $\underline{Y}^Y M$ son independientes por Teorema 1.2.7. En consecuencia, $\underline{Y}^T A\underline{Y} = \underline{Y}^T LL^T\underline{Y}$ y $\underline{Y}^T B\underline{Y} = \underline{Y}^T MM^T\underline{Y}$ son independientes.

■

Capítulo 2

Estimación por mínimos cuadrados

2.1. Estimadores

Considérense el modelo de regresión lineal dado en la expresión (1.1) y los supuestos descritos en la sección 1.1.2. En concordancia con el supuesto (II) se espera que la diferencia expresada en la ec. (1.5) sea *pequeña* pues $\mathbb{E}(\varepsilon_i) = 0$. En esta expresión y_i y las k variables explicativas $x_{i1}, x_{i2}, \ldots, x_{ik}$ son valores observados y $\beta_0, \beta_1, \ldots, \beta_k$ son parámetros desconocidos que, de saber su valor, podría calcularse el valor esperado de la variable de interés (obsérvese la ec. (1.4)).

Ahora bien, aunque no se conoce el valor real de estos parámetros, sí se pueden estimar, más aún, existen varios métodos para hacerlo y uno de los más comunes es el que se conoce como *mínimos cuadrados*.

La estimación por mínimos cuadrados consiste en minimizar la suma de cuadrados de la diferencia del valor observado y_i con su valor esperado $\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}$ (obsérvese el supuesto (IV)). Para ello se define la siguiente función:

$$\begin{aligned} S(\beta_0, \beta_1, \ldots, \beta_k) &= \sum_{i=1}^{n} (y_i - \mathbb{E}(Y_i))^2, \\ &= \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})\right)^2, \end{aligned} \tag{2.1}$$

y los estimadores para $\beta_0, \beta_1, \ldots, \beta_k$, denotados como $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, son

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) \in \underset{\beta_0, \beta_1, \ldots, \beta_k}{\arg\min}\ S, \tag{2.2}$$

es decir, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ son aquellos que minimizan la distancia entre el valor real observado y_i y el valor ajustado por el modelo de regresión lineal $\hat{y}_i$, donde

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}. \tag{2.3}$$

En otras palabras, el lado derecho de la expresión (2.3) no contiene parámetros desconocidos pues las $x_{i1}, x_{i2}, \ldots, x_{ik}$ son los valores observados de las variables explicativas y $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ son las estimaciones de los parámetros desconocidos a través de mínimos cuadrados. Al evaluar esta expresión se obtiene un valor ajustado que se denota como $\hat{y}_i$.

Para encontrar estos valores se hace uso de cálculo diferencial de varias variables y de la notación matricial del modelo[1], esto último facilita los cálculos.

Considerando la notación matricial descrita en la sección 1.1.1, la función (2.1) se puede reescribir como:

$$S(\underline{\beta}) = (\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}). \tag{2.4}$$

Teorema 2.1.1 *Sea $X \in \mathbb{R}^{n\times p}$ de rango p con $p < n$, entonces la función (2.4) se minimiza en $(X^TX)^{-1}X^T\underline{y}$.*

Demostración:

Sea $\hat{\underline{\beta}} \in \arg\min_{\underline{\beta}} S(\underline{\beta})$. Entonces $\hat{\underline{\beta}}$ satisface:

$$\frac{\partial}{\partial \underline{\beta}} S(\underline{\beta})\Big|_{\underline{\beta}=\hat{\underline{\beta}}} = 0. \tag{2.5}$$

Primero, obsérvese que:

$$\begin{aligned} S(\underline{\beta}) &= (\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) \\ &= (\underline{y}^T - \underline{\beta}^T X^T)(\underline{y} - X\underline{\beta}) \\ &= \underline{y}^T\underline{y} - \underline{y}^T X\underline{\beta} - \underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta} \\ &= \underline{y}^T\underline{y} - 2\underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta}. \end{aligned} \tag{2.6}$$

Dado que $\underline{\beta}^T X^T\underline{y}$ es una matriz de 1×1, o un escalar, es igual a su transpuesta $\underline{\beta}^T X^T\underline{y} = (\underline{\beta}^T X^T\underline{y})^T = \underline{y}^T X\underline{\beta}$, por lo tanto, se pueden reducir los términos justo como se hace en el último renglón.

Luego, se tiene:

$$\begin{aligned} \frac{\partial}{\partial \underline{\beta}} S(\underline{\beta}) &= \frac{\partial}{\partial \underline{\beta}} \left(\underline{y}^T\underline{y} - 2\underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta}\right) \\ &= -2X^T\underline{y} + 2X^T X\underline{\beta}, \end{aligned}$$

y considerando la ec. (2.5) se tiene:

$$-2X^T\underline{y} + 2X^T X\hat{\underline{\beta}} = 0,$$

[1]Existe también un argumento geométrico que permite deducir el estimador, véase el Anexo B.

lo que implica

$$\underline{\hat{\beta}} = \left(X^T X\right)^{-1} X^T \underline{y}, \tag{2.7}$$

como $X \in \mathbb{R}^{n\times p}$ con $p < n$ y rango p por hipótesis, aplicando el Teorema A.0.4, $(X^T X)^{-1}$ existe. De aquí surge la necesidad del supuesto (V) descrito en (1.1.2).

Luego, para probar que el valor (2.7) minimiza la función $S(\underline{\beta})$ obsérvese que la siguiente igualdad es cierta siempre que X sea de rango p:

$$(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) = (\underline{y} - X\underline{\hat{\beta}})^T(\underline{y} - X\underline{\hat{\beta}}) + (\underline{\hat{\beta}} - \underline{\beta})^T X^T X(\underline{\hat{\beta}} - \underline{\beta}), \tag{2.8}$$

para probarlo conviene definir $H = X\left(X^T X\right)^{-1} X^T$, nótese que H es simétrica pues $H^T = H$. Entonces:

$$\begin{aligned}
&(\underline{y} - X\underline{\hat{\beta}})^T(\underline{y} - X\underline{\hat{\beta}}) + (\underline{\hat{\beta}} - \underline{\beta})^T X^T X(\underline{\hat{\beta}} - \underline{\beta}) \\
&\quad = (\underline{y} - X(X^T X)^{-1}X^T\underline{y})^T(\underline{y} - X(X^T X)^{-1}X^T\underline{y}) \\
&\qquad + ((X^T X)^{-1}X^T\underline{y} - \underline{\beta})^T X^T X((X^T X)^{-1}X^T\underline{y} - \underline{\beta}) \\
&\quad = (\underline{y}^T - \underline{y}^T X(X^T X)^{-1}X^T)(\underline{y} - X(X^T X)^{-1}X^T\underline{y}) \\
&\qquad + (\underline{y}^T X(X^T X)^{-1} - \underline{\beta}^T)X^T X((X^T X)^{-1}X^T\underline{y} - \underline{\beta}) \\
&\quad = \underline{y}^T(I - X(X^T X)^{-1}X^T)(I - X(X^T X)^{-1}X^T)\underline{y} \\
&\qquad + (\underline{y}^T X(X^T X)^{-1}X^T X - \underline{\beta}^T X^T X)((X^T X)^{-1}X^T\underline{y} - \underline{\beta}) \\
&\quad = \underline{y}^T(I - X(X^T X)^{-1}X^T - X(X^T X)^{-1}X^T + X(X^T X)^{-1}X^T X(X^T X)^{-1}X^T)\underline{y} \\
&\qquad + (\underline{y}^T X I - \underline{\beta}^T X^T X)((X^T X)^{-1}X^T\underline{y} - \underline{\beta}) \\
&\quad = \underline{y}^T(I - X(X^T X)^{-1}X^T - X(X^T X)^{-1}X^T + XI(X^T X)^{-1}X^T)\underline{y} \\
&\qquad + \underline{y}^T X(X^T X)^{-1}X^T\underline{y} - \underline{y}^T X\underline{\beta} - \underline{\beta}^T X^T X(X^T X)^{-1}X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta} \\
&\quad = \underline{y}^T(I - X(X^T X)^{-1}X^T)\underline{y} \\
&\qquad + \underline{y}^T X(X^T X)^{-1}X^T\underline{y} - \underline{y}^T X\underline{\beta} - \underline{\beta}^T I X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta} \\
&\quad = \underline{y}^T\underline{y} - \underline{y}^T X(X^T X)^{-1}X^T\underline{y} \\
&\qquad + \underline{y}^T X(X^T X)^{-1}X^T\underline{y} - \underline{y}^T X\underline{\beta} - \underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta} \\
&\quad = \underline{y}^T\underline{y} - \underline{y}^T X\underline{\beta} - \underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta} \\
&\quad = \underline{y}^T(\underline{y} - X\underline{\beta}) - \underline{\beta}^T X^T(\underline{y} - X\underline{\beta}) \\
&\quad = (\underline{y}^T - \underline{\beta}^T X^T)(\underline{y} - X\underline{\beta}) \\
&\quad = (\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}).
\end{aligned}$$

Lo que prueba que la expresión (2.8) es cierta. Más aún, por el Teorema A.0.4 se sabe que $X^T X$ es d.p. por lo que el segundo sumando del lado derecho es no negativo, lo que permite concluir que la expresión (2.8) se minimiza cuando $\underline{\beta} = \underline{\hat{\beta}}$, lo que confirma que $\underline{\hat{\beta}}$ minimiza la función (2.4).

■

Definición 2.1.1 *Sean $\underline{y}$ y X los valores observados, como se definen en la sección 1.1.1. Entonces $\underline{\hat{\beta}}$ dado por la expresión (2.7) se conoce como el estimador por mínimos cuadrados del modelo*

$$\mathbb{E}(\underline{Y}) = X\beta.$$

El vector de valores ajustados queda:

$$\underline{\hat{y}} = X\underline{\hat{\beta}} = X\left(X^TX\right)^{-1}X^T\underline{y} = H\underline{y}, \tag{2.9}$$

donde

$$H = X\left(X^TX\right)^{-1}X^T, \tag{2.10}$$

es una matriz de $n \times n$ usualmente llamada la matriz sombrero y juega un papel importante ya que mapea el vector de valores observados al vector de valores ajustados.

Teorema 2.1.2 *Sea H definida como en la expresión (2.10), entonces H es simétrica e idempotente.*

Demostración:

Primero, obsérvese que:

$$H^T = \left(X(X^TX)^{-1}X^T\right)^T = X(X^TX)^{-1}X^T = H,$$

$\therefore$ H es simétrica.

Luego,

$$\begin{aligned} H^2 = H \cdot H &= \left(X(X^TX)^{-1}X^T\right)\left(X(X^TX)^{-1}X^T\right) \\ &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}IX^T \\ &= X(X^TX)^{-1}X^T \\ &= H. \end{aligned}$$

$\therefore$ H es idempotente.

■

Por último, se define el i-ésimo residual como:

$$e_i = y_i - \hat{y}_i, \tag{2.11}$$

donde y_i es el valor observado y $\hat{y}_i$ el valor ajustado como se define en ec. (2.3) ocupando los estimadores por mínimos cuadrados.

Entonces, el vector de residuales es:

$$\underline{e} = \underline{y} - \underline{\hat{y}}, \tag{2.12}$$

sin embargo, otra forma en que puede expresarse es:

$$\underline{e} = \underline{y} - X\underline{\hat{\beta}} = \underline{y} - H\underline{y} = (I - H)\underline{y}. \tag{2.13}$$

2.2. Propiedades de los estimadores

Teorema 2.2.1 *El estimador por mínimos cuadrados $\underline{\hat{\beta}}$ dado por la expresión (2.7) tiene las siguientes propiedades:*

i. es un estimador insesgado para $\underline{\beta}$,

ii. $Var(\underline{\hat{\beta}}) = \sigma^2(X^TX)^{-1}$.

Demostración:

i.

$$\begin{aligned}
\mathbb{E}(\underline{\hat{\beta}}) &= \mathbb{E}[(X^TX)^{-1}X^T\underline{y}] \\
&= \mathbb{E}[(X^TX)^{-1}X^T(X\underline{\beta} + \underline{\varepsilon})] \\
&= \mathbb{E}[(X^TX)^{-1}X^TX\underline{\beta} + (X^TX)^{-1}X^T\underline{\varepsilon}] \\
&= \mathbb{E}[\underline{\beta} + (X^TX)^{-1}X^T\underline{\varepsilon}] \\
&= \underline{\beta} + (X^TX)^{-1}X^T\,\mathbb{E}(\underline{\varepsilon}) \\
&= \underline{\beta}.
\end{aligned}$$

Ya que por el supuesto (II), $\mathbb{E}(\varepsilon) = 0$ y además $(X^TX)^{-1}X^TX = I$. Por lo tanto, $\underline{\hat{\beta}}$ es un estimador insesgado para $\underline{\beta}$ si el modelo es correcto.

ii. La propiedad de la varianza de $\underline{\hat{\beta}}$ se expresa por la matriz de varianzas y covarianzas

$$Cov(\underline{\hat{\beta}}) = \mathbb{E}\left(\left[\underline{\hat{\beta}} - \mathbb{E}(\underline{\hat{\beta}})\right]\left[\underline{\hat{\beta}} - \mathbb{E}(\underline{\hat{\beta}})\right]^T\right).$$

Esta matriz se obtiene al aplicar el operador varianza a $\underline{\hat{\beta}}$:

$$Cov(\underline{\hat{\beta}}) = Var(\underline{\hat{\beta}}) = Var\left[(X^TX)^{-1}X^T\underline{y}\right],$$

dado que $(X^TX)^{-1}X^T$ es una matriz de constantes y que la varianza de $\underline{y}$ es $\sigma^2 I$ por el supuesto (IV), se tiene

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

■

Obsérvese que si $C = (X^TX)^{-1}$, la varianza de $\hat{\beta}_i$ es $\sigma^2 c_{ii}$ y la covarianza entre $\hat{\beta}_i$ y $\hat{\beta}_j$ es $\sigma^2 c_{ij}$, donde c_{ij} es el elemento i, j de la matriz C.

Pero el estimador $\underline{\hat{\beta}}$ por mínimos cuadrados es, además de insesgado, el de mínima varianza. Esta característica es la que lo hace uno de los estimadores más comunes e importantes. Esto se enuncia y prueba en el siguiente teorema:

Teorema 2.2.2 *Bajo los supuestos descritos en la sección 1.1.2, el estimador por mínimos cuadrados para $\underline{\beta}$, $\underline{\hat{\beta}}$ definido en la expresión (2.7) es el mejor estimador lineal insesgado (BLUE, por las siglas en inglés de best linear unbiased estimator). Donde "mejor" significa que $\hat{\beta}_i$ tiene la mínima varianza de entre todos los estimadores lineales insesgados.*

Demostración:

Primero, obsérvese que $\underline{\hat{\beta}} = C\underline{y}$ con $C = (X^TX)^{-1}X^T$, donde C es una matriz de $p \times n$ de tal forma que $\hat{\beta}_{i-1} = c_i\underline{y}$ donde $\hat{\beta}_{i-1}$ es el estimador del parámetro β_{i-1} [2], c_i es el i-ésimo renglón de C y $\underline{y}$ es el vector de valores observados. Entonces:

$$\hat{\beta}_{i-1} = c_i\underline{y} = c_{i1}y_i + c_{i2}y_2 + \cdots + c_{in}y_n,$$

$\therefore$ $\underline{\hat{\beta}}$ es un estimador lineal pues cada uno de sus componentes son combinaciones lineales de los datos observados.

Luego, por el Teorema 2.2.1 se sabe que $\underline{\hat{\beta}}$ es un estimador insesgado para $\underline{\beta}$.

Y, por último, queda probar que $\underline{\hat{\beta}}$ tiene la mínima varianza. Para ello se propone otro estimador con las mismas propiedades y se prueba que su varianza no es más pequeña que la de $\underline{\hat{\beta}}$. Sea $\underline{\tilde{\beta}} = K\underline{y}$ otro estimador lineal insesgado para $\underline{\beta}$ con $K \neq (X^TX)^{-1}X^T$, es decir $\underline{\tilde{\beta}} \neq \underline{\hat{\beta}}$, sin embargo, se puede proponer una matriz $E \neq 0$ de $n \times p$ tal que $K = (X^TX)^{-1}X^T + D$. Se calcula

$Var(\underline{\tilde{\beta}}) = Var(K\underline{y})$

[2] Por notación se resta 1 al subíndice pues los parámetros son $\beta_0, \beta_1, \ldots, \beta_k$.

$$
\begin{aligned}
&= K\ Var(\underline{y})\ K^T \\
&= \sigma^2 K K^T \\
&= \sigma^2 \left((X^TX)^{-1}X^T + D\right)\left(X(X^TX)^{-1} + D^T\right) \\
&= \sigma^2 \left((X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DX(X^TX)^{-1} + DD^T\right) \\
&= \sigma^2 \left((X^TX)^{-1} + (X^TX)^{-1}(DX)^T + DX(X^TX)^{-1} + DD^T\right) \\
&= \sigma^2 \left((X^TX)^{-1} + DD^T\right) \\
&= \sigma^2(X^TX)^{-1} + \sigma^2 DD^T \\
&= Var(\underline{\hat{\beta}}) + \sigma^2 DD^T.
\end{aligned}
$$

La antepenúltima línea se da pues $DX = 0$, por condición de insesgamiento.

Se define la matriz $\tilde{D} = DD^T$ cuyos elementos de la diagonal principal son siempre positivos, entonces se tiene

$$Var(\tilde{\beta}_i) = Var(\hat{\beta}_i) + \sigma^2 \tilde{d}_{ii},$$

donde $\tilde{d}_{ii}$ es el i-ésimo elemento de la diagonal de la matriz $\tilde{D}$, el cual se sabe es positivo. Por lo tanto,

$$Var(\tilde{\beta}_i) > Var(\hat{\beta}_i) \qquad \forall\, i \in \{0, \ldots, k\}.$$

■

2.3. Estimadores con restricciones$(A\underline{\beta} = \underline{c})$

A veces, se tienen ciertas restricciones *a priori* sobre los parámetros desconocidos que deben considerarse al momento de estimarlos. Un ejemplo simple podría ser un modelo que sus parámetros β_i involucren los tres ángulos de un triángulo, la restricción sería que deben sumar 180°; o uno que involucren un peso total y sus componentes, tales como grasa, hueso, músculo y magro en una canal de res preparada, la restricción sería que la suma de los componentes debe ser el peso total [12]. Para ello se ocupa otro estimador distinto al de mínimos cuadrados[2].

2.3.1. Estimadores

Un conjunto de j restricciones en el vector $\underline{\beta}$ puede ser escrito como $A\underline{\beta} = \underline{c}$ donde A es una matriz de $j \times p$ de rango j y $\underline{c}$ es un vector de $j \times 1$.

Combinando estas restricciones con los datos muestrales, el método que se ocupa es el de los Multiplicadores de Lagrange, donde cada restricción es $a_i\underline{\beta} = c_i$ con $i = \{1, 2, \ldots, j\}$ con a_i el renglón i de la matriz A y c_i el i-ésimo elemento del vector $\underline{c}$.

Entonces, se busca minimizar $S(\underline{\beta})$, definida en la expresión (2.4) sujeta a la restricción $A\underline{\beta} = \underline{c}$. Como primer paso, al igualar las j restricciones a 0 queda $A\underline{\beta} - \underline{c} = 0$, por conveniencia para los cálculos se ocupa $2(A\underline{\beta} - \underline{c}) = 0$ que es una expresión equivalente.

Luego,

$$\sum_{i=1}^{j} \lambda_i \cdot 2(A\underline{\beta} - \underline{c}) = 2\underline{\lambda}^T(A\underline{\beta} - \underline{c}),$$

así que el lagrangiano es la función

$$\mathcal{L}(\underline{\beta}, \underline{\lambda}) = (\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) + 2\underline{\lambda}^T(A\underline{\beta} - \underline{c}),$$

considerando la simplificación hecha en la expresión (2.6) queda

$$\mathcal{L}(\underline{\beta}, \underline{\lambda}) = \underline{y}^T\underline{y} - 2\underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta} + 2\underline{\lambda}^T A\underline{\beta} - 2\underline{\lambda}^T\underline{c}. \tag{2.14}$$

Sean $\hat{\underline{\beta}}_R$ y $\hat{\underline{\lambda}}$ los valores que minimizan el lagrangiano (2.14), entonces satisfacen

$$\begin{cases} \frac{\partial}{\partial \underline{\beta}}\mathcal{L}(\underline{\beta}, \underline{\lambda})\Big|_{\underline{\beta}=\hat{\underline{\beta}}_R, \underline{\lambda}=\hat{\underline{\lambda}}} & = 0 \\ \frac{\partial}{\partial \underline{\lambda}}\mathcal{L}(\underline{\beta}, \underline{\lambda})\Big|_{\underline{\beta}=\hat{\underline{\beta}}_R, \underline{\lambda}=\hat{\underline{\lambda}}} & = 0. \end{cases}$$

Resolviendo las derivadas parciales y evaluando resulta el siguiente sistema de ecuaciones

$$\begin{cases} \frac{\partial}{\partial \underline{\beta}}\mathcal{L}(\underline{\beta}, \underline{\lambda})\Big|_{\underline{\beta}=\hat{\underline{\beta}}_R, \underline{\lambda}=\hat{\underline{\lambda}}} & = -2X^T\underline{y} + 2X^T X\hat{\underline{\beta}}_R + 2A^T\hat{\underline{\lambda}} = 0 \\ \frac{\partial}{\partial \underline{\lambda}}\mathcal{L}(\underline{\beta}, \underline{\lambda})\Big|_{\underline{\beta}=\hat{\underline{\beta}}_R, \underline{\lambda}=\hat{\underline{\lambda}}} & = 2A\hat{\underline{\beta}}_R - 2\underline{c} = 0. \end{cases}$$

Eliminando el factor 2 y acomodando las ecuaciones, el sistema queda de la siguiente forma:

$$\begin{bmatrix} X^T X & A^T \\ A & 0_{j\times j} \end{bmatrix} \cdot \begin{bmatrix} \hat{\underline{\beta}}_R \\ \hat{\underline{\lambda}} \end{bmatrix} = \begin{bmatrix} X^T\underline{y} \\ \underline{c} \end{bmatrix},$$

donde $0_{j\times j}$ es la matriz de ceros de dimensión $j \times j$. Para resolver el sistema por Gauss-Jordan se plantea la siguiente matriz

$$\left(\begin{array}{cc|c} X^T X & A^T & X^T\underline{y} \\ A & 0_{j\times j} & \underline{c} \end{array} \right),$$

se multiplica el primer renglón por $(X^TX)^{-1}$ por la izquierda ($R_1 \to (X^TX)^{-1}R_1$). Por el Teorema A.0.4 se sabe que $(X^TX)^{-1}$ existe, nótese también que las dimensiones de las matrices sí permiten el producto pues $(X^TX)^{-1}$ es de $p \times p$, A^T es de $p \times j$ y $X^T\underline{y}$ es de $p \times 1$. Entonces la matriz queda[3]

$$\left(\begin{array}{cc|c} I_{p\times p} & (X^TX)^{-1}A^T & (X^TX)^{-1}X^T\underline{y} \\ A & 0_{j\times j} & \underline{c} \end{array} \right),$$

restando al segundo renglón el producto de A por el primer renglón ($R_2 \to R_2 - AR_1$) resulta

$$\left(\begin{array}{cc|c} I_{p\times p} & (X^TX)^{-1}A^T & (X^TX)^{-1}X^T\underline{y} \\ 0_{j\times p} & -A(X^TX)^{-1}A^T & \underline{c} - A(X^TX)^{-1}X^T\underline{y} \end{array} \right),$$

dado que $(X^TX)^{-1}$ es d.p. (por los Teoremas A.0.4 y A.0.1), entonces $A(X^TX)^{-1}A^T$ también es definida positiva y, por lo tanto, invertible (por el Teorema A.0.3). Entonces, multiplicando al segundo renglón por $-\left\{A(X^TX)^{-1}A^T\right\}^{-1}$ ($R_2 \to -\left\{A(X^TX)^{-1}A^T\right\}^{-1}R_2$) se tiene

$$\left(\begin{array}{cc|c} I_{p\times p} & (X^TX)^{-1}A^T & (X^TX)^{-1}X^T\underline{y} \\ 0_{j\times p} & I_{j\times j} & \left\{A(X^TX)^{-1}A^T\right\}^{-1}\left(A(X^TX)^{-1}X^T\underline{y} - \underline{c}\right) \end{array} \right),$$

por último, se resta al primer renglón el producto de $(X^TX)^{-1}A^T$ por el segundo renglón ($R_1 \to R_1 - (X^TX)^{-1}A^TR_2$) y sustituyendo $\underline{\hat{\beta}}$ que es el estimador por mínimos cuadrados sin restricciones dado en la expresión (2.7), queda

$$\left(\begin{array}{cc|c} I_{p\times p} & 0_{p\times j} & \underline{\hat{\beta}} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\underline{\hat{\beta}} - \underline{c}) \\ 0_{j\times p} & I_{j\times j} & \left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\underline{\hat{\beta}} - \underline{c}) \end{array} \right),$$

$$\therefore \qquad \underline{\hat{\beta}}_R = \underline{\hat{\beta}} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\underline{\hat{\beta}} - \underline{c}) \tag{2.15}$$

$$\underline{\hat{\lambda}} = \left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\underline{\hat{\beta}} - \underline{c}), \tag{2.16}$$

o de forma equivalente

$$\underline{\hat{\beta}}_R = \underline{\hat{\beta}} - (X^TX)^{-1}A^T\underline{\hat{\lambda}}. \tag{2.17}$$

Para probar que $\underline{\hat{\beta}}_R$ realmente minimiza $S(\underline{\beta})$ sujeto a $A\underline{\beta} = \underline{c}$, nótese que

$$\begin{aligned} (\underline{\hat{\beta}} - \underline{\beta})^TX^TX(\underline{\hat{\beta}} - \underline{\beta}) &= (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R + \underline{\hat{\beta}}_R - \underline{\beta})^TX^TX(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R + \underline{\hat{\beta}}_R - \underline{\beta}) \\ &= \left[(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T + (\underline{\hat{\beta}}_R - \underline{\beta})^T\right]X^TX\left[(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R) + (\underline{\hat{\beta}}_R - \underline{\beta})\right] \\ &= \left[(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^TX^TX + (\underline{\hat{\beta}}_R - \underline{\beta})^TX^TX\right]\left[(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R) + (\underline{\hat{\beta}}_R - \underline{\beta})\right] \end{aligned}$$

[3]En las siguientes operaciones por renglón no se mencionará explícitamente que las dimensiones de las matrices sí permiten los productos, sin embargo, se invita al lector a verificarlo.

$$
\begin{aligned}
&= (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R) + (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}) \\
&\quad + (\underline{\hat{\beta}}_R - \underline{\beta})^T X^T X (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R) + (\underline{\hat{\beta}}_R - \underline{\beta})^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}) \\
&= (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R) + (\underline{\hat{\beta}}_R - \underline{\beta})^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}) \\
&\quad + 2(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}), \qquad (2.18)
\end{aligned}
$$

la última igualdad se da pues $(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta})$ es una matriz de 1×1, o un escalar y, por lo tanto, es igual a su transpuesta, esto es

$$
(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}) = \left\{ (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}) \right\}^T = (\underline{\hat{\beta}}_R - \underline{\beta})^T X^T X (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R),
$$

así que se pueden reducir los términos.

Ahora bien, trabajando en el tercer sumando de la expresión (2.18) y sustituyendo $\underline{\hat{\beta}}_R$ por (2.17) resulta

$$
\begin{aligned}
2(\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}) &= 2 \left[\underline{\hat{\beta}} - \underline{\hat{\beta}} + (X^T X)^{-1} A^T \underline{\hat{\lambda}} \right]^T X^T X \left[\underline{\hat{\beta}} - (X^T X)^{-1} A^T \underline{\hat{\lambda}} - \underline{\beta} \right] \\
&= 2\underline{\hat{\lambda}}^T A (X^T X)^{-1} X^T X \left[\underline{\hat{\beta}} - (X^T X)^{-1} A^T \underline{\hat{\lambda}} - \underline{\beta} \right] \\
&= 2\underline{\hat{\lambda}}^T A I \left[\underline{\hat{\beta}} - (X^T X)^{-1} A^T \underline{\hat{\lambda}} - \underline{\beta} \right] \\
&= 2\underline{\hat{\lambda}}^T \left[A\underline{\hat{\beta}} - A(X^T X)^{-1} A^T \underline{\hat{\lambda}} - A\underline{\beta} \right] \quad \textit{sustituyendo } \underline{\hat{\lambda}} \textit{ con (2.16)} \\
&= 2\underline{\hat{\lambda}}^T \left[A\underline{\hat{\beta}} - A(X^T X)^{-1} A^T \left\{ A(X^T X)^{-1} A^T \right\}^{-1} (A\underline{\hat{\beta}} - \underline{c}) - A\underline{\beta} \right] \\
&= 2\underline{\hat{\lambda}}^T \left[A\underline{\hat{\beta}} - I(A\underline{\hat{\beta}} - \underline{c}) - A\underline{\beta} \right] \qquad \textit{por restricción } A\underline{\beta} = \underline{c} \\
&= 2\underline{\hat{\lambda}}^T \left[A\underline{\hat{\beta}} - A\underline{\hat{\beta}} + \underline{c} - \underline{c} \right] \\
&= 0,
\end{aligned}
$$

entonces, la expresión (2.18) queda

$$
(\underline{\hat{\beta}} - \underline{\beta})^T X^T X (\underline{\hat{\beta}} - \underline{\beta}) = (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R) + (\underline{\hat{\beta}}_R - \underline{\beta})^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}). \qquad (2.19)
$$

Así que, retomando la expresión (2.8) y considerando (2.19), $S(\underline{\beta})$ se puede expresar como

$$
\begin{aligned}
S(\underline{\beta}) &= (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}) \\
&= (\underline{y} - X\underline{\hat{\beta}})^T (\underline{y} - X\underline{\hat{\beta}}) + (\underline{\hat{\beta}} - \underline{\beta})^T X^T X (\underline{\hat{\beta}} - \underline{\beta}) \\
&= (\underline{y} - X\underline{\hat{\beta}})^T (\underline{y} - X\underline{\hat{\beta}}) + (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R)^T X^T X (\underline{\hat{\beta}} - \underline{\hat{\beta}}_R) + (\underline{\hat{\beta}}_R - \underline{\beta})^T X^T X (\underline{\hat{\beta}}_R - \underline{\beta}),
\end{aligned}
$$

y alcanza su mínimo cuando $\underline{\beta} = \underline{\hat{\beta}}_R$.

Por último, es importante comentar lo siguiente: nótese que la expresión (2.15) es de la forma $\underline{\hat{\beta}}_R = \underline{\hat{\beta}} + ajuste$ donde el *ajuste* es un valor que depende de los datos observados

X, del estimador por mínimos cuadrados sin restricción $\hat{\underline{\beta}}$ y de los valores de la restricción A y $\underline{c}$. En otras palabras, $\hat{\underline{\beta}}_R$ es una actualización de $\hat{\underline{\beta}}$ con la información proporcionada por las restricciones.

2.3.2. Propiedades

Teorema 2.3.1 *Considerando la notación matricial descrita en la sección 1.1.1 y bajo los supuestos 1.1.2, si al modelo* $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ *se le agregan* j *restricciones de la forma* $A\underline{\beta} = \underline{c}$ *con* A *una matriz de* $j \times p$ *de rango* j *y* $\underline{c}$ *un vector de* $j \times 1$*. Entonces, el estimador por mínimos cuadrados restringidos* $\hat{\underline{\beta}}_R$ *dado por la expresión (2.15) tiene las siguientes propiedades:*

i. Restricción exacta, es decir, $A\hat{\underline{\beta}}_R = \underline{c}$.

ii. Es insesgado, esto es $\mathbb{E}(\hat{\underline{\beta}}_R) = \underline{\beta}$.

iii. Es eficiente, es decir, $Var(\hat{\underline{\beta}}_{Ri}) \leq Var(\hat{\underline{\beta}}_i)$.

Demostración:

i.

$$\begin{aligned} A\hat{\underline{\beta}}_R &= A\left[\hat{\underline{\beta}} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\hat{\underline{\beta}} - \underline{c})\right] \\ &= A\hat{\underline{\beta}} - A(X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\hat{\underline{\beta}} - \underline{c}) \\ &= A\hat{\underline{\beta}} - I(A\hat{\underline{\beta}} - \underline{c}) \\ &= A\hat{\underline{\beta}} - A\hat{\underline{\beta}} + \underline{c} \\ &= \underline{c}. \end{aligned}$$

ii.

$$\begin{aligned} \mathbb{E}\left(\hat{\underline{\beta}}_R\right) &= \mathbb{E}\left[\hat{\underline{\beta}} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\hat{\underline{\beta}} - \underline{c})\right] \\ &= \mathbb{E}(\hat{\underline{\beta}}) - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\,\mathbb{E}(\hat{\underline{\beta}}) - \underline{c}) \\ &= \underline{\beta} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\underline{\beta} - \underline{c}) \\ &= \underline{\beta} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(\underline{c} - \underline{c}) \\ &= \underline{\beta}, \end{aligned}$$

dado que $\mathbb{E}(\hat{\underline{\beta}}) = \underline{\beta}$ por Teorema 2.2.1 y $A\underline{\beta} = \underline{c}$ por hipótesis.

iii. Primero,

$$\begin{aligned}\underline{\hat{\beta}}_R &= \left[\underline{\hat{\beta}} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}(A\underline{\hat{\beta}} - \underline{c})\right] \\ &= \left[I - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A\right]\underline{\hat{\beta}} + M \\ &= F\underline{\hat{\beta}} + M,\end{aligned}$$

donde $F = \left[I - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A\right]$ y $M = (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}\underline{c}$. Entonces

$$\begin{aligned}Var\left(\underline{\hat{\beta}}_R\right) &= Var(F\underline{\hat{\beta}} + M) \\ &= F\ Var(\underline{\hat{\beta}})F^T \\ &= F(\sigma^2(X^TX)^{-1})F^T \\ &= \sigma^2\left[I - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A\right](X^TX)^{-1} \\ &\qquad \left[I - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A\right]^T \\ &= \sigma^2\left[I(X^TX)^{-1} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}\right] \\ &\qquad \left[I - A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}\right] \\ &= \sigma^2\Big[(X^TX)^{-1}I \\ &\quad - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1} \\ &\quad - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}I \\ &\quad + (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}\Big] \\ &= \sigma^2\Big[(X^TX)^{-1} \\ &\quad - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1} \\ &\quad - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1} \\ &\quad + (X^TX)^{-1}A^TI\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}\Big] \\ &= \sigma^2\left[(X^TX)^{-1} - (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}\right] \\ &= Var(\underline{\hat{\beta}}) - \sigma^2(X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1} \\ &= Var(\underline{\hat{\beta}}) - \sigma^2G,\end{aligned}$$

con $G = (X^TX)^{-1}A^T\left\{A(X^TX)^{-1}A^T\right\}^{-1}A(X^TX)^{-1}$. Entonces

$$Var(\underline{\hat{\beta}}) - Var(\underline{\hat{\beta}}_R) = \sigma^2G.$$

Se sabe que $(X^TX)^{-1}$ es d.p. (por los Teoremas A.0.4 y A.0.1), entonces $A(X^TX)^{-1}A^T$ también es d.p., invertible y su inversa $\left\{A(X^TX)^{-1}A^T\right\}^{-1}$ también es d.p. (por los Teo-

remas A.0.3 y A.0.1) y, aplicando repetidamente el Teorema A.0.3, primero con $C = A^T$ y luego con $C = (X^T X)^{-1}$, se tiene que G es d.p. En consecuencia

$$Var(\underline{\hat{\beta}}_i) \geq Var(\underline{\hat{\beta}}_{Ri}) \qquad \forall\, i = \{1, 2, \ldots, p\}.$$

■

Algunos comentarios acerca del estimador $\underline{\hat{\beta}}_R$ son:

- Es insesgado solo si la restricción es cierta ya que, como se puede ver en la segunda parte de la demostración del Teorema 2.3.1, de ser falsa habría un sesgo.
- Que $\underline{\hat{\beta}}_R$ tenga menor varianza que $\underline{\hat{\beta}}$ indica que tener información adicional (en este caso la restricción) hace más eficiente la estimación. Sin embargo, se puede probar que a pesar de que la restricción no fuera cierta el estimador restringido aún tendría menor varianza que el ordinario, aunque eso provocaría un sesgo como se menciona en el punto anterior.

Capítulo 3

Estimación por Máxima Verosimilitud suponiendo normalidad

Hasta ahora, los supuestos que se han asumido sobre el modelo lineal son los definidos en la sección 1.1.2, sin embargo, no se asignó ninguna distribución de probabilidad para los errores aleatorios. Esto, aunque permitió encontrar estimadores que minimizaran el error, no permite hacer inferencia estadística.

En este capítulo se desarrollan los cálculos que permiten hacer conclusiones o inferencias estadísticas sobre el modelo de regresión lineal múltiple.

3.1. Supuestos

Recordando, en los cinco supuestos dados en 1.1.2 para el modelo de regresión lineal (1.1) aunque sí se asume a los errores como un vector aleatorio no se define ninguna distribución de probabilidad en particular. Ahora bien, si ésta se define como la distribución normal multivariada (Definición 1.2.1), es decir, $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I_n)$, entonces se puede aplicar la teoría estadística que enriquezca al modelo de regresión.

Algunas consecuencias de definir $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I_n)$ son:

(i) $\varepsilon_i \sim N(0, \sigma^2)\ \forall\ i = \{1, 2, \ldots, n\}$, es decir, cada uno de los errores sigue una distribución normal con media 0 y varianza σ^2. Esto se concluye aplicando el Teorema 1.2.5 en donde la primera partición es únicamente el elemento ε_i.

Nótese que esto armoniza con el supuesto (II).

(ii) ε_i es independiente de ε_j. Esto se concluye aplicando dos veces el Teorema 1.2.6, una primera vez estableciendo a $(\varepsilon_i \quad \varepsilon_j)^T$ como la primera partición, y luego haciendo una nueva partición con el elemento ε_i.

Además, esto permite concluir que $Cov(\varepsilon_i, \varepsilon_j) = 0$, lo que concuerda con el supuesto (III).

(iii) Como $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, entonces $\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I_n)$ por el Teorema 1.2.4.
Lo que satisface el supuesto (IV).

En conclusión, definir la distribución conjunta de los errores como la normal multivariada satisface todos los supuestos que se asumen en el modelo de regresión lineal.

3.2. Estimadores

En la sección 2.1 se encontró una expresión para estimar el parámetro $\underline{\beta}$, este estimador encontrado, $\underline{\hat{\beta}}$ (expresión (2.7)), tiene la característica de minimizar el error, es decir, minimiza la distancia entre los valores reales observados $\underline{y}$ y los estimados $\underline{\hat{y}}$ (expresión (2.9)).

Ahora, sabiendo que $\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I_n)$, se pueden encontrar estimadores para los parámetros desconocidos, $\underline{\beta}$ y σ^2, que maximicen la probabilidad de tener los datos observados $\underline{y}$, es decir, $P(\underline{Y} = \underline{y})$. A estos estimadores se les conoce como *Máximos Verosímiles* (M.V.) y se denotan como $\underline{\hat{\beta}}_{MV}$ y $\hat{\sigma}^2_{MV}$. El análisis para encontrarlos es el siguiente:

Considerando el modelo (1.1) y bajo los supuestos 3.1 la función de verosimilitud es la función de probabilidad conjunta de los errores $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, esta es:

$$L(\underline{\epsilon}, \underline{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2} det(\sigma^2 I_n)^{1/2}} exp\Big\{ -\frac{1}{2}\underline{\varepsilon}^T (\sigma^2 I_n)^{-1} \underline{\varepsilon} \Big\},$$

pero considerando el modelo en su forma matricial (1.3) se tiene que $\underline{\varepsilon} = \underline{y} - X\underline{\beta}$, además $det(\sigma^2 I_n) = \sigma^{2n}$ y $(\sigma^2 I_n)^{-1} = \frac{1}{\sigma^2} I_n$. Entonces, la función de verosimilitud queda:

$$L(\underline{\beta}, \sigma^2; \underline{y}, X) = (2\pi\sigma^2)^{-n/2} exp\Big\{ -\frac{1}{2\sigma^2}(\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}) \Big\}, \tag{3.1}$$

esta función es la densidad conjunta de $\underline{Y}$ pues $\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I_n)$ por el supuesto (VI).

En la práctica, es más conveniente maximizar el logaritmo de la función de verosimilitud (comúnmente llamada *log verosimilitud*). Debido a que el logaritmo es una función monótonamente creciente, maximizar el logaritmo de una función equivale a maximizar la función misma, pues la log verosimilitud tiene la misma relación de orden que la función de verosimilitud:

$$P(\underline{Y} = \underline{y} | X, \underline{\hat{\beta}}_{MV}, \hat{\sigma}^2_{MV}) \geq P(\underline{Y} = \underline{y} | X, \underline{\beta}, \sigma^2)$$
$$\Longleftrightarrow$$

$$ln\left(P(\underline{Y}=\underline{y}|X,\underline{\hat{\beta}}_{MV},\hat{\sigma}^2_{MV})\right) \geq ln\left(P(\underline{Y}=\underline{y}|X,\underline{\beta},\sigma^2)\right).$$

Tomar el logaritmo no solo simplifica el análisis matemático posterior, sino que también ayuda numéricamente porque el producto de un gran número de pequeñas cantidades puede fácilmente sobrepasar la precisión numérica de la computadora y esto se resuelve calculando, en su lugar, la suma de los logaritmos de las probabilidades.

Entonces,

$$ln\ L(\underline{y},X,\underline{\beta},\sigma^2) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta}), \tag{3.2}$$

se conoce como la función log verosimilitud.

Teorema 3.2.1 *Sean $X \in \mathbb{R}^{n\times p}$ de rango p con $p < n$ y $\underline{y} \in \mathbb{R}^{n\times 1}$ una matriz y un vector de valores observados, respectivamente. Entonces, la función (3.2) se maximiza en $(\underline{\hat{\beta}}_{MV},\hat{\sigma}^2_{MV}) = ((X^TX)^{-1}X^T\underline{y},(\underline{y}-\underline{\hat{y}})^T(\underline{y}-\underline{\hat{y}})/n)$.*

Demostración:

Si $\underline{\hat{\beta}}_{MV}$ y $\hat{\sigma}^2_{MV}$ maximizan $ln\ L$, entonces satisfacen el siguiente sistema de ecuaciones:

$$\begin{cases} \frac{\partial}{\partial\underline{\beta}} ln\ L\Big|_{\underline{\beta}=\underline{\hat{\beta}}_{MV},\sigma^2=\hat{\sigma}^2_{MV}} = 0 \\ \\ \frac{\partial}{\partial\sigma^2} ln\ L\Big|_{\underline{\beta}=\underline{\hat{\beta}}_{MV},\sigma^2=\hat{\sigma}^2_{MV}} = 0, \end{cases} \tag{3.3}$$

entonces

$$\frac{\partial}{\partial\underline{\beta}} ln\ L\Big|_{\underline{\beta}=\underline{\hat{\beta}}_{MV},\sigma^2=\hat{\sigma}^2_{MV}} = -\frac{1}{2\hat{\sigma}^2_{MV}}(-2X^T\underline{y}+2X^TX\underline{\hat{\beta}}_{MV}) = 0 \tag{3.4}$$

$$\frac{\partial}{\partial\sigma^2} ln\ L\Big|_{\underline{\beta}=\underline{\hat{\beta}}_{MV},\sigma^2=\hat{\sigma}^2_{MV}} = -\frac{n}{2\hat{\sigma}^2_{MV}} + \frac{1}{2\hat{\sigma}^4_{MV}}(\underline{y}-X\underline{\hat{\beta}}_{MV})^T(\underline{y}-X\underline{\hat{\beta}}_{MV}) = 0, \tag{3.5}$$

trabajando la ecuación (3.4) se tiene

$$\begin{aligned} -\frac{1}{2\hat{\sigma}^2_{MV}}(-2X^T\underline{y}+2X^TX\underline{\hat{\beta}}_{MV}) &= 0 \\ -X^T\underline{y}+X^TX\underline{\hat{\beta}}_{MV} &= 0 \\ X^TX\underline{\hat{\beta}}_{MV} &= X^T\underline{y} \end{aligned}$$

$$\underline{\hat{\beta}}_{MV} = (X^T X)^{-1} X^T \underline{y}.$$

Resulta que el estimador M.V. para $\underline{\beta}$ es el mismo que por mínimos cuadrados

$$\underline{\hat{\beta}}_{MV} = \underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{y}, \tag{3.6}$$

lo que claramente maximiza la función log verosimilitud (3.2) pues $(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})$ es la suma de cuadrados del error y, por lo tanto, es mayor o igual a cero, así que el tercer término de (3.2) siempre restará, más aún, $(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})$ es la función $S(\underline{\beta})$ dada por (2.4) y por el Teorema 2.1.1, $\underline{\hat{\beta}}_{MV} = \underline{\hat{\beta}}$ minimiza S, así que

$$ln\ L(\underline{y}, X, \underline{\beta}, \sigma^2) \leq ln\ L(\underline{y}, X, \underline{\hat{\beta}}_{MV}, \sigma^2) \qquad \forall\ \sigma^2 > 0. \tag{3.7}$$

Luego, trabajando con la ecuación (3.5) se obtiene

$$\begin{aligned}
-\frac{n}{2\hat{\sigma}^2_{MV}} + \frac{1}{2\hat{\sigma}^4_{MV}}(\underline{y} - X\underline{\hat{\beta}}_{MV})^T(\underline{y} - X\underline{\hat{\beta}}_{MV}) &= 0 \\
\frac{1}{2\hat{\sigma}^4_{MV}}\left(-n\hat{\sigma}^2_{MV} + (\underline{y} - X\underline{\hat{\beta}}_{MV})^T(\underline{y} - X\underline{\hat{\beta}}_{MV})\right) &= 0 \\
-n\hat{\sigma}^2_{MV} + (\underline{y} - X\underline{\hat{\beta}}_{MV})^T(\underline{y} - X\underline{\hat{\beta}}_{MV}) &= 0
\end{aligned}$$

$$\therefore \qquad \hat{\sigma}^2_{MV} = \frac{(\underline{y} - X\underline{\hat{\beta}}_{MV})^T(\underline{y} - X\underline{\hat{\beta}}_{MV})}{n}. \tag{3.8}$$

Por último, para probar que el punto $(\underline{\hat{\beta}}_{MV}, \hat{\sigma}^2_{MV})$ efectivamente maximiza $ln\ L$ nótese que

$$\begin{aligned}
&ln\ L(\underline{y}, X, \underline{\hat{\beta}}_{MV}, \hat{\sigma^2}_{MV}) - ln\ L(\underline{y}, X, \underline{\hat{\beta}}_{MV}, \sigma^2) \\
&\quad = \left(-\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\hat{\sigma^2}_{MV}) - \frac{1}{2\hat{\sigma^2}_{MV}}(\underline{y} - X\underline{\hat{\beta}}_{MV})^T(\underline{y} - X\underline{\hat{\beta}}_{MV})\right) \\
&\qquad - \left(-\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(\underline{y} - X\underline{\hat{\beta}}_{MV})^T(\underline{y} - X\underline{\hat{\beta}}_{MV})\right) \\
&\quad = -\frac{n}{2}ln(\hat{\sigma^2}_{MV}) - \frac{n}{2} + \frac{n}{2}ln(\sigma^2) + \frac{n\hat{\sigma^2}_{MV}}{2\sigma^2} \qquad (3.9) \\
&\quad = -\frac{n}{2}\left(ln(\hat{\sigma^2}_{MV}) + 1 - ln(\sigma^2) - \frac{\hat{\sigma^2}_{MV}}{\sigma^2}\right) \\
&\quad = -\frac{n}{2}\left(ln(\frac{\hat{\sigma^2}_{MV}}{\sigma^2}) + 1 - \frac{\hat{\sigma^2}_{MV}}{\sigma^2}\right) \\
&\quad \geq 0, \qquad (3.10)
\end{aligned}$$

donde el renglón (3.9) se da por reducción algebraica y por sustituir $\hat{\sigma^2}_{MV}$ por (3.8), y la desigualdad (3.10) se da ya que se sabe que $v \leq e^{v-1}$ y, por lo tanto, $ln\ v \leq v-1\ \forall\ v \geq 0$ (con la igualdad en $v=1$). Haciendo $v = \hat{\sigma^2}_{MV}/\sigma^2$ se deduce el resultado.

Considerando lo anterior y la expresión (3.7), se tiene

$$ln\ L(\underline{y}, X, \underline{\beta}, \sigma^2) \leq ln\ L(\underline{y}, X, \underline{\hat{\beta}}_{MV}, \hat{\sigma^2}) \leq ln\ L(\underline{y}, X, \underline{\hat{\beta}}_{MV}, \hat{\sigma^2}_{MV}) \quad \forall\ \sigma^2 > 0$$
$$\therefore \qquad ln\ L(\underline{y}, X, \underline{\beta}, \sigma^2) \leq ln\ L(\underline{y}, X, \underline{\hat{\beta}}_{MV}, \hat{\sigma^2}_{MV}) \qquad \forall\ \sigma^2 > 0.$$

■

Definición 3.2.1 *Sean $\underline{y}$ y X los valores observados como se definen en la sección 1.1.1, y bajo los supuestos 3.1, $\underline{\hat{\beta}}_{MV}$ y $\hat{\sigma}^2_{MV}$ dados por las expresiones (3.6) y (3.8), respectivamente, se conocen como los estimadores máximo verosímiles del modelo*

$$\mathbb{E}(\underline{Y}) = X\underline{\beta}.$$

3.3. Propiedades de los estimadores

Dado que $\underline{\hat{\beta}}_{MV} = \underline{\hat{\beta}}$, entonces el estimador $\underline{\hat{\beta}}_{MV}$ tiene las mismas propiedades descritas en la sección 2.2, que son:

- $\underline{\hat{\beta}}_{MV}$ es un estimador insesgado, es decir, $\mathbb{E}(\underline{\hat{\beta}}_{MV}) = \underline{\beta}$, y su varianza es $\sigma^2(X^TX)^{-1}$, esto es $Var(\underline{\hat{\beta}}_{MV}) = \sigma^2(X^TX)^{-1}$, por el Teorema 2.2.1.
- $\underline{\hat{\beta}}_{MV}$ es BLUE, por el Teorema 2.2.2[1].

Además, considerando (VI) $\underline{\hat{\beta}}_{MV}$ y $\hat{\sigma}^2_{MV}$ tienen otras propiedades relacionadas con su distribución y la relación entre ellos que se presentan en el siguiente teorema.

Teorema 3.3.1 *Bajo los supuestos descritos en la sección 3.1, los estimadores M.V. para $\underline{\beta}$ y σ^2, $\underline{\hat{\beta}}$ y $\hat{\sigma}^2_{MV}$ definidos en las expresiones (3.6) y (3.8), respectivamente, tienen las siguientes propiedades:*

i. $\underline{\hat{\beta}} \sim N_p(\underline{\beta}, \sigma^2(X^TX)^{-1})$.

ii. $\frac{(\underline{Y}-\underline{\hat{Y}})^T(\underline{Y}-\underline{\hat{Y}})}{\sigma^2} \sim \chi^2_{n-p}$.

[1]Dado que $\underline{\hat{\beta}}_{MV} = \underline{\hat{\beta}}$, en lo sucesivo solo se ocupará $\underline{\hat{\beta}}$.

iii. $\hat{\sigma}^2_{MV}$ *es un estimador sesgado.*

Demostración:

i. Dado que $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y} = C\underline{y}$ donde C es una matriz de $p \times n$ con rango $rango(C) = rango(X^T) = rango(X) = p$. Por el Teorema 1.2.4, $\underline{\hat{\beta}}$ se distribuye normal p-variada, con media $C\,\mathbb{E}(\underline{Y}) = (X^TX)^{-1}X^TX\underline{\beta} = \underline{\beta}$ y varianza $C(\sigma^2 I_n)C^T = (X^TX)^{-1}X^T(\sigma^2 I_n)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$, lo que concuerda con las propiedades del Teorema 2.2.1. Entonces,

$$\underline{\hat{\beta}} \sim N_p(\underline{\beta}, \sigma^2(X^TX)^{-1}).$$

ii.

$$\begin{aligned}
\frac{(\underline{Y} - \underline{\hat{Y}})^T(\underline{Y} - \underline{\hat{Y}})}{\sigma^2} &= \frac{(\underline{Y} - H\underline{Y})^T(\underline{Y} - H\underline{Y})}{\sigma^2} \qquad \text{por (2.9)} \\
&= \frac{\underline{Y}^T(I_n - H)(I_n - H)\underline{Y}}{\sigma^2} \qquad \text{por Teorema 2.1.2} \\
&= \frac{\underline{Y}^T(I_n - H)\underline{Y}}{\sigma^2} \\
&= \left(\frac{\underline{Y}}{\sigma^2}\right)^T (I_n - H)\left(\frac{\underline{Y}}{\sigma^2}\right). \qquad (3.11)
\end{aligned}$$

Por Teorema 1.2.4, $\left(\frac{\underline{Y}}{\sigma^2}\right) \sim N_n(\frac{X\underline{\beta}}{\sigma}, I_n)$, además $I_n - H$ es simétrica e idempotente, así que aplicando el Teorema 1.2.8 resulta

$$\left(\frac{\underline{Y}}{\sigma^2}\right)^T (I_n - H)\left(\frac{\underline{Y}}{\sigma^2}\right) \sim \chi^2_m(\lambda),$$

donde

$$\begin{aligned}
\lambda &= \left(\frac{X\underline{\beta}}{\sigma}\right)^T (I_n - H)\left(\frac{X\underline{\beta}}{\sigma}\right) \\
&= \frac{1}{\sigma^2}\underline{\beta}^T(X^TX - X^THX)\underline{\beta} \\
&= \frac{1}{\sigma^2}\underline{\beta}^T(X^TX - X^TX(X^TX)^{-1}X^TX)\underline{\beta} \\
&= \frac{1}{\sigma^2}\underline{\beta}^T(0)\underline{\beta} \\
&= 0,
\end{aligned}$$

y

$$m = rango(I_n - H) = traza(I_n - H) = n - p$$

$$\therefore \qquad \frac{(\underline{Y}-\underline{\hat{Y}})^T(\underline{Y}-\underline{\hat{Y}})}{\sigma^2} \sim \chi^2_{n-p}.$$

iii.

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2_{MV}) &= \mathbb{E}\left[\frac{(\underline{Y}-\underline{\hat{Y}})^T(\underline{Y}-\underline{\hat{Y}})}{n}\right] \\
&= \frac{\sigma^2}{n}\,\mathbb{E}\left[\frac{(\underline{Y}-\underline{\hat{Y}})^T(\underline{Y}-\underline{\hat{Y}})}{\sigma^2}\right] \\
&= \frac{\sigma^2}{n}(n-p) \\
&= \sigma^2\frac{n-p}{n} \\
&\neq \sigma^2,
\end{aligned}$$

pues la esperanza de una v.a. χ^2 centrada son sus grados de libertad. Por lo tanto, $\hat{\sigma}^2_{MV}$ es un estimador sesgado.

■

Del Teorema 3.3.1 se puede proponer el siguiente estimador para σ^2:

$$\hat{\sigma}^2 = \frac{n}{n-p}\hat{\sigma}^2_{MV} = \frac{(\underline{Y}-\underline{\hat{Y}})^T(\underline{Y}-\underline{\hat{Y}})}{n-p} = \frac{\underline{Y}^T(I_n-H)\underline{\hat{Y}}}{n-p}, \tag{3.12}$$

(obsérvese (3.11)) el cual sí es insesgado pues

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{n}{n-p}\hat{\sigma}^2_{MV}\right) = \frac{n}{n-p}\,\mathbb{E}\left(\hat{\sigma}^2_{MV}\right) = \frac{n}{n-p}\left(\sigma^2\frac{n-p}{n}\right) = \sigma^2.$$

Por último, se tienen las siguientes propiedades que son útiles para la inferencia estadística.

Teorema 3.3.2 *Bajo los supuestos descritos en la sección 3.1 los estimadores para $\underline{\beta}$ y σ^2, $\underline{\hat{\beta}}$ definido en (3.6) o (2.7), y $\hat{\sigma}^2$ definido en (3.12). Tienen las siguientes propiedades:*

i. $\underline{\hat{\beta}}$ es independiente de $\hat{\sigma}^2$.

ii. $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

Demostración:

i. Es equivalente probar que $\underline{\hat{\beta}}$, el vector estimador, es independiente de $\underline{e} = \underline{Y} - \underline{\hat{Y}}$, el vector de residuos.

Se tiene que $\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{Y}$ y $\underline{Y} - \underline{\hat{Y}} = (I_n - H)\underline{Y}$ (obsérvese la expresión (2.13)). Dado que se asume que $\underline{y} \sim N_n(\underline{\mu}, \Sigma)$, entonces se puede definir $\underline{\hat{\beta}} = A\underline{Y}$ y $\underline{e} = B\underline{Y}$ con $A = (X^T X)^{-1} X^T$ y $B = I_n - H$, donde

$$\begin{aligned} A(\sigma^2 I_n)B &= \sigma^2 AB \\ &= \sigma^2 (X^T X)^{-1} X^T (I_n - H) \\ &= \sigma^2 \left[(X^T X)^{-1} X^T - (X^T X)^{-1} X^T X (X^T X)^{-1} X^T \right] \\ &= \sigma^2 \left[(X^T X)^{-1} X^T - (X^T X)^{-1} X^T \right] \\ &= 0, \end{aligned}$$

entonces, por el Teorema 1.2.7 se concluye que $\underline{\hat{\beta}}$ es independiente de $\hat{\sigma}^2$.

ii.

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} = \frac{n-p}{\sigma^2} \frac{(\underline{Y} - \underline{\hat{Y}})^T (\underline{Y} - \underline{\hat{Y}})}{n-p} = \frac{(\underline{Y} - \underline{\hat{Y}})^T (\underline{Y} - \underline{\hat{Y}})}{\sigma^2},$$

que, como se demostró en el Teorema 3.3.1 (ii), se distribuye χ^2 con $n - p$ g.l.

■

3.4. Estimadores con restricciones$(A\underline{\beta} = \underline{c})$

También pueden estimarse los valores de $\underline{\beta}$ y σ^2 a través de máxima verosimilitud sujetos a la restricción $A\underline{\beta} = \underline{c}$. Como se explica en la sección 2.3.1, $A\underline{\beta} = \underline{c}$ representa un conjunto de j restricciones en los parámetros del vector $\underline{\beta}$, pues A es una matriz de dimensiones $j \times p$ y de $rango(A) = j$, y $\underline{c}$ es un vector de dimensión $j \times 1$, ambos de constantes conocidas.

Esta estimación con j restricciones también se obtiene ocupando los Multiplicadores de Lagrange. Entonces, se busca maximizar la función log verosimilitud $ln\ L(\underline{y}, X, \underline{\beta}, \sigma^2)$ dada en la expresión (3.2) sujeta a que $A\underline{\beta} = \underline{c}$. Así que el lagrangiano es la función[2]:

$$\mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}) + \underline{\lambda}^T (A\underline{\beta} - \underline{c}),$$

[2]Obsérvese en la sección 2.3.1 el desarrollo de $\underline{\lambda}^T (A\underline{\beta} - \underline{c})$.

con $\underline{\lambda}$ el vector de multiplicadores de lagrange de dimensión $j \times 1$. Considerando la simplificación hecha en la expresión (2.6) queda

$$\mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \left(\underline{y}^T \underline{y} - 2\underline{\beta}^T X^T \underline{y} + \underline{\beta}^T X^T X \underline{\beta} \right) + \underline{\lambda}^T (A\underline{\beta} - \underline{c}). \tag{3.13}$$

Sean $\hat{\underline{\beta}}_{MVR}$, $\hat{\sigma}^2_{MVR}$ y $\hat{\underline{\lambda}}_{MV}$ los valores que maximizan el lagrangiano (3.13), entonces satisfacen

$$\begin{cases} \frac{\partial}{\partial \underline{\beta}} \mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) \Big|_{\underline{\beta}=\hat{\underline{\beta}}_{MVR}, \sigma^2=\hat{\sigma}^2_{MVR}, \underline{\lambda}=\hat{\underline{\lambda}}_{MV}} & = 0 \\ \frac{\partial}{\partial \sigma^2} \mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) \Big|_{\underline{\beta}=\hat{\underline{\beta}}_{MVR}, \sigma^2=\hat{\sigma}^2_{MVR}, \underline{\lambda}=\hat{\underline{\lambda}}_{MV}} & = 0 \\ \frac{\partial}{\partial \underline{\lambda}} \mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) \Big|_{\underline{\beta}=\hat{\underline{\beta}}_{MVR}, \sigma^2=\hat{\sigma}^2_{MVR}, \underline{\lambda}=\hat{\underline{\lambda}}_{MV}} & = 0. \end{cases}$$

Resolviendo las derivadas parciales y evaluando, resulta el siguiente sistema de ecuaciones

$$\begin{cases} \frac{\partial}{\partial \underline{\beta}} \mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) \Big|_{\underline{\beta}=\hat{\underline{\beta}}_{MVR}, \sigma^2=\hat{\sigma}^2_{MVR}, \underline{\lambda}=\hat{\underline{\lambda}}_{MV}} & = -\frac{1}{2\hat{\sigma}^2_{MVR}} \left(-2X^T \underline{y} + 2X^T X \hat{\underline{\beta}}_{MVR} \right) + A^T \hat{\underline{\lambda}}_{MV} = 0 \\ \frac{\partial}{\partial \sigma^2} \mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) \Big|_{\underline{\beta}=\hat{\underline{\beta}}_{MVR}, \sigma^2=\hat{\sigma}^2_{MVR}, \underline{\lambda}=\hat{\underline{\lambda}}_{MV}} & = -\frac{n}{2\hat{\sigma}^2_{MVR}} + \frac{1}{2\hat{\sigma}^4_{MVR}} (\underline{y} - X\hat{\underline{\beta}}_{MVR})^T (\underline{y} - X\hat{\underline{\beta}}_{MVR}) = 0 \\ \frac{\partial}{\partial \underline{\lambda}} \mathcal{L}(\underline{\beta}, \sigma^2, \underline{\lambda}) \Big|_{\underline{\beta}=\hat{\underline{\beta}}_{MVR}, \sigma^2=\hat{\sigma}^2_{MVR}, \underline{\lambda}=\hat{\underline{\lambda}}_{MV}} & = A\hat{\underline{\beta}}_{MVR} - \underline{c} = 0, \end{cases}$$

trabajando las ecuaciones resulta

$$\begin{cases} X^T \underline{y} - X^T X \hat{\underline{\beta}}_{MVR} + \hat{\sigma}^2_{MVR} A^T \hat{\underline{\lambda}}_{MV} & = 0 \\ -n\hat{\sigma}^2_{MVR} + (\underline{y} - X\hat{\underline{\beta}}_{MVR})^T (\underline{y} - X\hat{\underline{\beta}}_{MVR}) & = 0 \\ A\hat{\underline{\beta}}_{MVR} - \underline{c} & = 0, \end{cases} \tag{3.14}$$

de la segunda ecuación se tiene que

$$\hat{\sigma}^2_{MVR} = \frac{1}{n} (\underline{y} - X\hat{\underline{\beta}}_{MVR})^T (\underline{y} - X\hat{\underline{\beta}}_{MVR}). \tag{3.15}$$

Entonces, el sistema (3.14) queda de la forma

$$\begin{bmatrix} X^T X & -\hat{\sigma}^2_{MVR} A^T \\ A & 0_{j \times j} \end{bmatrix} \cdot \begin{bmatrix} \hat{\underline{\beta}}_{MVR} \\ \hat{\underline{\lambda}}_{MV} \end{bmatrix} = \begin{bmatrix} X^T \underline{y} \\ \underline{c} \end{bmatrix},$$

con $0_{j\times j}$ la matriz de ceros de dimensión $j \times j$. Para resolver el sistema por Gauss-Jordan se plantea la matriz

$$\left(\begin{array}{cc|c} X^TX & -\hat{\sigma}^2_{MVR}A^T & X^T\underline{y} \\ A & 0_{j\times j} & \underline{c} \end{array} \right),$$

se multiplica el primer renglón por $(X^TX)^{-1}$ por la izquierda $(R_1 \to (X^TX)^{-1}R_1)$. Por el Teorema A.0.4 se sabe que $(X^TX)^{-1}$ existe, nótese también que las dimensiones de las matrices sí permiten el producto pues $(X^TX)^{-1}$ es de $p \times p$, A^T es de $p \times j$ y $X^T\underline{y}$ es de $p \times 1$. Entonces la matriz queda[3]

$$\left(\begin{array}{cc|c} I_p & -\hat{\sigma}^2_{MVR}(X^TX)^{-1}A^T & (X^TX)^{-1}X^T\underline{y} \\ A & 0_{j\times j} & \underline{c} \end{array} \right),$$

sustituyendo $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y}$ (obsérvese la expresión (2.7)) y restando al segundo renglón el primero multiplicado por A por la izquierda $(R_2 \to R_2 - AR_1)$ resulta

$$\left(\begin{array}{cc|c} I_p & -\hat{\sigma}^2_{MVR}(X^TX)^{-1}A^T & \underline{\hat{\beta}} \\ 0_{j\times p} & \hat{\sigma}^2_{MVR}A(X^TX)^{-1}A^T & \underline{c} - A\underline{\hat{\beta}} \end{array} \right),$$

dado que $(X^TX)^{-1}$ es d.p. (por los Teoremas A.0.4 y A.0.1), entonces $A(X^TX)^{-1}A^T$ también es definida positiva y, por lo tanto, invertible (por el Teorema A.0.3). Entonces, multiplicando al segundo renglón por $\frac{1}{\hat{\sigma}^2_{MVR}}\big\{A(X^TX)^{-1}A^T\big\}^{-1}$ $(R_2 \to \frac{1}{\hat{\sigma}^2_{MVR}}\big\{A(X^TX)^{-1}A^T\big\}^{-1}R_2)$ se tiene

$$\left(\begin{array}{cc|c} I_p & -\hat{\sigma}^2_{MVR}(X^TX)^{-1}A^T & \underline{\hat{\beta}} \\ 0_{j\times p} & I_j & \frac{1}{\hat{\sigma}^2_{MVR}}\big\{A(X^TX)^{-1}A^T\big\}^{-1}(\underline{c} - A\underline{\hat{\beta}}) \end{array} \right),$$

por último, se suma al primer renglón el producto de $\hat{\sigma}^2_{MVR}(X^TX)^{-1}A^T$ por el segundo renglón $(R_1 \to R_1 + \hat{\sigma}^2_{MVR}(X^TX)^{-1}A^TR_2)$, queda

$$\left(\begin{array}{cc|c} I_p & 0_{p\times j} & \underline{\hat{\beta}} + (X^TX)^{-1}A^T\big\{A(X^TX)^{-1}A^T\big\}^{-1}(\underline{c} - A\underline{\hat{\beta}}) \\ 0_{j\times p} & I_j & \frac{1}{\hat{\sigma}^2_{MVR}}\big\{A(X^TX)^{-1}A^T\big\}^{-1}(\underline{c} - A\underline{\hat{\beta}}) \end{array} \right),$$

$$\begin{aligned} \therefore \qquad \underline{\hat{\beta}}_{MVR} &= \underline{\hat{\beta}} + (X^TX)^{-1}A^T\big\{A(X^TX)^{-1}A^T\big\}^{-1}(\underline{c} - A\underline{\hat{\beta}}) \\ &= \underline{\hat{\beta}} - (X^TX)^{-1}A^T\big\{A(X^TX)^{-1}A^T\big\}^{-1}(A\underline{\hat{\beta}} - \underline{c}) \qquad (3.16) \\ \underline{\hat{\lambda}}_{MV} &= \frac{1}{\hat{\sigma}^2_{MVR}}\big\{A(X^TX)^{-1}A^T\big\}^{-1}(\underline{c} - A\underline{\hat{\beta}}). \qquad (3.17) \end{aligned}$$

Obsérvese que, bajo la restricción $A\underline{\beta} = \underline{c}$, tanto el estimador máximo verosímil $\underline{\hat{\beta}}_{MVR}$ (3.16) como el estimador por mínimos cuadrados $\underline{\hat{\beta}}_R$ (2.15) están dados por la misma expresión, es decir, $\underline{\hat{\beta}}_{MVR} = \underline{\hat{\beta}}_R$.

[3]En las siguientes operaciones por renglón no se mencionará explícitamente que las dimensiones de las matrices sí permiten los productos, sin embargo, se invita al lector a verificarlo.

Ahora, solo queda probar que $\hat{\underline{\beta}}_R$ y $\hat{\sigma}^2_{MVR}$ dados por las expresiones (3.16) y (3.15), efectivamente maximizan la función log verosimilitud (3.2) cuando está sujeta a la restricción $A\underline{\beta} = \underline{c}$. Para ello, se hace un análisis similar al desarrollado en la sección 3.2.

Primero, en la sección 2.3.1 se probó que $(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})$ sujeta a $A\underline{\beta} = \underline{c}$ se minimiza cuando $\underline{\beta} = \hat{\underline{\beta}}_R$, como esta función es siempre positiva pues es la suma de cuadrados del error, entonces la función (3.2) sujeta a la misma restricción se maximiza en $\underline{\beta} = \hat{\underline{\beta}}_R = \hat{\underline{\beta}}_{MVR}$ $\forall\ \sigma^2 > 0$, es decir,

$$ln\ L(\underline{y}, X, \underline{\beta}, \sigma^2) \leq ln\ L(\underline{y}, X, \hat{\underline{\beta}}_R, \sigma^2) \qquad \forall\ \sigma^2 > 0. \tag{3.18}$$

Luego, nótese que

$$\begin{aligned}
&ln\ L(\underline{y},X, \hat{\underline{\beta}}_R, \hat{\sigma}^2_{MVR}) - ln\ L(\underline{y}, X, \hat{\underline{\beta}}_R, \sigma^2) \\
&\quad= \left(-\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\hat{\sigma}^2_{MVR}) - \frac{1}{2\hat{\sigma}^2_{MVR}}(\underline{y} - X\hat{\underline{\beta}}_R)^T(\underline{y} - X\hat{\underline{\beta}}_R)\right) \\
&\qquad - \left(-\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(\underline{y} - X\hat{\underline{\beta}}_R)^T(\underline{y} - X\hat{\underline{\beta}}_R)\right) \\
&\quad= -\frac{n}{2}ln(\hat{\sigma}^2_{MVR}) - \frac{n}{2} + \frac{n}{2}ln(\sigma^2) + \frac{n\hat{\sigma}^2_{MVR}}{2\sigma^2} \qquad (3.19) \\
&\quad= -\frac{n}{2}\left(ln(\hat{\sigma}^2_{MVR}) + 1 - ln(\sigma^2) - \frac{\hat{\sigma}^2_{MVR}}{\sigma^2}\right) \\
&\quad= -\frac{n}{2}\left(ln(\frac{\hat{\sigma}^2_{MVR}}{\sigma^2}) + 1 - \frac{\hat{\sigma}^2_{MVR}}{\sigma^2}\right) \\
&\quad\geq 0, \qquad (3.20)
\end{aligned}$$

el renglón (3.19) se da por reducción algebraica y por sustituir $\hat{\sigma}^2_{MVR}$ por (3.15), y la desigualdad (3.20) se da ya que se sabe que $v \leq e^{v-1}$ y, por lo tanto, $ln\ v \leq v - 1\ \forall\ v \geq 0$ (con la igualdad en $v = 1$). Haciendo $v = \hat{\sigma}^2_{MVR}/\sigma^2$ se deduce el resultado.

Considerando lo anterior y la expresión (3.18), se tiene

$$\begin{aligned}
ln\ L(\underline{y}, X, \underline{\beta}, \sigma^2) \leq ln\ L(\underline{y}, X, \hat{\underline{\beta}}_R, \hat{\sigma}^2_{MVR}) \leq ln\ L(\underline{y}, X, \hat{\underline{\beta}}_R, \hat{\sigma}^2_{MVR}) \quad \forall\ \sigma^2 > 0 \\
\therefore \qquad ln\ L(\underline{y}, X, \underline{\beta}, \sigma^2) \leq ln\ L(\underline{y}, X, \hat{\underline{\beta}}_R, \hat{\sigma}^2_{MVR}) \qquad \forall\ \sigma^2 > 0.
\end{aligned}$$

Dado que $\hat{\underline{\beta}}_{MVR} = \hat{\underline{\beta}}_R$, entonces el estimador máximo verosímil con restricciones tiene las propiedades descritas en el Teorema 2.3.1.

3.5. Pruebas de hipótesis

Una vez que se han estimado los parámetros desconocidos del modelo, sigue analizar los resultados. Entonces, surgen dos preguntas inminentes (Montgomery, 2012)[11]:

- ¿Es adecuado el modelo?
- ¿Cuáles variables explicativas parecen importantes y cuáles no?

Varios procedimientos de pruebas de hipótesis resultan útiles para abordar estas preguntas. Searle (1997)[12] menciona que hay varias que pueden ser de interés en muy diferentes campos de aplicación. Sin embargo, se concentra en cuatro. A continuación, se presentan estas pruebas más algunas otras:

1. $H_0 : \underline{\beta} = 0$. La hipótesis de que todos los parámetros son cero.

   No rechazar esta hipótesis implica que $\mathbb{E}(\underline{Y}) = 0$.

2. $H_0 : \underline{\beta} = \underline{b}$. Esta hipótesis establece que $\beta_i = b_i$ para $i = 0, \ldots, k$, esto es que cada β_i es igual a un valor en específico.

   Rechazar H_0 implica que al menos una β_i es diferente a la constante indicada.

3. $H_0 : A\underline{\beta} = \underline{c}$. Esta hipótesis establece que algunas combinaciones de los elementos de $\underline{\beta}$ son iguales a ciertas constantes.

   Rechazar H_0 indica que al menos una de las combinaciones lineales es diferente a la constante indicada.

4. $H_0 : \underline{\beta}_q = \underline{0}$. Esto es, algunos de los parámetros β_i, q de ellos con $q < k$, son cero.

   No rechazar H_0 indica que simultáneamente los q coeficientes no son diferentes de 0, por lo que habría evidencia estadística para quitarlas del modelo.

   Rechazar H_0 indica que al menos uno de los q coeficientes son diferentes de 0.

5. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. A ésta se le conoce como *prueba de significancia de la regresión.*

   Rechazar H_0 implica que al menos una variable explicativa $x_1, x_2, \ldots, x_k$ contribuye significativamente al modelo.

   La diferencia de esta prueba con la 1, es que aquí no se incluye al intercepto en la hipótesis nula.

6. $H_0 : \beta_i = 0$. Esta hipótesis prueba si un único coeficiente del modelo es cero.

   Si H_0 no se rechaza, entonces hay evidencia estadística que indica que la variable explicativa correspondiente al coeficiente de la hipótesis puede ser eliminada del modelo.

7. $H_0 : \sigma^2 = s$. Ésta permite probar si la varianza σ^2 es igual a cierta constante especificada.

Para desarrollar los cálculos, se requieren las siguientes definiciones.

Definición 3.5.1 *Distribución F no central. Sean* $X \sim \chi^2_{n_1}(\lambda_1)$ *y* $Y \sim \chi^2_{n_2}(\lambda_2)$ *dos v.a. independientes con distribución* χ^2 *con* n_1 *y* n_2 *g.l., y* λ_1 *y* λ_2 *parámetros de no centralidad, respectivamente. Entonces,*

$$F = \frac{X/n_1}{Y/n_2},$$

es una v.a. con distribución F no central con n_1 *y* n_2 *g.l., y* λ_1 *y* λ_2 *parámetros de no centralidad, y se denota* $F \sim F_{n_1,n_2}(\lambda_1, \lambda_2)$.

Definición 3.5.2 *Distribución t-Student. Sean* $X \sim N(0,1)$ *y* $Y \sim \chi^2_n$ *dos v.a. independientes, entonces*

$$T = \frac{X}{\sqrt{Y/n}},$$

es una v.a. con distribución t-Student con n g.l. y se denota como $T \sim t_n$.

Ahora bien, la estadística de las hipótesis 1 a 5 se puede calcular a través de un único procedimiento.

Considérese la hipótesis general

$$H_0 : K^T \underline{\beta} = \underline{m},$$

donde $\underline{\beta}$ es le vector de parámetros desconocidos de $p \times 1$ (con $p = k+1$), K^T es cualquier matriz de $s \times p$ de constantes. Se tienen dos condiciones: 1) que $rango(K^T) = s$. Lo que implica que todos los renglones de K^t sean linealmente independientes, es decir, la hipótesis debe estar compuesta de funciones lineales de los parámetros que sean linealmente independientes. Y 2) que el sistema de ecuaciones $K^T\underline{\beta} = \underline{m}$ sea consistente, lo que se logra si K^T es de rango s.

Ahora bien, se sabe que

$$\underline{\hat{\beta}} \sim N_p(\underline{\beta}, \sigma^2(X^TX)^{-1}), \qquad \text{por Teorema 3.3.1 (i)}$$

entonces,

$$K^T\underline{\hat{\beta}} - \underline{m} \sim N_s(K^T\underline{\beta} - \underline{m}, \sigma^2 K^T(X^TX)^{-1}K),$$

por Teorema 1.2.4.

Aplicando el Teorema 1.2.8 con $\underline{Y} = K^T\underline{\hat{\beta}} - \underline{m}$, entonces $\Sigma = \sigma^2 K^T(X^TX)^{-1}K$, y haciendo $A = \frac{1}{\sigma^2}\left[K^T(X^TX)^{-1}K\right]^{-1}$, se tiene que $A\Sigma = I_s$ es idempotente, así que

$$\left(K^T\underline{\hat{\beta}} - \underline{m}\right)^T \frac{1}{\sigma^2}\left[K^T(X^TX)^{-1}K\right]^{-1}\left(K^T\underline{\hat{\beta}} - \underline{m}\right),$$

es una v.a. con distribución χ^2 con s g.l., y parámetro de no centralidad

$$\lambda = \frac{\left(K^T\underline{\beta} - \underline{m}\right)^T \left[K^T(X^TX)^{-1}K\right]^{-1} \left(K^T\underline{\beta} - \underline{m}\right)}{\sigma^2},$$

definiendo $Q = \left(K^T\underline{\hat{\beta}} - \underline{m}\right)^T \left[K^T(X^TX)^{-1}K\right]^{-1} \left(K^T\underline{\hat{\beta}} - \underline{m}\right)$, entonces

$$\frac{Q}{\sigma^2} \sim \chi^2_s \left(\left(K^T\underline{\beta} - \underline{m}\right)^T \left[K^T(X^TX)^{-1}K\right]^{-1} \left(K^T\underline{\beta} - \underline{m}\right)/\sigma^2\right).$$

Ahora, sigue probar que $\frac{Q}{\sigma^2}$ y $\frac{n-p}{\sigma^2}\hat{\sigma}^2$ son independientes. Si se sustituye $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{Y}$ en Q, queda

$$\frac{Q}{\sigma^2} = \frac{1}{\sigma^2}\left(K^T(X^TX)^{-1}X^T\underline{Y} - \underline{m}\right)^T \left[K^T(X^TX)^{-1}K\right]^{-1} \left(K^T(X^TX)^{-1}X^T\underline{Y} - \underline{m}\right), \tag{3.21}$$

y dado que K^T es de $s \times p$ de rango s, por el Teorema A.0.4 $(K^TK)^{-1}$ existe. Entonces,

$$K^T(X^TX)^{-1}X^T\underline{Y} - \underline{m} = K^T(X^TX)^{-1}X^T\left[\underline{Y} - XK(K^TK)^{-1}\underline{m}\right], \tag{3.22}$$

sustituyendo (3.22) en (3.21) resulta

$$\begin{aligned}\frac{Q}{\sigma^2} &= \frac{1}{\sigma^2}\Big\{K^T(X^TX)^{-1}X^T\left[\underline{Y} - XK(K^TK)^{-1}\underline{m}\right]\Big\}^T \\ &\qquad \left[K^T(X^TX)^{-1}K\right]^{-1}\Big\{K^T(X^TX)^{-1}X^T\left[\underline{Y} - XK(K^TK)^{-1}\underline{m}\right]\Big\} \\ &= \frac{1}{\sigma^2}\left[\underline{Y} - XK(K^TK)^{-1}\underline{m}\right]^T \\ &\qquad X(X^TX)^{-1}K\left[K^T(X^TX)^{-1}K\right]^{-1}K^T(X^TX)^{-1}X^T \\ &\qquad \left[\underline{Y} - XK(K^TK)^{-1}\underline{m}\right].\end{aligned} \tag{3.23}$$

Por otro lado, considerando (3.12)se tiene

$$\frac{n-p}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}\underline{Y}^T(I_n - H)\underline{Y},$$

y dado que $X^T(I_n - H) = 0$ y $(I_n - H)X = 0$ pues $X^TH = X^T$ y $HX = X$ (obsérvese la definición de H en (2.10)), entonces

$$\frac{n-p}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}\left(\underline{Y} - XK(K^TK)^{-1}\underline{m}\right)^T(I_n - H)\left(\underline{Y} - XK(K^TK)^{-1}\underline{m}\right). \tag{3.24}$$

En las expresiones (3.22) y (3.24) se tiene a $\frac{Q}{\sigma^2}$ y $\frac{n-p}{\sigma^2}\hat{\sigma}^2$ como la forma cuadrática del vector $\underline{Y} - XK(K^TK)^{-1}\underline{m}$ que tiene una distribución normal. Donde ambas matrices en la forma cuadrática son simétricas y

$$(I_n - H)X(X^TX)^{-1}K\left[K^T(X^TX)^{-1}K\right]^{-1}K^T(X^TX)^{-1}X^T = 0,$$

por lo que aplicando el Teorema 1.2.9, $\frac{Q}{\sigma^2}$ y $\frac{n-p}{\sigma^2}\hat{\sigma}^2$ son independientes.

En resumen, tanto $\frac{Q}{\sigma^2}$ como $\frac{n-p}{\sigma^2}\hat{\sigma}^2$ tienen distribución χ^2 con s y $n-p$ g.l., respectivamente, y con $\left(K^T\underline{\beta}-\underline{m}\right)^T\left[K^T(X^TX)^{-1}K\right]^{-1}\left(K^T\underline{\beta}-\underline{m}\right)/\sigma^2$ y 0 parámetros de no centralidad, respectivamente.

Sea

$$F=\frac{(\frac{Q}{\sigma^2})/s}{\frac{n-p}{\sigma^2}\hat{\sigma}^2/(n-p)}=\frac{Q}{s\hat{\sigma}^2},$$

entonces, por la Definición 3.5.1

$$F\sim F_{s,n-p}(\left(K^T\underline{\beta}-\underline{m}\right)^T\left[K^T(X^TX)^{-1}K\right]^{-1}\left(K^T\underline{\beta}-\underline{m}\right)/\sigma^2,0),$$

y bajo la hipótesis nula $H_0: K^T\underline{\beta}=\underline{m}$, queda

$$F\sim F_{s,n-p}.$$

En resumen, la estadística que permite probar la hipótesis $H_0: K^T\underline{\beta}=\underline{m}$ es

$$F=\frac{Q}{s\hat{\sigma}^2}=\frac{\left(K^T\underline{\hat{\beta}}-\underline{m}\right)^T\left[K^T(X^TX)^{-1}K\right]^{-1}\left(K^T\underline{\hat{\beta}}-\underline{m}\right)}{s\hat{\sigma}^2}\sim F_{s,n-p}. \tag{3.25}$$

Retomando las hipótesis mencionadas al inicio de la sesión se tiene que:

1. $H_0: \underline{\beta}=0$. Se aplica la estadística F con $K^T=I_p$ y $\underline{m}=\underline{0}$. Por lo tanto, queda

$$F=\frac{\underline{\hat{\beta}}^TX^TX\underline{\hat{\beta}}}{p\hat{\sigma}^2}\sim F_{p,n-p}.$$

2. $H_0: \underline{\beta}=\underline{b}$. Se aplica la estadística F con $K^T=I_p$ y $\underline{m}=\underline{b}$. Por lo tanto, queda

$$F=\frac{(\underline{\hat{\beta}}-\underline{b})^TX^TX(\underline{\hat{\beta}}-\underline{b})}{p\hat{\sigma}^2}\sim F_{p,n-p}.$$

3. $H_0: A\underline{\beta}=\underline{c}$. Se aplica la estadística F con $K^T=A$ y $\underline{m}=\underline{c}$. Por lo tanto, queda

$$F=\frac{(A\underline{\hat{\beta}}-\underline{c})^T[A(X^TX)^{-1}A^T]^{-1}(A\underline{\hat{\beta}}-\underline{c})}{s\hat{\sigma}^2}\sim F_{s,n-p}.$$

4. $H_0 : \underline{\beta}_{\underline{q}} = \underline{0}$. Esto es, $\beta_i = 0$ para $i = 0, 1, \ldots, q$ para $q < p$. En este caso $K^T = (I_q \ 0)$, $\underline{m} = \underline{0}$, $s = q$, $\underline{\beta}_{\underline{q}} = (\beta_0 \ \beta_1 \ \cdots \beta_{q-1})^T$ y $\underline{\beta}_{\underline{p-q}} = (\beta_q \ \beta_{q+1} \ \cdots \beta_k)^T$. Entonces las particiones quedan

$$\underline{\beta} = \begin{bmatrix} \underline{\beta}_{\underline{q}} \\ \underline{\beta}_{\underline{p-q}} \end{bmatrix}, \underline{\hat{\beta}} = \begin{bmatrix} \underline{\hat{\beta}}_{\underline{q}} \\ \underline{\hat{\beta}}_{\underline{p-q}} \end{bmatrix} \ y \ (X^T X)^{-1} = \begin{bmatrix} T_{qq} & T_{qp} \\ T_{pq} & T_{pp} \end{bmatrix},$$

además $K^T \underline{\hat{\beta}} = \underline{\hat{\beta}}_{\underline{q}}$ y $[K^T (X^T X)^{-1} K]^{-1} = T_{qq}^{-1}$. Así que

$$F = \frac{\underline{\hat{\beta}}_{\underline{q}}^T T_{qq}^{-1} \underline{\hat{\beta}}_{\underline{q}}}{q\hat{\sigma}^2} \sim F_{q,n-p},$$

donde T_{qq}^{-1} es "la inversa de la parte de la inversa", es decir, se obtiene la inversa de $(X^T X)$ y luego se invierte la partición correspondiente.

5. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Esta prueba puede pensarse como un caso particular de la anterior. Pero en este caso $K^T = (0_{k\times 1} \ I_K)$, $\underline{m} = \underline{0}$, $s = k$, $\underline{\beta}_{\underline{1}} = (\beta_0)$ y $\underline{\beta}_{\underline{k}} = (\beta_1 \ \beta_2 \ \cdots \beta_k)^T$. Entonces las particiones quedan

$$\underline{\beta} = \begin{bmatrix} \underline{\beta}_{\underline{1}} \\ \underline{\beta}_{\underline{k}} \end{bmatrix}, \underline{\hat{\beta}} = \begin{bmatrix} \underline{\hat{\beta}}_{\underline{1}} \\ \underline{\hat{\beta}}_{\underline{k}} \end{bmatrix} \ y \ (X^T X)^{-1} = \begin{bmatrix} T_{11} & T_{1k} \\ T_{k1} & T_{kk}, \end{bmatrix},$$

además $K^T \underline{\hat{\beta}} = \underline{\hat{\beta}}_{\underline{k}}$ y $[K^T (X^T X)^{-1} K]^{-1} = T_{kk}^{-1}$. Así que

$$F = \frac{\underline{\hat{\beta}}_{\underline{k}}^T T_{kk}^{-1} \underline{\hat{\beta}}_{\underline{k}}}{k\hat{\sigma}^2} \sim F_{k,n-p},$$

donde, nuevamente T_{kk}^{-1} es "la inversa de la parte de la inversa".

6. $H_0 : \beta_i = 0$. Dado que $\underline{\hat{\beta}} \sim N_p(\underline{\beta}, \sigma^2 (X^T X)^{-1})$ (por el Teorema 3.3.1 (i)), haciendo $C = (X^T X)^{-1}$ y definiendo C_{ij} al elemento i, j de la matriz C, se tiene que

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 C_{(i+1)(j+1)})$$
$$\therefore \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 C_{(i+1)(j+1)}}} \sim N(0, 1).$$

Luego, dado que $\frac{n-p}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p}$ (Teorema 1.2.8 (ii)), entonces

$$T = \frac{\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 C_{(i+1)(j+1)}}}}{\sqrt{\frac{n-p}{\sigma^2}\hat{\sigma}^2/(n-p)}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(j+1)}}} \sim t_{n-p},$$

por Definición 3.5.2. Así que bajo la hipótesis nula $H_0 : \beta_i = 0$ la estadística es

$$T = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 C_{(i+1)(j+1)}}} \sim t_{n-p}.$$

7. $H_0 : \sigma^2 = s$. Del Teorema 1.2.8 (ii) se sabe que $\frac{n-p}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p}$. Luego, bajo la hipótesis nula $H_0 : \sigma^2 = s$ y, por lo tanto, la estadística queda

$$J = \frac{n-p}{s}\hat{\sigma}^2 \sim \chi^2_{n-p}.$$

Por último, una vez calculada la estadística para probar la hipótesis nula (F para las hipótesis 1 a 5, T para la hipótesis 6 y J para la 7), sigue obtener el $p - value$. Éste es, si la hipótesis nula es cierta, la probabilidad de tener el resultado obtenido o alguno más extremo, es decir, $P(F \geq F_{n_1,n_2})$, $P(|T|_n)$ y $P(J \geq \chi^2_{n-p})$, para las hipótesis 1 a 5, la 6 y la 7, respectivamente. Si esa probabilidad resulta un valor pequeño (usualmente se considera pequeño un $p - value$ menor a 0.05), entonces se concluye que hay evidencia estadística para rechazar la hipótesis nula. En otras palabras, el cálculo de la estadística de prueba permite obtener el $p - value$ y, en consecuencia, tomar una decisión con respecto a la prueba de interés.

Capítulo 4

La familia de distribuciones elípticas

La familia de distribuciones elípticas surge de una generalización de la distribución normal multivariante (d.n.m.). Propiamente, las distribuciones esféricas son la extensión de la distribución normal estándar multivariante $N_n(\underline{0}, I_n)$ y las distribuciones elípticas son una extensión del caso general $N_n(\underline{\mu}, \Sigma)$ (Fang et al., 1990)[7].

Así como la distribución normal estándar es un caso particular de la d.n.m., las distribuciones esféricas lo son de las elípticas.

Intuitivamente, para la normal bivariada la función de densidad de probabilidad resulta ser la campana de Gauss, cuyos gráficos de isodensidad[1] son: circunferencias en el caso estándar ($N_2(\underline{0}, I_2)$) y elipses en el caso general ($N_2(\underline{\mu}, \Sigma)$). Esto se ilustra en la Figura 4.1. Para la normal 3-variada, estos gráficos resultan ser esferas o elipsoides, para $N_3(\underline{0}, I_3)$ o $N_3(\underline{\mu}, \Sigma)$, respectivamente.

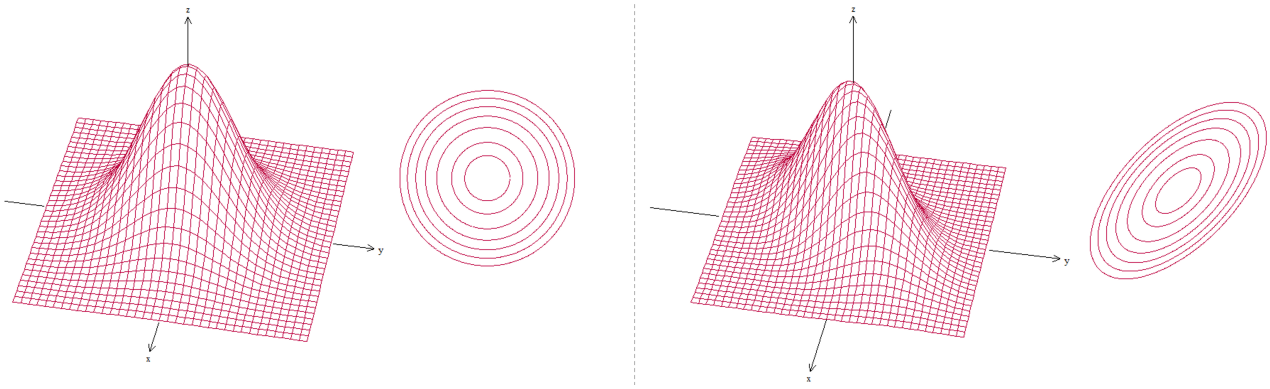

Figura 4.1: Gráficas de la función de densidad normal bivariada y sus curvas de nivel, a la izquierda $N_2(\underline{0}, I_2)$ y a la derecha $N_2(\underline{\mu}, \Sigma)$.

[1] También conocidos como curvas de nivel.

Por tal motivo, también se conocen como distribuciones contorneadas elípticamente.

El interés por las distribuciones contorneadas elípticamente surge de la necesidad de buscar alternativas a la d.n.m. Como bien explican Gupta et al. (2013)[9] y Fang et al. (1990)[7], esta familia de distribuciones ha sido ampliamente estudiada a través de varios artículos publicados en la época moderna (segunda mitad del s. XX), sin embargo, ellos se dieron a la tarea de recolectar y organizar la información en un libro (cada grupo de autores publicó su propio libro). En referencia a esto, Gupta et al. (2013)[9] dice que "ellos (los investigadores) publicaron material muy rico sobre el tema, pero los resultados están dados en *papers*, lo que no proporciona un tratamiento unificado de la teoría. Por lo tanto, parecía apropiado recoger los resultados más importantes sobre la teoría de las distribuciones matriciales de contorno elíptico disponibles en la literatura y organizarlos de una manera unificada que pueda servir como una introducción al tema."

La explicación presente en este capítulo está basada en el libro de Fang et al. (1990)[7], el artículo de Galea et al. (2000)[8] y mayormente en el libro de Gupta et al. (2013)[9], sin embargo, la teoría de este último está adaptada pues ellos generalizan y desarrollan la teoría para matrices aleatorias y, como se ha venido trabajando aquí, se presentará y desarrollará la teoría para vectores aleatorios.

4.1. Definición de la familia de distribuciones elípticas

Una manera común y sencilla de definir la familia de distribuciones elípticas es a través de su función de densidad de probabilidad (f.d.p.), sin embargo, la definición formal se da a través de función característica (f.c.) y la f.d.p. es una consecuencia.

Definición 4.1.1 ***Vector aleatorio con distribución contorneada elípticamente.*** *Sea $\underline{Y}$ un vector aleatorio de dimensiones $n \times 1$. Entonces, se dice que $\underline{Y}$ es un vector aleatorio con distribución contorneada elípticamente (d.c.e.) y absolutamente continua si su f.c. es de la forma*

$$\phi_{\underline{Y}}(\underline{t}) = exp(i\underline{t}^T\underline{\mu})\psi(\underline{t}^T\Sigma\underline{t}),$$

donde i es la unidad imaginaria, $\underline{t}$ y $\underline{\mu}$ son vectores de $n \times 1$, Σ es una matriz d.p. de $n \times n$, y $\psi : [0, \infty) \to \mathbb{R}$. Y se denota como

$$\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi).$$

Aunque en Gupta et al. (2013) la definición se da con Σ semi-definida positiva, si $\underline{Y}$ es un vector aleatorio con distribución absolutamente continua, entonces Σ debe ser estrictamente d.p. La explicación de esto se da en la siguiente sección. Por tal motivo, la Definición 4.1.1 es para vectores aleatorios con distribución absolutamente continua. Así que, en lo sucesivo por d.c.e., es decir, $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ se referirá al caso absolutamente continuo.

Definición 4.1.2 ***Vector aleatorio con distribución esférica.*** *Sea* $\underline{Y} \sim E_n(\underline{0}, I_n, \psi)$ *su distribución se conoce como distribución esférica o distribución con contorno esférico, y se denota como* $\underline{Y} \sim S_n(\psi)$.

4.2. Propiedades de distribuciones elípticas

4.2.1. Función lineal de un vector aleatorio con d.c.e.

El siguiente teorema muestra que una función lineal de un vector aleatorio con d.c.e. sigue también una d.c.e., es decir, son iguales en distribución.

Teorema 4.2.1 *Sean* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, *y* A *y* $\underline{b}$ *una matriz y un vector de constantes de dimensiones* $q \times n$ *y* $q \times 1$, *respectivamente. Entonces,*

$$A\underline{Y} + \underline{b} \sim E_q(A\underline{\mu} + \underline{b}, A\Sigma A^T, \psi).$$

Demostración:

Se define $\underline{w} = A\underline{Y} + \underline{b}$. La función característica de $\underline{w}$ es:

$$\begin{aligned}
\phi_{\underline{w}}(\underline{t}^T) &= \mathbb{E}\left[exp(i\underline{t}^T\underline{w})\right] \\
&= \mathbb{E}\left[exp\{i\underline{t}^T(A\underline{Y} + \underline{b})\}\right] \\
&= \mathbb{E}\left[exp\{i\underline{t}^T A\underline{Y}\}\right] exp\{i\underline{t}^T\underline{b}\} \\
&= \phi_{\underline{Y}}(A^T\underline{t}) exp\{i\underline{t}^T\underline{b}\} \\
&= exp\{i\underline{t}^T A\underline{\mu}\}\psi(\underline{t}^T A\Sigma A^T\underline{t}) exp\{i\underline{t}^T\underline{b}\} \\
&= exp\{i\underline{t}^T(A\underline{\mu} + \underline{b})\}\psi(\underline{t}^T A\Sigma A^T\underline{t}),
\end{aligned}$$

lo que la f.c. de un vector aleatorio con distribución $E_q(A\underline{\mu} + \underline{b}, A\Sigma A^T, \psi)$.

■

Del teorema anterior se deriva el siguiente corolario.

Corolario 4.2.1.1 *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, *entonces*

$$(\Sigma^{1/2})^{-1}(\underline{Y} - \underline{\mu}) \sim S_n(\psi). \tag{4.1}$$

Por el contrario, si $\underline{Y} \sim S_n(\psi)$, *entonces*

$$A\underline{Y} + \underline{\mu} \sim E_n(\underline{\mu}, \Sigma, \psi), \tag{4.2}$$

con $\Sigma = AA^T$.

Demostración:

Por el Teorema A.0.2 se sabe que para Σ d.p. existe $\Sigma^{1/2}$ simétrica e invertible. Aplicando el Teorema 4.2.1 el vector aleatorio $(\Sigma^{1/2})^{-1}(\underline{Y} - \underline{\mu})$ tiene d.c.e. con parámetros

$$(\Sigma^{1/2})^{-1}(\underline{\mu}) - (\Sigma^{1/2})^{-1}(\underline{\mu}) = \underline{0},$$

y

$$(\Sigma^{1/2})^{-1}\Sigma(\Sigma^{1/2})^{-1} = (\Sigma^{1/2})^{-1}(\Sigma^{1/2}\Sigma^{1/2})(\Sigma^{1/2})^{-1} = [(\Sigma^{1/2})^{-1}\Sigma^{1/2}][\Sigma^{1/2}(\Sigma^{1/2})^{-1}] = I_n.$$

En cambio, si $\underline{Y} \sim E_n(\underline{0}, I_n, \psi)$, entonces $A\underline{Y} + \underline{\mu}$ tiene d.c.e. con parámetros

$$\Sigma^{1/2}\underline{0} + \underline{\mu} = \underline{\mu},$$

y

$$AI_nA^T = AA^T = \Sigma.$$

■

Además, si $\underline{Y} \sim S_n(\psi)$ del Teorema 4.2.1 se tiene que si G es una matriz ortogonal de $n \times n$, entonces $G\underline{Y} \sim S_n(\psi)$. En varios artículos es justo así como se define la distribución esférica, como aquella que no cambia de distribución o que es invariante en distribución bajo rotaciones y reflexiones.

4.2.2. Distribución marginal

Teorema 4.2.2 *Sean $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ y las particiones de $\underline{Y}$, $\underline{\mu}$ y Σ de la forma (1.11). Entonces,*

$$\underline{Y}_1 \sim E_{n_1}(\underline{\mu}_1, \Sigma_{11}, \psi).$$

Demostración:

Se define $A = [I_{n_1}\ 0_{n_1 \times n_2}]$ una matriz de dimensiones $n_1 \times n$, entonces $A\underline{Y} = \underline{Y}_1$. Del Teorema 4.2.1 se tiene que

$$\underline{Y}_1 \sim E_{n_1}(A\underline{\mu}, A\Sigma A^T, \psi),$$

es decir,

$$\underline{Y}_1 \sim E_{n_1}(\underline{\mu}_1, \Sigma_{11}, \psi).$$

■

Del teorema anterior se puede deducir que si $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, entonces $Y_i \sim E_1(\mu_i, \sigma_{ii}, \psi)$ con μ_i el i-ésimo elemento del vector $\underline{\mu}$ y σ_{ii} el i-ésimo elemento de la diagonal de Σ.

A continuación, se explica por qué en la Definición 4.1.1 Σ debe ser d.p. para que la distribución sea absolutamente continua. Supóngase que Σ es semi-definida positiva, entonces la descomposición espectral es $\Sigma = GDG^T$ con G una matriz ortogonal y $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ con $\lambda_1, \lambda_2, \ldots, \lambda_n$ los valores propios de Σ. Ahora, como Σ es semi-definida positiva al menos un valor propio es cero, digamos $\lambda_1 = 0$. Luego, sea $\underline{w} = G^T(\underline{Y} - \underline{\mu})$, entonces $\underline{w}$ es también de distribución absolutamente continua. Por otro lado, del Teorema 4.2.2 se sabe que $w_1 \sim S_1(\psi)$ esto es $w_1 \sim E_1(0, 0, \psi)$ así que w_1 es degenerada. Pero la marginal de una distribución absolutamente continua es una distribución absolutamente continua. Por lo tanto, Σ debe ser d.p. para que el vector aleatorio tenga d.c.e. absolutamente continua.

4.2.3. Función de densidad de probabilidad

Teorema 4.2.3 *Sea $\underline{Y}$ un vector aleatorio de $n \times 1$ con distribución absolutamente continua. Entonces, $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ si y solo si su f.d.p. tiene la forma*

$$f_{\underline{Y}}(\underline{y}) = c_i |\Sigma|^{-1/2} g\big\{(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\big\}, \tag{4.3}$$

con c_i la constante de normalización y $g(\cdot)$ una función no negativa de la forma $g(\underline{y}^T \underline{y})$ llamada función generadora de densidad de probabilidad[2].

Demostración:

Se sabe que:

1. El teorema de Vinograd establece que sean A y B matrices de $n \times q$ y $q \times r$, respectivamente, con $q \leq r$. Entonces, $AA^T = BB^T$ si y solo si existe una matriz H de $n \times r$ con $HH^T = I_n$ tal que $B = AH$.

Primero, se prueba que si $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, entonces su f.d.p. es de la forma (4.3).

Para el caso en que $\underline{Z} \sim S_n(\psi)$ del Corolario 4.2.1.1 se sabe que si H es matriz ortogonal de dimensiones $n \times n$, entonces $H\underline{Z} \sim S_n(\psi)$ así que la distribución de $\underline{Z}$ es invariante bajo transformaciones ortogonales, por (1) se concluye que la f.d.p. depende únicamente de $\underline{Z}^T \underline{Z}$, digamos que es $f_{\underline{Z}}(\underline{z}) = g(\underline{z}^T \underline{z})$, donde g solo depende de n y ψ.

Ahora bien, para el caso $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, también del Corolario 4.2.1.1 se tiene que $\underline{Z} = \Sigma^{-1/2}(\underline{Y} - \underline{\mu}) \sim S_n(\psi)$, entonces el jacobiano de la transformación $\underline{Z} \to \underline{Y}$ es $|\Sigma^{-1/2}|$.

[2] Al término de este subtema se discuten con más detalle las características de c_i y $g(\cdot)$.

Así que la f.d.p. de $\underline{Y}$ es

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= |\Sigma^{-1/2}| g(\underline{z}^T \underline{z}) \\ &= |\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T \Sigma^{-1/2}\Sigma^{-1/2}(\underline{y}-\underline{\mu})\right) \\ &= |\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T \Sigma^{-1}(\underline{y}-\underline{\mu})\right). \end{aligned}$$

Luego, queda probar que si $\underline{Y}$ tiene su f.d.p. de la forma (4.3), entonces su distribución es elípticamente contorneada.

Se asume que $\underline{Y}$ es un vector aleatorio cuya f.d.p. es (4.3), ahora se busca probar que $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$. Sea $\underline{Z} = \Sigma^{-1/2}(\underline{Y} - \underline{\mu})$ y sea $h(\underline{z}^T \underline{z})$ su f.d.p. Entonces, su función característica es

$$\phi_{\underline{Z}}(\underline{t}) = \int_{\mathbb{R}^n} exp(i\underline{t}^T \underline{z}) h(\underline{z}^T \underline{z})\, d\underline{z},$$

con $\underline{t}$ de $n \times 1$.

Ahora, se prueba que si $\underline{t}_1$ y $\underline{t}_2$ son vectores de $n \times 1$ tales que $\underline{t}_1^T \underline{t}_1 = \underline{t}_2^T \underline{t}_2$, entonces $\phi(\underline{t}_1) = \phi(\underline{t}_2)$. Usando (1), se sabe que existe H ortogonal de $n \times n$ tal que $\underline{t}_1^T H = \underline{t}_2^T$. Entonces,

$$\begin{aligned} \phi_{\underline{Z}}(\underline{t}_2) &= \int_{\mathbb{R}^n} exp(i\underline{t}_2^T \underline{z}) h(\underline{z}^T \underline{z})\, d\underline{z} \\ &= \int_{\mathbb{R}^n} exp(i\underline{t}_1^T H\underline{z}) h(\underline{z}^T \underline{z})\, d\underline{z}. \end{aligned}$$

Sea $\underline{W} = H\underline{Z}$. El jacobiano de la transformación es $|H^T|^n = 1$. Entonces,

$$\begin{aligned} \phi_{\underline{Z}}(\underline{t}_2) &= \int_{\mathbb{R}^n} exp(i\underline{t}_1^T \underline{w}) h(\underline{w}^T H H^T \underline{w})\, d\underline{w} \\ &= \int_{\mathbb{R}^n} exp(i\underline{t}_1^T \underline{w}) h(\underline{w}^T \underline{w})\, d\underline{w} \\ &= \phi_{\underline{Z}}(\underline{t}_1). \end{aligned}$$

Por consiguiente, $\phi_{\underline{Z}}(\underline{t})$ es función de $\underline{t}^T \underline{t}$, por lo que existe una función ψ tal que $\phi_{\underline{Z}}(\underline{t}) = \psi(\underline{t}^T \underline{t})$, es decir, $\underline{Z} \sim S_n(\psi) = E_n(\underline{0}, I_n, \psi)$. Y por Corolario 4.2.1.1 se concluye que $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$.

■

En la expresión (4.3) la función g es comúnmente conocida como *función generadora de densidad de probabilidad* o simplemente *función generadora*. Como se lee en la demostración del teorema anterior la función de densidad de $\underline{Y} \sim S_n(\psi)$ debe ser de la forma

$g(\underline{Y}^T\underline{Y})$ para alguna función no negativa $g(\cdot)$, en Fang et al. (1990)[7] en la sección de conocimientos preliminares se puede encontrar la demostración a la siguiente igualdad

$$\int g(\underline{y}^T\underline{y})d\underline{y} = \frac{\pi^{n/2}}{\Gamma(n/2)}\int_0^\infty y^{n/2-1}g(y)dy.$$

Entonces, dado que la igualdad anterior debe ser 1, se puede concluir que cualquier función $g(\cdot)$ no negativa puede ser usada para definir una función de densidad $g(\underline{y}^T\underline{y})$ de alguna distribución esférica si y solo si

$$\int_0^\infty y^{n/2-1}g(y)dy < \infty.$$

También, en Fang et al.(1990)[7] se prueba un teorema, a través de la representación estocástica de un vector aleatorio con d.c.e., que establece la relación entre la función generadora g y la función de densidad, digamos f. Gracias a esto se concluye que la constante de normalización c_i en la función generadora de densidad g está dada por la expresión:

$$c_i = \frac{\Gamma(n/2)}{2\pi^{n/2}\int_0^\infty y^{n-1}g(y^2)dy}. \tag{4.4}$$

4.2.4. Valor esperado y covarianza

Teorema 4.2.4 *Sea $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, y Y_i y μ_i el i-ésimo elemento de los vectores $\underline{Y}$ y $\underline{\mu}$, respectivamente, y σ_{ij} el elemento ij de la matriz Σ. Se cumple que:*

(i) Si $\underline{Y}$ tiene primer momento finito, entonces $\mathbb{E}(\underline{Y}) = \underline{\mu}$.

(ii) Si $\underline{Y}$ tiene segundo momento finito, entonces $Var(\underline{Y}) = c\Sigma$.

(iii) Si $\mathbb{E}(Y_iY_j)$ existe, entonces $\mathbb{E}(Y_iY_j) = c\sigma_{ij} + \mu_i\mu_j$.

(iv) Si $\mathbb{E}(\underline{Y}^TA\underline{Y})$ existe, entonces $\mathbb{E}(\underline{Y}^TA\underline{Y}) = c\,tr(A\Sigma) + \underline{\mu}^TA\underline{\mu}$.

Con $c = -2\psi'(0)$, $tr(\cdot)$ la traza de una matriz y A cualquier matriz de constantes de dimensión $n \times n$.

Demostración:

(i) y (ii)

Primero, sea $\underline{Z} \sim S_n(\phi) = E_n(\underline{0}, I_n, \psi)$, del Teorema 4.2.1 se sabe que

$$-I_n\underline{Z} \sim E_n(\underline{0}, I_n, \psi).$$

Por lo tanto, $\mathbb{E}(\underline{Z}) = \mathbb{E}(-\underline{Z}) = \underline{0}$.

La matriz de varianzas y covarianzas puede construirse ocupando la función característica, donde los elementos de la diagonal son

$$(-i)^2 \frac{\partial^2}{\partial t_i^2} \phi_{\underline{Z}}(\underline{t}) \Big|_{\underline{t}=\underline{0}},$$

y los demás elementos son

$$(-i)^2 \frac{\partial^2}{\partial t_i \partial t_j} \phi_{\underline{Z}}(\underline{t}) \Big|_{\underline{t}=\underline{0}}.$$

Además, la función característica de $\underline{Z}$ es de la forma $\phi(\underline{t}) = \psi(\underline{t}^T \underline{t})$ por la Definición 4.1. Entonces,

$$\begin{aligned} \frac{\partial}{\partial t_i} \phi_{\underline{Z}}(\underline{t}) &= \frac{\partial}{\partial t_i} \psi(\underline{t}^T \underline{t}) \\ &= \frac{\partial}{\partial t_i} \psi \left(\sum_{i=1}^{n} t_i^2 \right) \\ &= 2t_i \psi' \left(\sum_{i=1}^{n} t_i^2 \right), \end{aligned}$$

obteniendo la segunda derivada parcial resulta

$$\begin{aligned} \frac{\partial^2}{\partial t_i^2} \phi_{\underline{Z}}(\underline{t}) &= \frac{\partial}{\partial t_i} \left[2t_i \psi' \left(\sum_{i=1}^{n} t_i^2 \right) \right] \\ &= 2\psi' \left(\sum_{i=1}^{n} t_i^2 \right) + 4t_i^2 \psi'' \left(\sum_{i=1}^{n} t_i^2 \right), \end{aligned}$$

y si $i \neq j$

$$\begin{aligned} \frac{\partial^2}{\partial t_j \partial t_i} \phi_{\underline{Z}}(\underline{t}) &= \frac{\partial}{\partial t_j} \left[2t_i \psi' \left(\sum_{i=1}^{n} t_i^2 \right) \right] \\ &= 4t_i t_j \psi'' \left(\sum_{i=1}^{n} t_i^2 \right). \end{aligned}$$

Entonces,

$$(-i)^2 \frac{\partial^2}{\partial t_i^2} \phi_{\underline{Z}}(\underline{t}) \Big|_{\underline{t}=\underline{0}} = -2\psi'(0), \qquad y \tag{4.5}$$

$$(-i)^2 \frac{\partial^2}{\partial t_i \partial t_j} \phi_{\underline{Z}}(\underline{t}) \Big|_{\underline{t}=\underline{0}} = 0 \tag{4.6}$$

$$\therefore \qquad Var(\underline{Z}) = -2\psi'(0)\, I_n. \tag{4.7}$$

Sea $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, con $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$. Entonces $\underline{Z} = \Sigma^{1/2}(\underline{Y} - \underline{\mu}) \sim E_n(\underline{0}, I_n, \psi)$ y $\underline{Y} = \Sigma^{1/2}\underline{Z} + \underline{\mu}$, así que

$$\begin{aligned} \mathbb{E}(\underline{Y}) &= \mathbb{E}(\Sigma^{1/2}\underline{Z} + \underline{\mu}) = \underline{\mu}, \qquad y \\ Var(\underline{Y}) &= Var(\Sigma^{1/2}\underline{Z} + \underline{\mu}) \\ &= \Sigma^{1/2} Var(\underline{Z}) \Sigma^{1/2} \\ &= -2\psi'(0)\, \Sigma. \end{aligned}$$

(iii)

Sea $\underline{Y} = \underline{Z} + \underline{\mu}$ con $\underline{Z} \sim E_n(\underline{0}, \Sigma, \psi)$, del Teorema 4.2.2 se sabe que $Z_i \sim E_1(0, \sigma_{ii}, \psi)$ con σ_{ii} el i-ésimo elemento de la diagonal de Σ. Entonces

$$\begin{aligned} \mathbb{E}(Y_i Y_j) &= \mathbb{E}\left((Z_i - \mu_i)(Z_j - \mu_j)\right) \\ &= \mathbb{E}(Z_i Z_j - Z_i\mu_j - \mu_i Z_j + \mu_i\mu_j) \\ &= \mathbb{E}(Z_i Z_j) - \mu_j\, \mathbb{E}(Z_i) - \mu_i\, \mathbb{E}(Z_j) + \mu_i\mu_j \\ &= Cov(Z_i, Z_j) + \mu_i\mu_j \\ &= c\sigma_{ij} + \mu_i\mu_j, \end{aligned}$$

pues de (i) y (ii) sabemos que $\mathbb{E}(Z_i) = 0$ y $Cov(Z_i, Z_j) = c\sigma_{ij}$.

(iv)

Primero, obsérvese que

$$\begin{aligned} \underline{Y}^T A \underline{Y} &= \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n Y_i a_{i1} & \sum_{i=1}^n Y_i a_{i2} & \cdots & \sum_{i=1}^n Y_i a_{in} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \sum_{i=1}^n Y_i a_{i1} Y_1 + \sum_{i=1}^n Y_i a_{i2} Y_2 + \cdots + \sum_{i=1}^n Y_i a_{in} Y_n \\ &= \sum_{i=1}^n \sum_{j=1}^n Y_i a_{ij} Y_j. \end{aligned} \tag{4.8}$$

Entonces,

$$\begin{aligned}\mathbb{E}(\underline{Y}^T A\underline{Y}) &= \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} Y_i a_{ij} Y_j\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\,\mathbb{E}(Y_i Y_j)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(c\sigma_{ij} + \mu_i\mu_j)\\ &= c\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\sigma_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n}\mu_i a_{ij}\mu_j.\end{aligned}$$

Ahora bien, con el mismo desarrollo con el que se obtuvo (4.8) se puede concluir que $\sum_{i=1}^{n}\sum_{j=1}^{n}\mu_i a_{ij}\mu_j = \underline{\mu}^T A\underline{\mu}$. Por otro lado, obsérvese que

$$\begin{aligned}tr(A\Sigma) &= tr\left(\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{bmatrix}\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n}\\ \vdots & \vdots & & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}\end{bmatrix}\right)\\ &= \sum_{j=1}^{n} a_{1j}\sigma_{j1} + \sum_{j=1}^{n} a_{2j}\sigma_{j2} + \cdots + \sum_{j=1}^{n} a_{nj}\sigma_{jn}\\ &= \sum_{j=1}^{n}\sum_{j=1}^{n} a_{ij}\sigma_{ji}\\ &= \sum_{j=1}^{n}\sum_{j=1}^{n} a_{ij}\sigma_{ij},\end{aligned}$$

pues Σ es simétrica. Entonces

$$\mathbb{E}(\underline{Y}^T A\underline{Y}) = c\; tr(A\Sigma) + \underline{\mu}^T A\underline{\mu}.$$

■

Corolario 4.2.4.1 *Bajo las condiciones del Teorema 4.2.4*

$$cor(Y_i, Y_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \tag{4.9}$$

con σ_{ij} el elemento i, j de la matriz Σ.

Demostración:

Del Teorema 4.2.4 se tiene que la $Cov(Y_i, Y_j) = c\sigma_{ij}$, $Var(Y_i) = c\sigma_{ii}$ y $Var(Y_j) = c\sigma_{jj}$, con $c = -2\psi'(0)$. Entonces,

$$cor(Y_i, Y_j) = \frac{c\sigma_{ij}}{\sqrt{c^2\sigma_{ii}\sigma_{jj}}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

■

Es decir, que la correlación entre dos elementos del vector aleatorio $\underline{Y}$ depende únicamente de Σ y no de ψ.

4.2.5. Forma cuadrática

Definición 4.2.1 *Sea $\underline{Y} \sim E_n(\underline{0}, \Sigma, \psi)$ con Σ d.p. Entonces, $G_{1,1}(\Sigma, \frac{n}{2}, \psi)$ denota la distribución de $\underline{Y}^T\underline{Y}$.*

Para poder expresar la f.d.p. de la distribución G, se ocupa el siguiente lema probado en Anderson (2003)[1].

Lema 4.2.5 *Sea X una matriz aleatoria de $n \times p$ con f.d.p. $f(XX^T)$. Sea $A = XX^T$, entonces la f.d.p. de A es*

$$\frac{\pi^{pn/2}}{\Gamma_p(\frac{n}{2})}|A|^{\frac{n-p-1}{2}} f(A),$$

donde $\Gamma_p(t) = \pi^{\frac{p(p-1)}{4}}\Pi_{i=1}^{p}\Gamma(t - \frac{i-1}{2})$ y A d.p.

Como se comentó al inicio de este capítulo, esta definición fue tomada del libro Gupta et al. (2013)[9] quienes definen la d.c.e. desde un punto de vista matricial. En consecuencia, este lema se adaptará a vectores aleatorios en el siguiente teorema.

Teorema 4.2.6 *Sea $\underline{Y} \sim E_n(\underline{0}, \Sigma, \psi)$ cuya f.d.p. es*

$$f_{\underline{Y}}(\underline{y}) = \frac{c_i}{|\Sigma|^{1/2}} g(\underline{y}^T\Sigma^{-1}\underline{y}),$$

con c_i la constante de normalización y sea $W = \underline{Y}^T\underline{Y}$, entonces su f.d.p. es

$$f_W(w) = \frac{c_i\pi^{n/2}|\Sigma|^{-1/2}}{\Gamma_1(\frac{n}{2})} w^{\frac{n-2}{2}} g\left(tr(\Sigma^{-1}w\right) \qquad con\ w > 0. \tag{4.10}$$

con $tr(\cdot)$ la traza de la matriz.

Demostración:

Dado que $\underline{Y}$ es un vector aleatorio de dimensiones $n \times 1$ y su f.d.p. es de la forma $c_i|\Sigma|^{1/2}g(\underline{y}^T\Sigma^{-1}\underline{y})$, aplicando el Lema 4.2.5 con $p = 1$ y $n = n$ se obtiene (4.10).

■

Del teorema anterior se deduce el siguiente corolario:

Corolario 4.2.6.1 *Sean $\underline{Y} \sim E_n(\underline{0}, \sigma^2 I_n, \psi)$ y $W = \underline{Y}^T\underline{Y}$, entonces su f.d.p. es*

$$f_W(w) = \frac{c_i\pi^{n/2}(\sigma^2)^{-n/2}}{\Gamma(\frac{n}{2})}w^{\frac{n}{2}-1}g\left(\frac{1}{\sigma^2}w\right) \qquad con\ w > 0. \tag{4.11}$$

Demostración:

Si $\underline{Y} \sim E_n(\underline{0}, \sigma^2 I_n, \psi)$, entonces su f.d.p. es:

$$f_{\underline{Y}}(\underline{y}) = \frac{c_i}{|\sigma^2 I_n|^{1/2}}g(\underline{y}^T(\sigma^2 I_n)^{-1}\underline{y}) = \frac{1c_i}{(\sigma^2)^{n/2}}g\left(\frac{1}{\sigma^2}\underline{y}^T\underline{y}\right).$$

Y por el Teorema (4.2.6) se tiene que:

$$\begin{aligned} f_W(w) &= \frac{c_i\pi^{n/2}|\sigma^2 I_n|^{-1/2}}{\Gamma_1(\frac{n}{2})}w^{\frac{n-2}{2}}g\left(\frac{1}{\sigma^2}w\right) \\ &= \frac{c_i\pi^{n/2}(\sigma^2)^{-n/2}}{\Gamma(\frac{n}{2})}w^{\frac{n}{2}-1}g\left(\frac{1}{\sigma^2}w\right) \qquad con\ w > 0. \end{aligned}$$

■

Teorema 4.2.7 *Sea $\underline{Y} \sim E_n(\underline{0}, \Sigma, \psi)$ con Σ d.p. Si[3] A es una matriz simétrica e idempotente de dimensión $n \times n$ y rango$(A) = k$ con $k \leq n$. Entonces,*

$$\underline{Y}^T A\underline{Y} \sim G_{1,1}(\Sigma, \frac{k}{2}, \psi).$$

Demostración:

[3]El teorema se extiende en la otra dirección también, es decir, es *si y solo si*. Sin embargo, para desarrollar la teoría inferencial en el presente trabajo solo se requiere de la suficiencia. La demostración de necesidad puede encontrarse en Gupta et al. (2013) pág. 131[9], aunque para probarla se requiere asumir que existe un vector v de dimensión $n \times 1$ tal que $P(v^T\Sigma^{-1}\underline{Y} = 0) = 0$.

Es suficiente considerar el caso $\Sigma = I_n$ ya que de otra forma basta con definir $\underline{Z} = \Sigma^{-1/2}\underline{Y}$ y probar $\underline{Y}^T A\underline{Y} \sim G_{1,1}(\Sigma, \frac{k}{2}, \psi)$ es equivalente a probar $\underline{Z}^T A\underline{Z} \sim G_{1,1}(\Sigma, \frac{k}{2}, \psi)$.

Asumiendo $A^2 = A$, se sabe que existe una matriz ortogonal G tal que

$$A = G \begin{bmatrix} I_n & 0_{(k)\times(n-k)} \\ 0_{(k-n)\times(k)} & 0_{(k-n)\times(n-k)} \end{bmatrix} G^T,$$

donde 0_{ij} denota la matriz de ceros de dimensión $i \times j$.

Defínanse $C = \begin{bmatrix} I_k \\ 0_{(k-n)\times(k),} \end{bmatrix}$ y $\underline{Z} = C^T G\underline{Y}$ con $\underline{Y} \sim E_n(\underline{0}, I_n, \psi)$, una matriz y un vector de dimensiones $n \times k$ y $k \times 1$, respectivamente. Obsérvese que $A = GCC^T G$ y $C^T C = I_k$. Entonces, por el Teorema 4.2.1

$$\underline{Z} \sim E_k(C^T G\underline{0}, C^T G I_n GC, \psi) = E_k(\underline{0}, I_k, \psi),$$

pues G es ortogonal. Además,

$$\underline{Z}^T\underline{Z} = \underline{Y}^T GCC^T G\underline{Y} = \underline{Y}^T A\underline{Y},$$

y por la Definición 4.2.1 se tiene que $\underline{Z}^T\underline{Z} \sim G_{1,1}(I_k, \frac{k}{2}, \psi)$, por lo tanto,

$$\underline{Y}^T A\underline{Y} \sim G_{1,1}(I_k, \frac{k}{2}, \psi).$$

■

4.3. Ejemplos de distribuciones elípticas

En esta sección se presentan subclases de la familia de distribuciones elípticas.

En el caso multivariante, para cada una se da la definición formal a través de la función generadora de densidad, se encuentra la constante de normalización usando (4.4), se muestra la función de densidad y se ofrecen algunos comentarios.

Inspirada en las tablas que presentan Fang et al. (1990)[7] y Galea et al. (2000)[8], la Tabla (C.1) del Apéndice resume en un listado algunas distribuciones elípticas, indicando su nombre, notación y su función generadora de densidad g. Recuérdese que la función de densidad de probabilidad f se obtiene usando (4.3). La notación $c_1, c_2, \ldots, c_{10}$ es usada para denotar la constante de normalización, la cual se obtiene por medio de la expresión (4.4).

4.3.1. Caso de una dimensión

En el caso en que $n = 1$, las distribuciones $E_1(\mu, \sigma, \psi)$ coinciden con aquellas que son simétricas alrededor de un punto. De forma más precisa, $Y \sim E_1(\mu, \sigma, \psi)$ si y solo si $P(Y \leq r) = P(Y \geq \mu - r)$ para todo $r \in \mathbb{R}$. Algunas de ellas son las distribuciones: uniforme, Cauchy, doble exponencial, t-Student, y la distribución cuya f.d.p. es:

$$f_Y(y) = \frac{\sqrt{2}}{\pi\sigma\left(1 + (\frac{y}{\sigma})^4\right)}, \qquad \sigma > 0.$$

4.3.2. Distribución Uniforme multivariante

El vector aleatorio U de dimensión $n \times 1$ se dice que tiene distribución Uniforme multivariante si se distribuye uniformemente en la esfera unitaria en $\mathbb{R}^n$.

4.3.3. Distribución tipo Kotz simétrica multivariante

Definición 4.3.1 ***Distribución tipo Kotz.*** *Si $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ y su función generadora es de la forma*

$$g(u) = u^{N-1} exp(-ru^s),\ r, s > 0,\ 2N + n > 2, \tag{4.12}$$

se dice que $\underline{Y}$ tiene una distribución simétrica tipo Kotz y se denota $\underline{Y} \sim K_n(r, s, N, \underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_1 |\Sigma|^{-1/2} g\left((\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right) \\ &= c_1 \left[(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right]^{N-1} exp\{ - r \left[(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right]^s \}, \end{aligned} \tag{4.13}$$

con

$$c_1 = \frac{s\Gamma\left(\frac{n}{2}\right)}{\pi^{n/2}\Gamma\left(\frac{2N+n-2}{2s}\right)} r^{\frac{2N+n-2}{2s}}.$$

Cuando $s = 1$, es la distribución Kotz original propuesta por Kotz (1975). Cuando $N = 1$, $s = 1$ y $r = 1/2$, la distribución se reduce a la normal multivariada. Es decir, esta distribución es una generalización de la normal multivariada y es útil en la contrucción de modelos en los que el supuesto de normalidad no aplica (Fang et al., 1990)[7].

4.3.4. Distribución Normal multivariante

Definición 4.3.2 ***Distribución Normal.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = exp(-u/2), \tag{4.14}$$

se dice que $\underline{Y}$ *posee una distribución normal multivariante y se denota* $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$.

Su función de densidad es (1.6). La constante de normalización, digamos c_2, se puede obtener usando (4.4).

Si bien en la sección 1.2 se prueban y comentan algunas de las propiedades más comunes de esta distribución, hasta este punto se puede agregar una más y es que pertenece a la familia de d.c.e.

4.3.5. Distribución Pearson tipo VII simétrica multivariante

Definición 4.3.3 ***Distribución Pearson tipo VII.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = \left(1 + \frac{u}{m}\right)^{-N}, \; N > \frac{n}{2}, \; m > 0, \tag{4.15}$$

entonces se dice que $\underline{Y}$ *tiene una distribución Pearson tipo VII simétrica multivariante con parámetros* $N, m \in \mathbb{R}$, $\underline{\mu}$ *un vector de* $n \times 1$ *y* Σ *una matriz d.p. de dimensiones* $n \times n$, *y se denota* $\underline{Y} \sim MPVII_n(N, m, \underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_3 |\Sigma|^{-1/2} g\left((\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right) \\ &= \frac{\Gamma(N)}{(\pi m)^{n/2} \Gamma\left(N - \frac{n}{2}\right) |\Sigma|^{1/2}} \left(1 + \frac{(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})}{m}\right)^{-N}. \end{aligned} \tag{4.16}$$

Un caso especial es cuando $N = \frac{m+n}{2}$, $\underline{Y}$ se decide que tiene una distribución t multivariante con m g.l. Más aún, si $m = 1$, entonces se tiene la distribución Cauchy multivariante $(m = 1, N = (m + n)/2)$.

4.3.6. Distribución t-Student multivariante

Ésta es una caso particular de la distribución Pearson tipo VII con $N = \frac{1}{2}(m + n)$ (obsérvese la Definición 4.3.3).

Definición 4.3.4 ***Distribución t-Student.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = \left(1 + \frac{u}{m}\right)^{-\frac{m+n}{2}}, \ u \geq 0, \tag{4.17}$$

entonces se dice que $\underline{Y}$ *tiene una distribución t-Student multivariante con* m *g.l. y se denota* $\underline{Y} \sim t_n(m, \underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_4 |\Sigma|^{-1/2} g\left((\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right) \\ &= \frac{\Gamma(\frac{m+n}{2})}{(\pi m)^{n/2} \Gamma\left(\frac{m}{2}\right) |\Sigma|^{1/2}} \left(1 + \frac{(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})}{m}\right)^{-\frac{m+n}{2}}. \end{aligned} \tag{4.18}$$

Dos propiedades de un vector aleatorio con distribución t-Student, son que la distribución de una función lineal y de una partición, esto es, la distribución marginal, son también t-Student.

4.3.7. Distribución Cauchy multivariante

La distribución $t_n(1, \underline{\mu}, \Sigma)$ se conoce como la distribución Cauchy multivariante.

Definición 4.3.5 ***Distribución Cauchy.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = (1 + u)^{-(n+1)/2}, \ u \geq 0, \tag{4.19}$$

entonces se dice que $\underline{Y}$ *tiene una distribución Cauchy multivariante y se denota como* $\underline{Y} \sim C_n(\underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_5 |\Sigma|^{-1/2} g\left((\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right) \\ &= \frac{\Gamma(\frac{n+1}{2})}{(\pi)^{n/2} \Gamma\left(\frac{1}{2}\right) |\Sigma|^{1/2}} \left(1 + (\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right)^{-\frac{n+1}{2}}. \end{aligned} \tag{4.20}$$

4.3.8. Distribución Pearson tipo II simétrica multivariante

Definición 4.3.6 ***Distribución Pearson tipo II.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = (1 - u)^m, \qquad 0 \leq u \leq 1, \ m > -1, \tag{4.21}$$

entonces se dice que $\underline{Y}$ *tiene una distribución Pearson tipo II simétrica multivariante y se denota* $\underline{Y} \sim MPII_n(m, \underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_6|\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right) \\ &= \frac{\Gamma(\frac{n}{2}+m+1)}{(\pi)^{n/2}\Gamma(m+1)\,|\Sigma|^{1/2}}\left[1-(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right]^m. \end{aligned} \tag{4.22}$$

4.3.9. Distribución Logística simétrica multivariante

Definición 4.3.7 ***Distribución Logística.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = \frac{e^{-u}}{(1+e^{-u})^2}, \qquad u \geq 0, \tag{4.23}$$

entonces se dice que $\underline{Y}$ *tiene una distribución logística elípticamente simétrica y se denota* $\underline{Y} \sim L_n(\underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_7|\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right) \\ &= \frac{c_7}{|\Sigma|^{-1/2}}\frac{exp\big\{-(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\big\}}{\left(1+exp\big\{-(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\big\}\right)^2} \end{aligned} \tag{4.24}$$

con

$$c_7 = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)}\int_0^\infty z^{\frac{n}{2}-1}\frac{e^{-z}}{(1+e^{-z})^2}\,dz.$$

Fang et al., (1990) comentan que varios autores han estudiado esta distribución usando diferentes definiciones, por eso en 4.3.7 la definen como "distribución logística elípticamente simétrica" para diferenciarla de las demás.

4.3.10. Distribución Bessel simétrica multivariante

Definición 4.3.8 ***Distribución Bessel.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = \left(\frac{u^{1/2}}{\beta}\right)^\alpha k_\alpha\left(\frac{u^{1/2}}{\beta}\right), \qquad \alpha > -\frac{n}{2},\ \beta > 0, \tag{4.25}$$

donde $k_\alpha(\cdot)$ la función modificada de Bessel del tercer tipo, es decir,

$$k_\alpha(z) = \frac{\pi}{2}\frac{I_{-\alpha}(z) - I_\alpha(z)}{sen(\alpha\pi)}, \qquad |arg(z)| < \pi,\ \alpha = 0, \pm 1, \pm 2, \cdots,$$

donde

$$I_\alpha(z) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+\alpha+1)} \left(\frac{z}{2}\right)^{\alpha+2k}, \qquad |z| < \infty,$$

entonces se dice que $\underline{Y}$ tiene una distribución Bessel simétrica multivariante y se denota $\underline{Y} \sim Bessel_n(\alpha, \beta, \underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_8|\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right) \\ &= \frac{\left[(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right]^{\alpha/2}}{2^{\alpha+n-1}\pi^{n/2}\beta^{n+\alpha}|\Sigma|^{-1/2}\Gamma\left(\alpha+\frac{n}{2}\right)} k_\alpha\left(\frac{\left[(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right]^{1/2}}{\beta}\right). \end{aligned} \tag{4.26}$$

4.3.11. Distribución Laplace multivariante

Es el caso particular de la distribución Bessel simétrica multivariante con $\alpha = 0$ y $\beta = \frac{\sigma}{2}$ con $\sigma > 0$.

Definición 4.3.9 ***Distribución Laplace.*** *Si $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ y su función generadora es de la forma*

$$g(u) = k_0\left(\frac{\sqrt{2}u^{1/2}}{\sigma}\right), \qquad \sigma > 0, \tag{4.27}$$

donde $k_0(\cdot)$ es la función modificada de Bessel del tercer tipo dada en la Definición 4.3.8. Entonces se dice que el vector $\underline{Y}$ tiene una distribución Laplace multivariante y se denota $\underline{Y} \sim Laplace_n(\sigma, \underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_9|\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right) \\ &= \frac{1}{2^{n/2-1}\pi^{n/2}\sigma^n|\Sigma|^{-1/2}\Gamma\left(\frac{n}{2}\right)} k_0\left(\sqrt{2}\frac{\left[(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right]^{1/2}}{\sigma}\right). \end{aligned} \tag{4.28}$$

4.3.12. Distribución Exponencial potencia multivariante

Definición 4.3.10 ***Distribución Exponencial Potencia.*** *Si* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *y su función generadora es de la forma*

$$g(u) = e^{-u^{\alpha}/2}, \qquad u \geq 0, \tag{4.29}$$

entonces se dice que el vector $\underline{Y}$ *tiene una distribución Exponencial Potencia y se denota* $\underline{Y} \sim EP_n(\alpha, \underline{\mu}, \Sigma)$.

Su función de densidad es de la forma

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_1|\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right) \\ &= c_{10}|\Sigma|^{-1/2} exp\left\{ -\frac{\left[(\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right]^{\alpha}}{2} \right\} \end{aligned} \tag{4.30}$$

con

$$c_{10} = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{n/2}\int_0^{\infty} z^{n-1} g(z^2)\, dz}.$$

4.4. Estimación

Un teorema que será útil para el análisis de regresión lineal sumiendo una d.c.e. en los errores es el siguiente:

Teorema 4.4.1 *Supóngase que se tiene una observación del vector* $\underline{Y}$ *de distribución* $E_n(\underline{\mu}, \Sigma, \psi)$*, donde* $(\underline{\mu}, \Sigma) \in \Omega \subset \mathbb{R}^{n\times 1} \times \mathbb{R}^{n\times n}$*. Además, supóngase que* Ω *tiene la propiedad que si* $(Q, S) \in \Omega$ *con* $Q \in \mathbb{R}^{n\times 1}$ *y* $S \in \mathbb{R}^{n\times n}$*, entonces* $(Q, cS) \in \Omega$ *para cualquier* $c > 0$*. Del Teorema 4.2.3 se tiene que la f.d.p. de* $\underline{Y}$ *es (4.3) con* $g(\cdot)$ *la función generadora. Defínase* $l(z) = z^{n/2}g(z)$ *para* $z > 0$ *y sea* z_h *el valor en el que* $l(\cdot)$ *alcanza su máximo. Más aún, supóngase que bajo el supuesto de que* $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$*, con* $(\underline{\mu}, \Sigma) \in \Omega$*, los estimadores M.V. de* $\underline{\mu}$ *y* Σ *son* $\underline{\mu}^*$ *y* Σ^**, que son únicos y* Σ^* *es d.p. con probabilidad 1. Entonces, bajo la condición* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$*, con* $(\underline{\mu}, \Sigma) \in \Omega$*, los estimadores M.V. de* $\underline{\mu}$ *y* Σ *son* $\underline{\hat{\mu}}_E = \underline{\mu}^*$ *y* $\hat{\Sigma}_E = \frac{n}{z_h}\Sigma^*$ *y el valor máximo de esta verosimilitud es* $c_i|\hat{\Sigma}_E|^{-1/2} g(z_h)$.

Demostración:

Se definen

$$\Sigma_1 = \frac{\Sigma}{|\Sigma|^{1/n}} \tag{4.31}$$

y

$$z = (\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}). \tag{4.32}$$

Entonces,

$$\begin{aligned} z &= (\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) \\ &= (\underline{y} - \underline{\mu})^T [|\Sigma|^{1/n} \Sigma_1]^{-1} (\underline{y} - \underline{\mu}) \\ &= \frac{1}{|\Sigma|^{1/n}} (\underline{y} - \underline{\mu})^T \Sigma_1^{-1} (\underline{y} - \underline{\mu}). \end{aligned} \tag{4.33}$$

Así que la f.d.p. de $\underline{Y}$ se puede expresar como

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_i |\Sigma|^{-1/2} g\big\{(\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\big\} \\ &= \frac{c_i}{|\Sigma|^{1/2}} g(z) \\ &= \frac{c_i}{|\Sigma|^{1/2}} g(z) \frac{z^{n/2}}{z^{n/2}} \\ &= \frac{c_i}{|\Sigma|^{1/2}} g(z) \frac{z^{n/2}}{\left[\frac{1}{|\Sigma|^{1/n}} (\underline{y} - \underline{\mu})^T \Sigma_1^{-1} (\underline{y} - \underline{\mu})\right]^{n/2}} \\ &= c_i \; z^{n/2} g(z) \frac{|\Sigma|^{1/2}}{|\Sigma|^{1/2}} \left[(\underline{y} - \underline{\mu})^T \Sigma_1^{-1} (\underline{y} - \underline{\mu})\right]^{-n/2} \\ &= c_i \; z^{n/2} g(z) \left[(\underline{y} - \underline{\mu})^T \Sigma_1^{-1} (\underline{y} - \underline{\mu})\right]^{-n/2} \\ &= c_i \; l(z) \left[(\underline{y} - \underline{\mu})^T \Sigma_1^{-1} (\underline{y} - \underline{\mu})\right]^{-n/2}. \end{aligned}$$

De lo anterior se puede concluir que maximizar $f_{\underline{Y}}(\underline{y})$ es equivalente a maximizar $l(z) = z^{n/2} g(z)$ y $\left[(\underline{y} - \underline{\mu})^T \Sigma_1^{-1} (\underline{y} - \underline{\mu})\right]^{-n/2}$.

Ahora bien, si $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, entonces $g(z) = (2\pi)^{-n/2} e^{-z/2}$ (obsérvese la Definición 4.3.2) y

$$l(z) = (2\pi)^{-n/2} z^{n/2} e^{-z/2},$$

cuyo dominio es $z \leq 0$. Así que aplicando cálculo diferencial para obtener el valor que lo maximiza, se tiene

$$\frac{d}{dz} l(z) = (2\pi)^{-n/2} \left(z^{n/2} e^{-z/2} \left(-\frac{1}{2} \right) + \frac{n}{2} e^{-z/2} z^{\frac{n}{2}-1} \right) = \frac{1}{2} (2\pi)^{-n/2} z^{\frac{n}{2}-1} e^{-z/2} (n - z)$$

en consecuencia, $l(z)$ alcanza su máximo en $z_h = n$. De las condiciones del teorema, bajo normalidad $\left[(\underline{y} - \underline{\mu})^T \Sigma_1^{-1} (\underline{y} - \underline{\mu})\right]^{-n/2}$ se maximiza en $\underline{\mu} = \underline{\mu}^*$ y $\Sigma_1 = \Sigma_1^* = \Sigma^* / |\Sigma^*|^{1/n}$.

Dado que $(\underline{y}-\underline{\mu})^T\Sigma_1^{-1}(\underline{y}-\underline{\mu})$ no depende de g, en el caso de $\underline{Y}\sim E_n(\underline{\mu},\Sigma,\psi)$ también alcanza su máximo en $\underline{\mu}=\underline{\hat{\mu}}_E=\underline{\mu}^*$ y $\Sigma_1=\hat{\Sigma}_1=\Sigma_1^*$. Por otra parte, $l(z)$ se maximiza en $z_h=n$. Entonces, usando (4.31), (4.32) y (4.33) se tiene

$$\begin{aligned}\hat{\Sigma}_E &= |\hat{\Sigma}|^{1/n}\hat{\Sigma}_1\\ &= \frac{(\underline{y}-\underline{\hat{\mu}})^T\hat{\Sigma}_1^{-1}(\underline{y}-\underline{\hat{\mu}})}{z_h}\hat{\Sigma}_1\\ &= \frac{n}{z_h}\frac{(\underline{y}-\underline{\mu}^*)^T\Sigma_1^{*-1}(\underline{y}-\underline{\mu}^*)}{n}\Sigma_1^*\\ &= \frac{n}{z_h}\Sigma^*.\end{aligned}$$

Y el valor máximo de esta verosimilitud es $\frac{c_i}{|\hat{\Sigma}_E|^{1/2}}g(z_h)=c_i|\hat{\Sigma}_E|^{-1/2}g(z_h)$.

■

Conviene explicar el Teorema 4.4.1, éste señala que si $\underline{Y}\sim E_n(\underline{\mu},\Sigma,\psi)$, entonces los estimadores máximo verosímiles para $\underline{\mu}$ y Σ son los mismos que en el caso normal pues el estimador para $\underline{\mu}$ sería el mismo y para Σ sería un múltiplo. En otras palabras, para obtener los estimadores máximo verosímiles para $\underline{\mu}$ y Σ cuando $\underline{Y}\sim E_n(\underline{\mu},\Sigma,\psi)$, digamos $\underline{\hat{\mu}}_E$ y $\hat{\Sigma}_E$, primero se obtienen para el caso particular normal $\underline{Y}\sim N_n(\underline{\mu},\Sigma)$, digamos $\underline{\mu}^*$ y Σ^*, y luego se tiene que

$$\underline{\hat{\mu}}_E=\underline{\mu}^*$$

y

$$\hat{\Sigma}_E=\frac{n}{z_h}\Sigma^*$$

con z_h el valor que maximiza la función $l(z)=z^{n/2}g(z)$ con $z\leq 0$ y $g(\cdot)$ la función generadora de densidad de $\underline{Y}\sim E_n(\underline{\mu},\Sigma,\psi)$.

Además, obsérvese que si $g(\cdot)$ es continua y decreciente, entonces z_h existe y es positivo y finito. Para ciertas funciones se puede obtener analíticamente (por ejemplo, las distribuciones Normal, Exponencial potencia, t), sin embargo, para muchas otras solo a través de métodos numéricos (Díaz, 2003)[5].

4.5. Pruebas de hipótesis

En esta sección se presentan algunos teoremas útiles en el análisis de regresión en la parte de las pruebas de hipótesis de los estimadores. Para demostrarlos se requiere el siguiente lema demostrado en Gupta et al. (2013) pág. 139[9]:

Lema 4.5.1 *Sea* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *con* $P(\underline{Y} = \underline{0}) = 0$. *Asuma* $\underline{Z} \sim N_n(\underline{0}, \Sigma)$. *Sea* $\mathscr{F}$ *un subconjunto de vectores reales de* $n \times 1$, *tal que si* $A \in \mathbb{R}^{n\times 1}$, $A \in \mathscr{F}$ *y* $a > 0$, *entonces* $aA \in \mathscr{F}$ *y* $P(\underline{Y} \notin \mathscr{F}) = P(\underline{Z} \notin \mathscr{F}) = 0$. *Sea* $K(A)$ *una función definida en* $\mathscr{F}$, *tal que si* $A \in \mathscr{F}$ *y* $a > 0$, *entonces* $K(A) = K(aA)$. *Entonces,* $K(\underline{Y})$ *y* $K(\underline{Z})$ *son definidas con probabilidad 1 y* $K(\underline{Y})$ *y* $K(\underline{Z})$ *son idénticamente distribuidas.*

Teorema 4.5.2 *Supóngase que se tienen los valores de una observación* $\underline{y}$ *del vector aleatorio* $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, *donde* $(\underline{\mu}, \Sigma) \in \Omega \subset \mathbb{R}^{n\times 1} \times \mathbb{R}^{n\times n}$ *y se desea probar la hipótesis*

$$H_0 : (\underline{\mu}, \Sigma) \in \omega \quad vs \quad H_a : (\underline{\mu}, \Sigma) \in \Omega - \omega, \tag{4.34}$$

donde $\omega \subset \Omega$. *Supóngase que* Ω *y* ω *tienen la propiedad de que si* $\underline{Q} \in \mathbb{R}^{n\times 1}$ *y* $S \in \mathbb{R}^{n\times n}$, *entonces* $(\underline{Q}, S) \in \Omega$ *implica* $(\underline{Q}, cS) \in \Omega$ *y* $(\underline{Q}, S) \in \omega$ *implica* $(\underline{Q}, cS) \in \omega$ *para cualquier escalar positivo c. Además, la f.d.p. de* $\underline{Y}$ *es de la forma*

$$f_{\underline{Y}}(\underline{y}) = c_i|\Sigma|^{-1/2} g\left((\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu})\right),$$

donde $l(z) = z^{n/2} g(z)$ *con* $z > 0$ *alcanza su máximo finito en* $z = z_h > 0$. *Más aún, supóngase que bajo el supuesto de que* $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, $(\underline{\mu}, \Sigma) \in \Omega$, *los estimadores máximo verosímiles de* $\underline{\mu}$ *y* Σ *son* $\underline{\mu}^*$ *y* Σ^* *que son únicos y* Σ^* *es d.p. con probabilidad 1.*

Asúmase también que bajo el supuesto de que $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, $(\underline{\mu}, \Sigma) \in \omega$, *los estimadores máximo verosímiles de* $\underline{\mu}$ *y* Σ *son* $\underline{\mu}_0^*$ *y* Σ_0^* *que son únicos y* Σ_0^* *es d.p. con probabilidad 1.*

Entonces, la estadística del cociente de verosimilitudes para probar la hipótesis 4.34 bajo el supuesto $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ *es la misma que bajo el supuesto* $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$, *digamos* $\frac{|\Sigma^*|}{|\Sigma_0^*|}$.

Demostración:

Del Teorema 4.4.1 se tiene que bajo la condición $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, con $(\underline{\mu}, \Sigma) \in \Omega$, los estimadores máximo verosímiles de $\underline{\mu}$ y Σ son $\underline{\hat{\mu}}_E = \underline{\mu}^*$ y $\hat{\Sigma}_E = \frac{n}{z_h}\Sigma^*$ y el valor máximo de la verosimilitud es

$$c_i|\hat{\Sigma}_E^{-1/2}|g(z_h).$$

De forma similar, bajo la condición $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ con $(\underline{\mu}, \Sigma) \in \omega$, los estimadores máximo verosímiles de $\underline{\mu}$ y Σ son $\underline{\hat{\mu}}_{E0} = \underline{\mu}_0^*$ y $\hat{\Sigma}_{E0} = \frac{n}{z_h}\Sigma_0^*$ y el valor máximo de la verosimilitud es

$$c_i|\hat{\Sigma}_{E0}^{-1/2}|g(z_h).$$

Por lo tanto, la estadística del cociente de verosimilitudes es

$$\frac{c_i|\hat{\Sigma}_{E0}^{-1/2}|g(z_h)}{c_i|\hat{\Sigma}_{E}^{-1/2}|g(z_h)} = \frac{|\frac{n}{z_h}\Sigma_0^*|^{-1/2}}{|\frac{n}{z_h}\Sigma^*|^{-1/2}} = \left(\frac{|\Sigma^*|}{|\Sigma_0^*|}\right)^{1/2},$$

lo que es equivalente a la estadística $\frac{|\Sigma^*|}{|\Sigma_0^*|}$.

■

Teorema 4.5.3 *Supóngase que se tienen los valores de una observación $\underline{y}$ del vector aleatorio $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$, donde $(\underline{\mu}, \Sigma) \in \Omega = \Omega_1 \times \Omega_2$ con $\Omega_1 \subset \mathbb{R}^{n\times 1}$ y $\Omega_2 \subset \mathbb{R}^{n\times n}$ y se desea probar la hipótesis*

$$H_0 : (\underline{\mu}, \Sigma) \in \omega \quad vs \quad H_a : (\underline{\mu}, \Sigma) \in \Omega - \omega,$$

donde $\omega = \omega_1 \times \omega_2$, $\omega_1 \subset \Omega_1$ y $\omega_2 \subset \Omega_2$. Asúmase que $\underline{0} \in \omega_1$. Sea $f(\underline{z})$ una estadística de prueba tal que $f(c\underline{z}) = f(\underline{z})$ para cualquier escalar $c > 0$. Entonces:

(i) Si $f(\underline{z}) = f(\underline{z} - \underline{\mu})$ para cada $\underline{\mu} \in \omega_1$, entonces la distribución de $f(\underline{y})$ si la hipótesis nula es cierta, es la misma bajo $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ que bajo $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$.

(ii) Si $f(\underline{z}) = f(\underline{z} - \underline{\mu})$ para cada $\underline{\mu} \in \Omega_1$, entonces la distribución de $f(\underline{y})$ si la hipótesis nula es cierta, es la misma bajo $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ que bajo $\underline{Y} \sim N_n(\underline{\mu}, \Sigma)$.

Demostración:

(i) Sea $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ con $\underline{\mu} \in \omega_1$. Defínase $\underline{W} = \underline{Y} - \underline{\mu}$. Así que $\underline{W} \sim E_n(\underline{0}, \Sigma, \psi)$ y $f(\underline{Y}) = f(\underline{Y} - \underline{\mu}) = f(\underline{W})$. Entonces, la distribución de $f(\underline{Y})$ es la misma que la distribución de $f(\underline{W})$.

Del Lema 4.5.1 se tiene que la distribución de $f(\underline{W})$ es la misma bajo $\underline{W} \sim E_n(\underline{0}, \Sigma, \psi)$ que bajo $\underline{W} \sim N_n(\underline{0}, \Sigma)$. Dado que $f(\underline{W}) = f(\underline{W} + \underline{\mu})$ se da que la distribución de $f(\underline{W})$ es la mima bajo $\underline{W} \sim N_n(\underline{0}, \Sigma)$ que bajo $\underline{W} \sim N_n(\underline{\mu}, \Sigma)$.

(ii) Sea $\underline{Y} \sim E_n(\underline{\mu}, \Sigma, \psi)$ con $\underline{\mu} \in \Omega_1$. Defínase $\underline{W} = \underline{Y} - \underline{\mu}$. Entonces, $\underline{W} \sim E_n(\underline{0}, \Sigma, \psi)$ y la prueba puede completarse exactamente del mismo modo que la prueba de (i).

■

Comentando los teoremas anteriores, del Teorema 4.5.2 se sabe que la estadística del cociente de verosimilitud para probar una hipótesis nula sobre los parámetros $\underline{\mu}$ y Σ de una variable elíptica es la misma que en el caso normal, y el Teorema 4.5.3 dice que, cumpliendo con ciertas condiciones, la distribución de la estadística de prueba es la misma que en el caso normal.

Capítulo 5

El uso de las distribuciones elípticas en regresión

En este capítulo se desarrolla la teoría para el modelo lineal dado en la expresión (1.1) con los supuestos descritos en la sección 1.1.2 pero ahora definiendo la distribución de los errores aleatorios como una elíptica.

Además de la estimación de los parámetros, en este capítulo se presentan las propiedades de los estimadores y los cálculos para hacer el análisis de regresión, es decir, las pruebas de hipótesis.

5.1. Supuestos

Considérese el modelo de regresión lineal descrito en la sección 1.1 y los cinco supuestos dados en la sección 1.1.2.

Si se define la distribución de los errores como una perteneciente a la familia de d.c.e., es decir, si $\underline{\varepsilon} \sim E_n(\underline{0}, \sigma^2 I_n, \psi)$, entonces al modelo se le conoce como *modelo de regresión lineal elíptico* o *modelo de regresión lineal con errores contorneados elípticamente*.

Ahora bien, algunas consecuencias de asumir $\underline{\varepsilon} \sim E_n(\underline{0}, \sigma^2 I_n, \psi)$ son:

(i) $\varepsilon_i \sim E_1(0, \sigma^2, \psi)$, es decir, cada uno de los errores sigue una distribución elíptica con media 0 y varianza σ^2. Esto se concluye aplicando el Teorema 4.2.2 donde la partición es únicamente el elemento ε_i. Además, por el Teorema 4.2.4 se sabe que si $\varepsilon_i \sim E_1(0, \sigma^2, \psi)$, entonces $\mathbb{E}(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = c\sigma^2$ con $c = -2\psi'(0)$.

Nótese que esto armoniza con le supuesto (II) pues los errores son idénticamente distribuidos con media 0 y varianza finita.

(ii) Del Teorema 4.2.4 y del Corolario 4.2.4.1 se sabe que si $\underline{\varepsilon} \sim E_n(\underline{0}, \sigma^2 I_n, \psi)$, entonces $Cov(\varepsilon_i, \varepsilon_j) = 0$ y, en consecuencia, $Cor(\varepsilon_i, \varepsilon_j) = 0$.

Lo que concuerda con el supuesto (III) pues los errores son no correlacionados.

(iii) Bajo el supuesto (I) se tiene que $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, entonces por el Teorema 4.2.1 $\underline{Y} \sim E_n(X\underline{\beta}, \sigma^2 I_n, \psi)$.

Lo que satisface el supuesto (IV).

En conclusión, si se tiene el modelo lineal (1.1) o en su forma matricial (1.3) y bajo los supuestos descritos en 1.1.2, bien puede definirse la distribución conjunta de los errores como una d.c.e., esto es, $\underline{\varepsilon} \sim E_n(\underline{0}, \sigma^2 I_n, \psi)$, pues se satisfacen todos los supuestos.

5.2. Estimadores

Ahora, del Teorema 4.2.3 se sabe que la f.d.p. conjunta de $\underline{Y}$ es (4.3) y lo que se busca es, basados en una única observación de $\underline{Y}$, esto es $\underline{y}$, estimar los parámetros $\underline{\beta}$ y σ^2 que maximicen esta función. Estos estimadores se obtienen con el siguiente teorema.

Teorema 5.2.1 *Sean $X \in \mathbb{R}^{n\times p}$ de rango p con $p < n$ y $\underline{y} \in \mathbb{R}^{n\times 1}$ una matriz y un vector de valores observados, respectivamente. Entonces, la función (4.3) se maximiza en*

$$\underline{\hat{\beta}}_E = (X^T X)^{-1} X^T \underline{y} \tag{5.1}$$

y

$$\hat{\Sigma}_E = \hat{\sigma}_E^2 I_n,$$

con

$$\hat{\sigma}_E^2 = \frac{1}{z_h}(\underline{y} - X\hat{\beta}_E)^T(\underline{y} - X\hat{\beta}_E), \tag{5.2}$$

donde z_h el valor que maximiza la función $l(z) = z^{n/2}g(z)$, con $g(\cdot)$ la función generadora de $\underline{Y}$.

Demostración:

Si $\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I_n)$, entonces los estimadores M.V. para $\underline{\beta}$ y σ^2 son $\underline{\hat{\beta}}_{MV}$ y $\hat{\sigma}^2_{MV}$ dados por las expresiones (3.6) y (3.8), respectivamente (obsérvense el Teorema 3.2.1 y la Definición 3.2.1).

Entonces, del Teorema 4.4.1 se tiene que los estimadores M.V. para $\underline{\beta}$ y Σ son:

$$\underline{\hat{\beta}}_E = \underline{\hat{\beta}}_{MV} = (X^T X)^{-1} X^T \underline{y}$$

y

$$\begin{aligned}\hat{\Sigma}_E &= \frac{n}{z_h}\hat{\Sigma}_{MV} \\ &= \frac{n}{z_h}\hat{\sigma}^2_{MV} I_n \\ &= \frac{n}{z_h}\cdot\frac{1}{n}(\underline{Y} - X\underline{\hat{\beta}}_E)^T(\underline{Y} - X\underline{\hat{\beta}}_E)I_n \\ &= \frac{1}{z_h}(\underline{Y} - X\underline{\hat{\beta}}_E)^T(\underline{Y} - X\underline{\hat{\beta}}_E)I_n \\ &= \hat{\sigma}^2_E I_n,\end{aligned}$$

con $\hat{\sigma}^2_E = \frac{1}{z_h}(\underline{Y} - X\underline{\hat{\beta}}_E)^T(\underline{Y} - X\underline{\hat{\beta}}_E)$.

■

Definición 5.2.1 *Sean $\underline{y}$ y X los valores observados como se describen en la sección 1.1.1 y definiendo $\underline{\varepsilon} \sim E_n(\underline{0}, \sigma^2 I_n, \psi)$ lo que implica que $\underline{Y} \sim E_n(X\underline{\beta}, \sigma^2 I_n, \psi)$, entonces $\underline{\hat{\beta}}_E$ y $\hat{\sigma}^2_E$ dados por las expresiones (5.1) y (5.2), respectivamente, se conocen como los estimadores M.V. del modelo de regresión lineal con errores contorneados elípticamente.*

5.3. Propiedades de los estimadores

Teorema 5.3.1 *Los estimadores $\underline{\hat{\beta}}_E$ y $\hat{\sigma}^2_E$ como se describen en la Definición 5.2.1 tienen las siguientes propiedades:*

(i) $\underline{\hat{\beta}}_E \sim E_p(\underline{\beta}, \sigma^2(X^TX)^{-1}, \psi)$.

(ii) $z_h\hat{\sigma}^2_E \sim G_{1,1}(\sigma^2 I_n, \frac{n-p}{2}, \psi)$.

(iii) $\underline{\hat{\beta}}_E$ *es un estimador insesgado para* $\underline{\beta}$.

(iv) $\hat{\sigma}^2_E$ *es un estimador sesgado para* σ^2.

Demostración:

(i)

Por el Teorema 4.2.1 se tiene que

$$\underline{\hat{\beta}}_E \sim E_p(X^TX)^{-1}X^TX(\underline{\beta}, (X^TX)^{-1}X^T(\sigma^2 I_n)X(X^TX)^{-1}, \psi) = E_p(\underline{\beta}, \sigma^2(X^TX)^{-1}, \psi).$$

Recuérdese que del Teorema A.0.4 se sabe que $(X^TX)^{-1}$ es d.p.

(ii)

Primero, obsérvese que $z_h\hat{\sigma}_E^2$ puede expresarse de la forma

$$z_h\hat{\sigma}_E^2 = \underline{\varepsilon}^T(I_n - H)\underline{\varepsilon},$$

con H la matriz dada en la expresión (2.10). Para ello se sustituye $\underline{\hat{\beta}}_E$ en $Z_h\hat{\sigma}_E^2$, entonces queda

$$\begin{aligned} z_h\hat{\sigma}_E^2 &= (\underline{Y} - X\underline{\hat{\beta}}_E)^T(\underline{Y} - X\underline{\hat{\beta}}_E) \\ &= (\underline{Y} - X(X^TX)^{-1}X^T\underline{Y})^T(\underline{Y} - X(X^TX)^{-1}X^T\underline{Y}) \\ &= \underline{Y}^T(I_n - H)(I_n - H)\underline{Y} \\ &= \underline{Y}^T(I_n - H)\underline{Y}, \end{aligned}$$

pues del Teorema 2.1.2 se sabe que H es simétrica y, en consecuencia, $I_n - H$ también lo es. Luego, sustituyendo $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ queda

$$z_h\hat{\sigma}_E^2 = (X\underline{\beta} + \underline{\varepsilon})^T(I_n - H)(X\underline{\beta} + \underline{\varepsilon}),$$

distribuyendo las multiplicaciones y considerando que $HX = X$ y $X^TH = X^T$, queda

$$\begin{aligned} Z_h\hat{\sigma}_E^2 &= (\underline{\beta}^TX^T - \underline{\beta}^TX^TH - \underline{\varepsilon}^T + \underline{\varepsilon}^TH)(X\underline{\beta} + \underline{\varepsilon}) \\ &= \underline{\varepsilon}^T(-I_n + H)(X\underline{\beta} + \underline{\varepsilon}) \\ &= \underline{\varepsilon}^T(-X\underline{\beta} + HX\underline{\beta} + \underline{\varepsilon} - H\underline{\varepsilon} \\ &= \underline{\varepsilon}^T(I_n - H)\underline{\varepsilon}. \end{aligned} \tag{5.3}$$

Dado que el vector $\underline{\varepsilon} \sim E_n(\underline{0}, \sigma^2I_n, \psi)$ y $(I_n - H)$ es simétrica e idempotente con $rango(I_n - H) = traza(I_n - H) = traza(I_n) - traza(H) = n - p$, por el Teorema 4.2.7 se concluye que

$$z_h\hat{\sigma}_E^2 \sim G_{1,1}(\sigma^2I_n, \frac{n-p}{2}, \psi).$$

(iii)

Dado que ya se probó que $\underline{\hat{\beta}}_E \sim E_p(\underline{\beta}, \sigma^2(X^TX)^{-1}, \psi)$ y por el Teorema 4.2.4 (i) se sabe que $\mathbb{E}(\underline{\hat{\beta}}_E) = \underline{\beta}$ y, por lo tanto, $\underline{\hat{\beta}}_E$ es un estimador insesgado para $\underline{\beta}$.

(iv)

Dado en (ii) se probó que $z_h\hat{\sigma}_E^2$ puede expresarse de la forma $z_h\hat{\sigma}_E^2 = \underline{\varepsilon}^T(I_n - H)\underline{\varepsilon}$ con $\underline{\varepsilon} \sim E_n(\underline{0}, \sigma^2I_n, \psi)$ y $I_n - H$ idempotente. Entonces, del Teorema 4.2.4 (iv) se tiene que

$$\mathbb{E}(\hat{\sigma}_E^2) = \mathbb{E}\left(\frac{1}{z_h}\underline{\varepsilon}^T(I_n - H)\underline{\varepsilon}\right)$$

$$
\begin{aligned}
&= \frac{c}{z_h} traza\left((I_n - H)(\sigma^2 I_n)\right) + \frac{1}{z_h}\underline{0}^T(I_n - H)\underline{0} \\
&= \frac{c}{z_h}\sigma^2 traza\,(I_n - H) \\
&= \frac{c}{z_h}\sigma^2(n-p),
\end{aligned}
$$

con $c = -2\psi'(0)$, por lo tanto, $\hat{\sigma}^2_E$ es un estimador sesgado para σ^2.

■

De la demostración anterior se puede proponer un estimador insesgado para σ^2, digamos $\hat{\sigma}^2_{EU}$, cuya expresión es:

$$\hat{\sigma}^2_{EU} = \frac{z_h}{c(n-p)}\hat{\sigma}^2_E = \frac{z_h}{c(n-p)} \cdot \frac{\underline{Y}^T(I_n - H)\underline{Y}}{z_h} = -\frac{1}{2\psi'(0)}\underline{Y}^T(I_n - H)\underline{Y}. \tag{5.4}$$

Obsérvese que

$$\mathbb{E}(\hat{\sigma}^2_{EU}) = \mathbb{E}\left(\frac{z_h}{c(n-p)}\hat{\sigma}^2_E\right) = \frac{z_h}{c(n-p)}\,\mathbb{E}(\hat{\sigma}^2_E) = \frac{z_h}{c(n-p)} \cdot \frac{c}{z_h}\sigma^2(n-p) = \sigma^2.$$

Obsérvese que, considerando $\hat{\sigma}^2$ (el estimador insesgado de σ^2 suponiendo normalidad en los errores) dado en la expresión (3.12) y $c = -2\psi'(0)$, entonces el estimador $\hat{\sigma}^2_{EU}$ dado en la expresión (5.4) puede escribirse como $\hat{\sigma}^2_{EU} = \frac{1}{c}\hat{\sigma}^2$, o de forma equivalente $\hat{\sigma}^2 = c\hat{\sigma}^2_{EU}$. Ahora bien, en la sección 5.1 se explicó que $\underline{Y} \sim E_n(X\underline{\beta}, \sigma^2 I_n, \phi)$ y por el Teorema 4.2.4 se sabe que $Var(\underline{Y}) = c\sigma^2 I_n$ así que el estimador $\hat{\sigma}^2$ en realidad estima la varianza del vector, y $\hat{\sigma}^2_{EU}$ estima el parámetro σ^2 de la densidad $E_n(X\underline{\beta}, \sigma^2 I_n, \phi)$. En otras palabras, si los errores tienen una d.c.e. el estimador $\hat{\sigma}^2_{EU}$ estima el parámetro σ^2 y el estimador $\hat{\sigma}^2$ estima la varianza del vector aleatorio $\underline{Y}$.

También es importante mencionar que dado que el estimador M.V. del modelo de regresión lineal elíptico es el mismo que el de mínimos cuadrados, es decir, $\hat{\underline{\beta}}_E = \hat{\underline{\beta}}$ (compárense las expresiones (2.7) y (5.1)), entonces por el Teorema 2.2.2 es BLUE, pues el teorema se demostró para alguna distribución sin definir ninguna en particular.

5.4. Pruebas de hipótesis

Como se comentó en la sección 3.5, luego de estimar los parámetros desconocidos sigue analizar los resultados.

Considérese el modelo de regresión lineal elíptico justo como se describe en la sección 5.1. Supóngase que se tiene una observación $\underline{y}$ de la variable aleatoria $\underline{Y} \sim E_n(X\underline{\beta}, \sigma^2 I_n, \psi)$ y que se desea probar la hipótesis:

$$H_0 : K^T\underline{\beta} = \underline{m} \qquad vs \qquad H_a : K^T\underline{\beta} \neq \underline{m}, \tag{5.5}$$

con K^T una matriz de $s \times p$ y de rango s, y $\underline{m}$ un vector de $s \times 1$.

Esta prueba puede hacerse a través del cociente de verosimilitudes, defínase la letra λ para expresarlo

$$\lambda = \frac{\sup_{\underline{\beta} \in \mathbb{R}^{p\times 1} \ s.a. \ H_0} f_{\underline{Y}}(\underline{y}; X, \underline{\beta}, \sigma^2)}{\sup_{\underline{\beta} \in \mathbb{R}^{p\times 1}} f_{\underline{Y}}(\underline{y}; X, \underline{\beta}, \sigma^2)}.$$

Nótese que en el denominador del cociente se está maximizando a la verosimilitud sobre todo el espacio parametral y se sabe que ese supremo se obtiene evaluándola en los parámetros M.V. Por otro lado, en el numerador se maximiza la verosimilitud en todo el espacio parametral sujeta a H_0, es decir, sujeta a $K^T\underline{\beta} = \underline{m}$.

Del Teorema 4.5.2 se tiene que la estadística de prueba del cociente de verosimilitudes de una d.c.e. es la misma que en el caso normal. Luego, en el caso normal el máximo sobre todo el espacio parametral es:

$$\begin{aligned} L_{\underline{\beta} \in \mathbb{R}^{p\times 1}}(\underline{y}; X, \hat{\underline{\beta}}, \hat{\sigma}^2_{MV}) &= (2\pi\hat{\sigma}^2_{MV})^{-n/2} exp\Big\{ -\frac{1}{2\hat{\sigma}^2_{MV}}(\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})\Big\} \\ &= (2\pi\hat{\sigma}^2_{MV})^{-n/2} exp\Big\{ -\frac{n}{2}\Big\}, \end{aligned}$$

obsérvense el Teorema 3.2.1 y la Definición 3.2.1. Y el máximo sobre todo el espacio parametral sujeto a $K^T\underline{\beta} = \underline{m}$ es:

$$\begin{aligned} L_{\underline{\beta} \in \mathbb{R}^{p\times 1} \ s.a. \ K^T\underline{\beta} = \underline{m}}(\underline{y}; X, \hat{\underline{\beta}}_R, \hat{\sigma}^2_{MVR}) &= (2\pi\hat{\sigma}^2_{MVR})^{-n/2} exp\Big\{ -\frac{1}{2\hat{\sigma}^2_{MVR}}(\underline{y} - X\hat{\underline{\beta}}_R)^T(\underline{y} - X\hat{\underline{\beta}}_R)\Big\} \\ &= (2\pi\hat{\sigma}^2_{MVR})^{-n/2} exp\Big\{ -\frac{n}{2}\Big\}, \end{aligned}$$

según lo probado en la sección 3.4, obsérvense las expresiones (3.15) y (3.16).

Entonces, el cociente de verosimilitudes queda:

$$\lambda = \frac{(2\pi\hat{\sigma}^2_{MVR})^{-n/2} exp\{-\frac{n}{2}\}}{(2\pi\hat{\sigma}^2_{MV})^{-n/2} exp\{-\frac{n}{2}\}} = \left(\frac{\hat{\sigma}^2_{MV}}{\hat{\sigma}^2_{MVR}}\right)^{n/2} = \left(\frac{(\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})}{(\underline{y} - X\hat{\underline{\beta}}_R)^T(\underline{y} - X\hat{\underline{\beta}}_R)}\right)^{n/2},$$

que sustituyendo $\hat{\underline{\beta}}_R$ por (3.16), H por (2.10) y recordando que $(\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})$ es equivalente a la expresión $\underline{y}(I_n - H)\underline{y}$ (obsérvese la demostración de (ii) del Teorema 5.3.1), λ puede simplificarse y expresarse como

$$\lambda = \left(\frac{\underline{y}(I_n - H)\underline{y}}{\underline{y}(I_n - H)\underline{y} + (K^T\hat{\underline{\beta}} - \underline{m})^T \left[K^T(X^TX)^{-1}K\right]^{-1}(K^T\hat{\underline{\beta}} - \underline{m})}\right)^{n/2}, \qquad (5.6)$$

que es la misma expresión que se muestra en el artículo de Díaz-García et al., 2003[5].

Si sucede que H_0 es cierta, entonces $K^T\underline{\beta} \approx \underline{m}$ lo que implica $\lambda \approx 1$. Por tal motivo, valores pequeños de λ llevan a rechazar H_0.

Nótese que existe una relación entre λ y F dadas por las expresiones (5.6) y (3.25), respectivamente, pues (Seber y Lee, 2003)[13]:

$$F = \frac{n-p}{s}(\lambda^{-2/n} - 1), \tag{5.7}$$

la cual se sabe que se distribuye $F_{s,n-p}$ si H_0 es cierta y $\underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I_n)$.

Ahora bien, si $\underline{\beta}$ y X se particionan de la forma $\underline{\beta} = \begin{bmatrix}\underline{\beta}_1 \\ \underline{\beta}_2\end{bmatrix}$ y $X = \begin{bmatrix}X_1 & X_2\end{bmatrix}$, con $\underline{\beta}_1$ y $\underline{\beta}_2$ vectores de dimensiones $s \times 1$ y $(p-s) \times 1$, respectivamente, y X_1 y X_2 matrices de dimensiones $n \times s$ y $n \times (p-s)$, respectivamente. Y si K^T es tal que $K^T\underline{\beta} = \underline{\beta}_1$, entonces la hipótesis:

$$H_0 : K^T\underline{\beta} = \underline{0} \qquad vs \qquad H_a : K^T\underline{\beta} \neq \underline{0}, \tag{5.8}$$

es equivalente a la prueba 5.5 pues si se define $\underline{Y}^* = \underline{Y} - X_1\underline{m}$ por el Teorema 4.2.1 se tiene que

$$\underline{Y}^* \sim E_n(X\underline{Y} - X_1\underline{m}, \sigma^2 I_n, \psi) =\sim E_n\left(\begin{bmatrix}X_1 & X_2\end{bmatrix}\begin{bmatrix}\underline{\beta}_1 - \underline{n} \\ \underline{\beta}_2\end{bmatrix}, \sigma^2 I_n, \psi\right),$$

y $\underline{\beta}_1 - \underline{m} = 0$ equivale a $\underline{\beta}_1 = \underline{m}$ y esto es $K^T\underline{\beta} = \underline{m}$.

Díaz-García et al. (2003)[5] mencionan en su artículo que las hipótesis (5.5) y (5.8) son equivalentes para toda matriz K^T genere o no una partición del vector $\underline{\beta}$.

Dado que las pruebas (5.5) y (5.8) son equivalentes el análisis se puede enfocar en (5.8). Obsérvese que la estadística F, ya sea expresada como (5.7) o como (3.25), vista como una función de la observación $\underline{y}$, digamos $f(\underline{y}) = F$, satisface que

$$f(c\underline{y}) = f(\underline{y}) \qquad \forall\, c > 0, \qquad y \qquad f(\underline{y} - \underline{0}) = f(\underline{y}),$$

por lo tanto, se satisfacen las condiciones de la parte (i) del Teorema 4.5.3 (dado que $\underline{0} \in \omega_1$) y se puede concluir que dada una observación $\underline{y}$ de $\underline{Y} \sim E_n(X\underline{\beta}, \sigma^2 I_n, \psi)$ la estadística para probar la hipótesis (5.5) es[1]:

$$F = \frac{n-p}{s}(\lambda^{-2/n} - 1) \sim F_{s,n-p}.$$

Como concluyen Gupta et al., 2013[9]: "las estadísticas de prueba de razón de verosimilitudes son las mismas que las desarrolladas en la teoría de la distribución normal y sus distribuciones bajo la hipótesis nula y regiones críticas son también iguales que en el caso normal".

[1]Galea et al. (2000)[8] concluyen lo mismo pero basándose en el Teorema 22.2 de Fang et al. (1990)[7] pág. 51 que establece que si $\underline{y}$ de $\underline{Y} \sim E_n(\underline{0}, I_n, \psi)$ y sea $T(\underline{Y})$ una estadística, entonces $T(k\underline{Y}) \overset{d}{=} T(\underline{Y})$ $\forall\, k > 0$ y $T(\underline{Y}) \overset{d}{=} T(\underline{Z})$ con $\underline{Z} \sim N_n(\underline{0}, I_n)$ donde el operador $\overset{d}{=}$ indica igualdad en distribución.

5.5. Aplicaciones

Una aplicación importante de los modelos de regresión es la predicción de nuevas observaciones que corresponden a específicos valores de las variables explicativas. Dado un conjunto de valores observados de X, la distribución de futuros valores respuesta de $\underline{Y}$ de un modelo estadístico es conocida como distribución de predicción. La inferencia basada en la distribución de predicción es conocida como inferencia predictiva. Al lector interesado en profundizar en este tema se le sugiere el artículo *"Predictive Inference for the Elliptical Linear Model"* de Kibria y Haq (1998)[10] donde concluyen, a través de probabilidades condicionales, que la distribución de predicción de respuestas futuras del modelo lineal con errores elípticos sigue una t-Student multivariante.

Capítulo 6

Simulaciones

En este capítulo se generan valores aleatorios centrados en cero con distribución elíptica diferente a la normal, esto es, $\varepsilon \sim E_1(0, \sigma^2, \psi)$ y con ellos se simula un modelo como se describe en la sección 1.1.2. El objetivo de esta simulación es, a través de varios ejemplos, comparar la inferencia estadística de la regresión lineal, por un lado, sabiendo que los errores son elípticos y, por otro, ignorando ésto y asumiendo erróneamente normalidad.

De ahí que se simulan dos de las distribuciones definidas en la sección 4.3 que son la Exponencial Potencia (4.3.10) y la t-Student multivariada (4.3.4). Para cada una de ellas se seleccionan dos valores distintos de σ^2 tales que el primero genere una distribución de colas pesadas y el segundo una de colas ligeras.

Ahora bien, los datos que se comparan son, primeramente, la estimación de σ^2, es decir, $\hat{\sigma}^2$ y $\hat{\sigma}^2_{EU}$ dados por las expresiones (3.12) y (5.4), luego, las pruebas de hipótesis para los parámetros desconocidos, esto es, $H_0 : \beta_i = 0$ vs $H_a : \beta_i \neq 0 \; \forall \; i$ que, asumiendo normalidad, se ocupa la prueba 6 descrita en la sección 3.5 (en los sucesivo se referirá a ésta como *prueba normal*) y, en el caso elíptico, se ocupa la descrita en la sección 5.4 (en los sucesivo se referirá a ésta como *prueba elíptica*). El experimento tiene el siguiente diseño:

1. Se definen los parámetros $\beta_0 = -5$ y $\beta_1 = 11$, y los vectores $\underline{x}_1$ y $\underline{x}_2$ de variables explicativas[1]. Estos valores se mantienen fijos durante el análisis.

2. Se define el valor de σ^2. Para ambas distribuciones, la Exponencial Potencia y la t-Student, se ocupó $\sigma^2 = \{19, 1/19\}$.

3. Se generan n errores aleatorios con distribución $E_1(0, \sigma^2, \psi)$ y con ellos se obtiene el vector de la variable respuesta con el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i,$$

[1] $\underline{x}_1$ se simuló como el valor absoluto de una v.a. $N(100, 10)$ y $\underline{x}_2$ como el *arc cos* de una v.a. $Unif(-1, 1)$.

con $i = 1, 2, \ldots, n$ y x_{1i} el i-ésimo elemento del vector $\underline{x_1}$. Obsérvese que no se ocupa $\underline{x_2}$ en el modelo. Conviene mencionar que esta forma de generar aleatoriamente los valores satisface los supuestos descritos en la sección 1.1.2.

4. Se ajusta un modelo de regresión lineal con $\underline{x}_1$ y $\underline{x}_2$ como variables explicativas, es decir, se ajusta:
$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$
En consecuencia, se obtiene $\underline{\hat{\beta}} = (\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2)^T$ ocupando la expresión (2.7). Se estiman $\hat{\sigma}^2$ y $\hat{\sigma}^2_{EU}$. Se obtienen los $p - value$ de las tres pruebas $H_0 : \beta_j = 0$ para $j = 0, 1, 2$ asumiendo normalidad. Por último, respetando la distribución elíptica, se calculan los tres cocientes de verosimilitudes para probar $H_0 : K^T\underline{\beta} = \underline{m}$ con $K^T = \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ y $\underline{m} = 0$.

   Nótese que, debido a cómo fue simulado el vector de observaciones en el paso anterior, se espera que las hipótesis $H_0 : \beta_0 = 0$ y $H_0 : \beta_1 = 0$ se rechacen, pero no así $H_0 : \beta_2 = 0$.

5. Se repiten m veces los pasos 3 y 4.

En otras palabras, se repite m veces el mismo experimento de generar n observaciones aleatorias con las cuales se simula la variable respuesta y, entonces, se estiman los parámetros desconocidos y se hacen las pruebas de hipótesis.

Las simulaciones se hicieron con $n = \{30, 100, 500, 1000\}$ observaciones y $m = \{100, 300, 1000\}$ repeticiones, esto con el fin de mantener conclusiones objetivas. Sin embargo, dado que los resultados son parecidos entre los diferentes números de repeticiones, únicamente se reportan los resultados obtenidos para una m, digamos $m = 300$.

Por último, conviene explicar cómo se resume la información en las gráficas:

- Para las gráficas de la estimación de σ^2: dado que cada estimación con n observaciones se está repitiendo m veces, en esta gráfica se presenta la media de la estimación de σ^2 para cada n, y en las barras se muestra una desviación estándar.
- Para las gráficas de las pruebas de hipótesis: se muestra el cociente del número de decisiones correctas asumiendo errores elípticos sobre normales. Por ejemplo, dado que se sabe que $\beta_0 \neq 0$, pues desde el principio así está definido, entonces si la prueba de hipótesis indica que hay evidencia para rechazar $H_0 : \beta_0 = 0$, se dice que la decisión es correcta, y de igual forma con la $H_0 : \beta_1 = 0$, caso contrario, con β_2 pues se sabe que en el modelo vale 0, entonces si no se rechaza la hipótesis nula, se está tomando la decisión correcta. Ahora bien, en las gráficas se muestra el número de decisiones correctas con la prueba de hipótesis elíptica sobre la prueba normal. Así que, si el cociente es mayor 1, se dice que la prueba suponiendo a los errores con distribución elíptica es más adecuada, pues es mayor el número de veces que se toma la decisión correcta; si es igual a 1, entonces ambas pruebas resultan en la misma decisión; y si es menor a 1, entonces es mejor la prueba asumiendo normalidad.

6.1. Modelo de regresión lineal con errores con distribución Exponencial Potencia

Considérese la distribución Exponencial Potencia definida en la sección 4.3.10 y tomada del artículo de Galea et al. (2000)[8].

Se define para la simulación a los errores como $\varepsilon_i \sim EP_1(1/2, 0, \sigma^2)$, es decir, $\alpha = 1/2$, $\mu = 0$ y $\Sigma = \sigma^2$, entonces su función generadora es:

$$g(u) = \frac{1}{4} e^{-\frac{\sqrt{u}}{2}},$$

y la constante de normalización se obtiene de la expresión (4.4):

$$c_{10} = \frac{\Gamma\frac{1}{2}}{2\pi^{1/2} \int_0^\infty y^{1-1} g(y^2)\, dy} = \frac{1}{2\int_0^\infty e^{-\frac{|y|}{2}}\, dy} = \frac{1}{2 \cdot 2} = \frac{1}{4},$$

entonces la función de densidad es[2]

$$\begin{aligned} f_\varepsilon(x) &= c_{10}|\Sigma|^{-1/2} g((x-\mu)^T \Sigma^{-1}(x-\mu)) \\ &= \frac{1}{4} \cdot \frac{1}{\sqrt{\sigma^2}} \cdot e^{-\frac{1}{2}\sqrt{\frac{x^2}{\sigma^2}}} \\ &= \frac{1}{4\sqrt{\sigma^2}} e^{-\frac{1}{2\sqrt{\sigma^2}}|x|}, \end{aligned} \tag{6.1}$$

la cual se puede simular con la función `rLaplace()` de la librería `ExtDist`[3] con parámetros `mu = 0` y `b =` $2\sqrt{\sigma^2}$.

Ahora bien, para el cálculo del estimador $\hat{\sigma}^2_{EU}$ se requiere conocer z_h y $c = -2\psi'(0)$.

Por un lado, se sabe que la función característica de la distribución Laplace[4] es:

$$\phi_\varepsilon(t) = \frac{exp(\mu i t)}{1 + b^2 t^2} = exp(\mu i t)\psi(\sigma^2 t^2),$$

con $\psi(u) = (1 + 4u)^{-1}$, entonces

$$c = -2\psi'(0) = -2(-4)(1+4u)^{-2}\Big|_{u=0} = 8.$$

Y, por otro lado, en el Teorema 5.2.1 se da la definición de la función $l(\cdot)$ y de z_h, entonces

$$l(z) = z^{1/2} g(z) = z^{1/2} e^{-\frac{\sqrt{z}}{2}} \; con \; z > 0,$$

[2]Se puede comprobar que efectivamente integra 1 para $x \in \mathbb{R}$.

[3]Obsérvese la documentación de la librería páginas 30 y 31[3].

[4]Aunque a la función de densidad 6.1 se le llama Exponencial Potencia en el artículo de Galea et. al (2000)[8], también es conocida como Laplace.

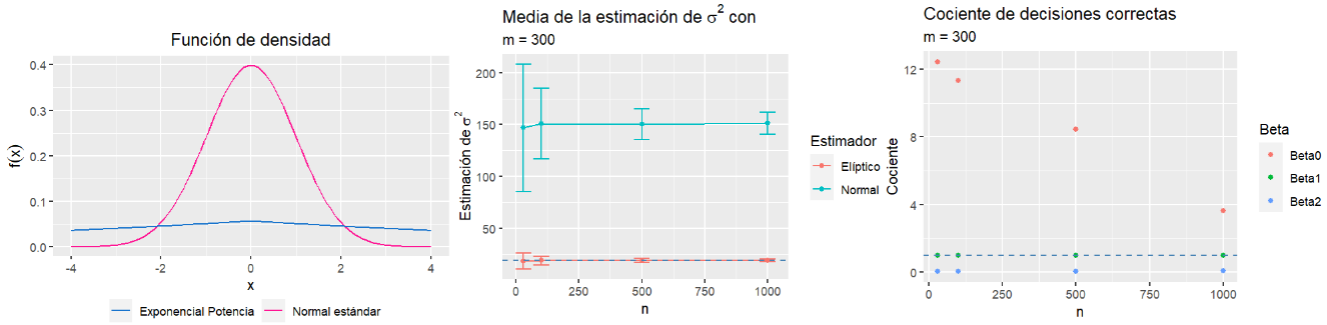

Figura 6.1: Izquierda: Funciones de densidad de las distribuciones $EP_1(1/2, 0, 19)$ y $N(0, 1)$, en azul y rosa respectivamente. Centro: Media de m valores de σ^2 estimados con n observaciones generadas con errores $\varepsilon \sim EP_1(1/2, 0, 19)$, en azul se estimó asumiendo normalidad, es decir, con la expresión (3.12) y en rosa asumiendo una distribución elíptica, es decir, con la expresión (5.4). Las barras señalan una desviación estándar. Derecha: Cociente del número de decisiones correctas tomadas bajo realizar las pruebas de hipótesis elípticas sobre las normales.

con z_h el valor que la maximiza, así que z_h satisface que

$$\frac{d}{dz} l(z)\Big|_{z=z_h} = -\frac{e^{-\frac{\sqrt{z}}{2}}(\sqrt{z_h} - 2)}{\sqrt{z_h}} = 0,$$

por lo tanto, en $z_h = 4$ se obtiene el máximo de $l(\cdot)$.

Ahora bien, como se comentó en la introducción se seleccionaron dos valores para σ^2. A continuación, se muestran los resultados obtenidos.

CASO 1: Simulación de colas pesadas con $\sigma^2 = 19$. Se simularon errores con la función de densidad dada en (6.1) con parámetro $\sigma^2 = 19$. En la Figura 6.1 se muestra, primeramente, su gráfica comparada con la función de densidad normal estándar, lo que permite verificar de forma visual que, efectivamente, se trata de una distribución de colas pesadas. Con respecto a la estimación de σ^2 en la gráfica central de la Figura 6.1 se concluye que a mayor número de observaciones menor es la varianza de las estimaciones y que con $n \geq 100$ observaciones se mantiene constante el valor estimado. Se puede ver que $\hat{\sigma}^2_{EU}$ da una estimación muy precisa, sin embargo, $\hat{\sigma}^2$ sobre estima el valor del parámetro. Y, en la última gráfica, se observa que la prueba elíptica es considerablemente mejor para el parámetro β_0, sin embargo, para el parámetro β_1 se concluye exactamente lo mismo en la prueba elíptica que en la normal, y, por último, con el parámetro β_2 la prueba normal resulta mejor. También se puede notar que a mayor número de observaciones (n) las decisiones tienden a ser iguales bajo las dos pruebas para el parámetro β_0, es decir, se nota un acercamiento a 1.

CASO 2: Simulación de colas ligeras $\sigma^2 = 1/19$. Se simularon errores con la función de densidad dada en (6.1) con parámetro $\sigma^2 = 1/19$. En la Figura 6.2 se muestra, primeramente, su gráfica comparada con la función de densidad normal estándar, lo que

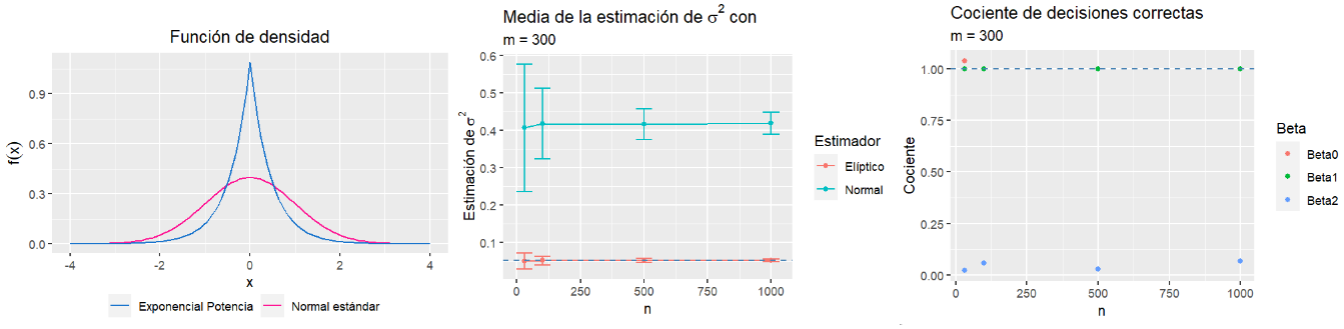

Figura 6.2: Izquierda: Funciones de densidad de las distribuciones $EP_1(1/2, 0, 1/19)$ y $N(0, 1)$, en azul y rosa respectivamente. Centro: Media de m valores de σ^2 estimados con n observaciones generadas con errores $\varepsilon \sim EP_1(1/2, 0, 1/19)$, en azul se estimó asumiendo normalidad, es decir, con la expresión (3.12) y en rosa asumiendo una distribución elíptica, es decir, con la expresión (5.4). Las barras señalan una desviación estándar. Derecha: Cociente del número de decisiones correctas tomadas bajo realizar las pruebas de hipótesis elípticas sobre las normales.

permite verificar de forma visual que, efectivamente, se trata de una distribución de colas ligeras. Con respecto a la estimación de σ^2 en la gráfica central de la Figura 6.2 se concluye que a mayor número de observaciones menor es la varianza de las estimaciones y que con $n \geq 100$ observaciones se mantiene constante el valor estimado. Se puede ver que $\hat{\sigma}^2_{EU}$ da una estimación muy precisa, sin embargo, $\hat{\sigma}^2$ sobrestima el valor del parámetro. Y, en la gráfica de la derecha, se observa que tanto la prueba elíptica como la normal son iguales[5] para los parámetros β_0 y β_1 y, por último, con el parámetro β_2 la prueba normal resulta mejor.

6.2. Modelo de regresión lineal con errores con distribución t-Student multivariante

Considérese la distribución elíptica t-Student multivariante definida en la sección 4.3.4.

Se define a los errores como $\varepsilon_i \sim t_1(3, 0, \sigma^2)$, es decir, $m = 3$, $\mu = 0$ y $\Sigma = \sigma^2$, entonces su función generadora es

$$g(u) = \left(1 + \frac{u}{3}\right)^{-\frac{3+1}{2}},$$

y la constante de normalización es:

$$c_4 = \frac{\Gamma\left(\frac{3+1}{2}\right)}{(3\pi)^{1/2}\Gamma\left(\frac{3}{2}\right)} = \frac{2}{\pi\sqrt{3}},$$

[5]Para $n = \{100, 500, 1000\}$ el cociente de β_0 también da 1, en la gráfica no se aprecia porque los puntos se encimaron.

entonces la función de densidad es[6]:

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= c_4|\Sigma|^{-1/2} g\left((\underline{y}-\underline{\mu})^T\Sigma^{-1}(\underline{y}-\underline{\mu})\right) \\ &= \frac{2}{\pi\sqrt{3}}\cdot\frac{1}{\sqrt{\sigma^2}}\left(1+\frac{x^2/\sigma^2}{3}\right)^{-\frac{3+1}{2}} \\ &= \frac{2}{\pi\sqrt{3\sigma^2}}\left(1+\frac{x^2}{3\sigma^2}\right)^{-2}, \end{aligned} \tag{6.2}$$

la cual se puede simular con la función `rpearsonVII` de la librería `PearsonDS`[7] con parámetros `df = 3`, `location = 0` y `scale =` $\sqrt{\sigma^2}$.

Ahora bien, para el cálculo del estimador $\hat{\sigma}^2_{EU}$ se requiere conocer z_h y $c = -2\psi'(0)$.

Por un lado, la función característica de la distribución t-Student multivariante cuando los grados de libertad son número impar es (Sutradhar, 1986)[14]:

$$\phi_\varepsilon(t) = \frac{\sqrt{\pi}\,\Gamma\left(\frac{m+1}{2}\right) exp\left(\mu it - \sqrt{m}\sqrt{\sigma^2t^2}\right)}{2^{m-1}\Gamma\left(\frac{m}{2}\right)} \times \sum_{r=1}^{d}\left[\binom{2d-r-1}{d-r}\,\frac{\left(2\sqrt{m}\,\sqrt{t^2\sigma^2}\right)^{r-1}}{(r-1)!}\right],$$

donde $d = \frac{m+1}{2}$. Sustituyendo $m = 3$, desarrollando la suma y simplificando algebraicamente la expresión resulta

$$\begin{aligned} \phi_\varepsilon(t) &= exp(\mu it)\left(\frac{1+\sqrt{3}\sqrt{\sigma^2t^2}}{exp(\sqrt{3}\sqrt{\sigma^2t^2})}\right) \\ &= exp(it\mu)\psi(\sigma^2t^2), \end{aligned}$$

con $\psi(u) = \frac{1+\sqrt{3u}}{exp(\sqrt{3u})}$, entonces

$$c = -2\psi'(0) = -2\frac{d}{du}\frac{1+\sqrt{3u}}{exp(\sqrt{3u})}\bigg|_{u=0} = 3.$$

Por otro lado, en el Teorema 5.2.1 se da la definición de la función $l(\cdot)$ y de z_h, entonces

$$l(z) = z^{1/2}g(z) = z^{1/2}\left(1+\frac{u}{3}\right)^{-2} \; con\; z > 0,$$

con z_h el valor que la maximiza, así que z_h satisface que

$$\frac{d}{dz}l(z)\bigg|_{z=z_h} = -\frac{27(z-1)}{2\sqrt{z}(z+3)^3} = 0,$$

por lo tanto, en $z_h = 1$ se obtiene el máximo de $l(\cdot)$.

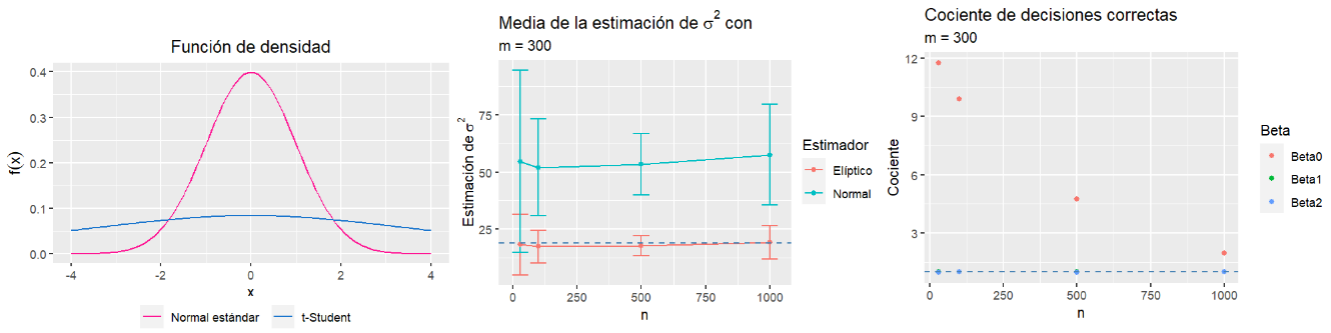

Figura 6.3: Izquierda: Funciones de densidad de las distribuciones $t_1(3, 0, 19)$ y $N(0, 1)$, en azul y rosa respectivamente. Centro: Media de m valores de σ^2 estimados con n observaciones generadas con errores $\varepsilon \sim t_1(3, 0, 19)$, en azul se estimó asumiendo normalidad, es decir, con la expresión (3.12) y en rosa asumiendo una distribución elíptica, es decir, con la expresión (5.4). Las barras señalan una desviación estándar. Derecha: Cociente del número de decisiones correctas tomadas bajo realizar las pruebas de hipótesis elípticas sobre las normales.

Ahora bien, como se comentó en la introducción se seleccionaron dos valores para σ^2. A continuación, se muestran los resultados obtenidos.

CASO 1: Simulación de colas pesadas con $\sigma^2 = 19$. Se simularon errores con la función de densidad dada en (6.2) con parámetro $\sigma^2 = 19$. En la Figura 6.3 se muestra, primeramente, su gráfica comparada con la función de densidad normal estándar, lo que permite verificar de forma visual que, efectivamente, se trata de una distribución de colas pesadas. Con respecto a la estimación de σ^2 en la gráfica central de la Figura 6.3 se ve que para el tamaño de muestra $n = 30$ observaciones la varianza de la estimación es bastante grande y, en general, los valores estimados varían sin importar el número de datos. Se puede ver que $\hat{\sigma}^2_{EU}$ da una estimación muy precisa, sin embargo, $\hat{\sigma}^2$ sobrestima el valor del parámetro. Y, en la última gráfica, se observa que la prueba elíptica es considerablemente mejor para el parámetro β_0 y conforme menor es el número de observaciones, más adecuada es, sin embargo, para los parámetros β_1 y β_2 se concluye exactamente lo mismo en la prueba elíptica que en la normal[8] pues el cociente da 1. También se puede notar que a mayor número de observaciones (n) las decisiones tienden a ser iguales bajo las dos pruebas para el parámetro β_0, es decir, se nota un acercamiento a 1.

CASO 2: Simulación de colas ligeras $\sigma^2 = 1/19$. Se simularon errores con la función de densidad dada en (6.2) con parámetro $\sigma^2 = 1/19$. En la Figura 6.4 se muestra, primeramente, su gráfica comparada con la función de densidad normal estándar, lo que permite verificar de forma visual que, efectivamente, se trata de una distribución de colas ligeras. Con respecto a la estimación de σ^2 en la gráfica central de la Figura 6.4 se ve que

[6]Se puede comprobar que efectivamente integra 1 para $x \in \mathbb{R}$.

[7]Obsérvese en la documentación de la librería las páginas 24 y 25[4].

[8]Los puntos verdes correspondientes al cociente de β_1 quedaron debajo de los puntos azules, por ese motivo no se aprecian en la gráfica.

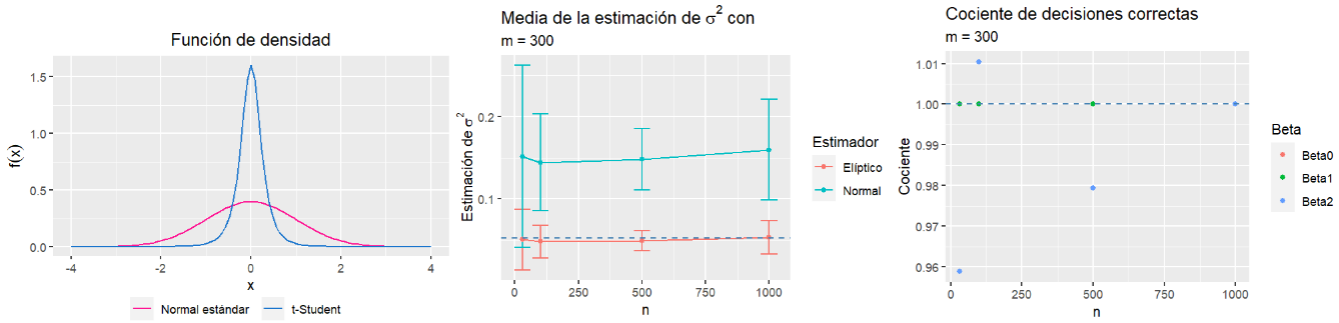

Figura 6.4: Izquierda: Funciones de densidad de las distribuciones $t_1(3, 0, 1/19)$ y $N(0, 1)$, en azul y rosa respectivamente. Centro: Media de m valores de σ^2 estimados con n observaciones generadas con errores $\varepsilon \sim t_1(3, 0, 1/19)$, en azul se estimó asumiendo normalidad, es decir, con la expresión (3.12) y en rosa asumiendo una distribución elíptica, es decir, con la expresión (5.4). Las barras señalan una desviación estándar. Derecha: Cociente del número de decisiones correctas tomadas bajo realizar las pruebas de hipótesis elípticas sobre las normales.

para el tamaño de muestra $n = 30$ observaciones la varianza de la estimación es bastante grande y, en general, los valores estimados varían sin importar el número de datos. Se puede ver que $\hat{\sigma}^2_{EU}$ da una estimación muy precisa, sin embargo, $\hat{\sigma}^2$ sobrestima el valor del parámetro. Y, en la gráfica de la derecha, se observa que tanto la prueba elíptica como la normal son iguales[9] para β_0 y β_1. Para el parámetro β_2 en general resulta mejor la prueba normal que la elíptica.

[9]Nuevamente los puntos rojos correspondientes al cociente de β_0 quedaron debajo de los puntos verdes y por eso no se aprecian en la gráfica.

Capítulo 7

Conclusiones

Considerando el modelo de regresión lineal dado en la expresión (1.1) o en forma matricial (1.3) sujeto a los supuestos descritos en la sección 1.1.2, se puede concluir que:

- El estimador para $\underline{\beta}$ es igual por mínimos cuadrados (2.7), que por máxima verosimilitud asumiendo normalidad en los errores (3.6) o asumiendo en los errores una distribución elíptica (5.1).
- Los dos estimadores máximo verosímiles para σ^2, es decir, $\hat{\sigma}^2_{MV}$ y $\hat{\sigma}^2_{E}$, el primero asumiendo normalidad y dado en la expresión (3.8) y el segundo asumiendo una distribución elíptica y dado en la expresión (5.2), son ambos sesgados.
- Los estimadores insesgados para σ^2 obtenidos a partir de los máximo verosímiles, es decir, $\hat{\sigma}^2$ y $\hat{\sigma}^2_{EU}$, el primero obtenido a partir de $\hat{\sigma}^2_{MV}$ y dado en la expresión (3.12) y el segundo obtenido a partir de $\hat{\sigma}^2_{E}$ y dado en la expresión (5.4), son múltiplo uno del otro pues pueden expresarse como $\hat{\sigma}^2_{EU} = \frac{1}{c}\hat{\sigma}^2$ con $c = -2\psi'(0)$ y $\psi(\cdot)$ como se define en 4.1.1. Ya que en el caso normal la varianza de los errores es el parámetro σ^2 presente en la f.d.p., es decir, estos dos valores son iguales, puede pensarse que $\hat{\sigma}^2$ estima, de hecho, la varianza de los errores, sin embargo, si la distribución es elíptica estos dos valores no son iguales (obsérvese el Teorema 4.2.4) y $\hat{\sigma}^2_{EU}$ estima el parámetro σ^2, no la varianza.
- Con respecto a las pruebas de hipótesis se puede concluir que, dado que en la sección 5.4 se prueba que en el caso elíptico la estadística F para probar la hipótesis $H_0 : A\underline{\beta} = \underline{c}$, con A y $\underline{c}$ de las dimensiones apropiadas, se distribuye $F_{s,n-p}$ justo como en el caso normal, entonces se puede concluir que aún si los errores no siguen una distribución normal pero sí una distribución simétrica (obsérvese la sección 4.3.1) puede ocuparse la típica prueba F en el modelo de regresión. En otras palabras, si se está haciendo un análisis de regresión en el cual los errores no siguen una distribución normal pero sí una simétrica, puede entonces ocuparse la estadística F para realizar las pruebas de hipótesis sobre los parámetros $\underline{\beta}$ del modelo.

Apéndice A

Algunos resultados de álgebra matricial

Las siguientes definiciones, propiedades y teoremas son útiles para el desarrollo de la teoría en el presente trabajo.

Definición A.0.1 *Una matriz* Σ *de* $n \times n$ *se dice definida positiva (d.p.) si es simétrica y además* $\underline{y}^T \, \Sigma \, \underline{y} > 0$ *para todo vector* $\underline{y}$ *de* $n \times 1$ *con* $\underline{y} \neq \underline{0}$.

Donde $\underline{y}^T$ denota la transpuesta de $\underline{y}$.

Teorema A.0.1 *Si* Σ *es una matriz d.p., entonces* Σ^{-1} *existe y es d.p.*

Demostración:

Primero, si Σ es definida positiva, entonces por definición (A.0.1) $\underline{v}^T \, \Sigma \, \underline{v} > 0$ para todo $\underline{v} \neq \underline{0}$, lo que implica que $\Sigma \, \underline{v} \neq 0$ para todo $\underline{v} \neq \underline{0}$, así que Σ es de rango completo y, por lo tanto, es invertible.

Luego, para toda matriz A invertible de $n \times n$ sucede $(A^{-1})^T = (A^T)^{-1}$. Esto es claro pues nótese que $(A^T(A^{-1})^T)^T = A^{-1}A = I_n$, con I_n la matriz identidad de dimensión $n \times n$, entonces $A^T(A^{-1})^T = I_n^T = I_n$ así que $(A^{-1})^T$ es la inversa de A^T, esto es $(A^{-1})^T = (A^T)^{-1}$.

De lo anterior, si A es simétrica e invertible, entonces $(A^{-1})^T = (A^T)^{-1} = A^{-1}$ y, por lo tanto, A^{-1} es simétrica.

Por último, queda por mostrar que Σ^{-1} es definida positiva. Considere $\underline{v}^T \, \Sigma^{-1} \, \underline{v}$ para cualquier $\underline{v} \neq \underline{0}$

$$\begin{aligned} \underline{v}^T \, \Sigma^{-1} \, \underline{v} &= \underline{v}^T \, \Sigma^{-1} \Sigma \Sigma^{-1} \, \underline{v} \\ &= (\Sigma^{-1}\underline{v})^T \Sigma (\Sigma^{-1}\underline{v}) \\ &> 0, \end{aligned}$$

la igualdad se da pues Σ es simétrica e invertible, entonces $(\Sigma^{-1})^T = (\Sigma^T)^{-1} = \Sigma^T$, y la última desigualdad del hecho de que Σ es definida positiva.

■

Teorema A.0.2 *Sea Σ una matriz definida positiva, existe una raíz cuadrada definida positiva $\Sigma^{1/2}$, tal que $(\Sigma^{1/2})^2 = \Sigma$.*

Demostración:

Sea $\Sigma = TDT^T$ la descomposición espectral de Σ, donde $D = diag(\lambda_1, \ldots, \lambda_n)$ es una matriz diagonal con λ_i los eigenvalores de Σ, y T es una matriz ortogonal, esto es que $T^T = T^{-1}$. Dado que Σ es d.p. $\lambda_i > 0 \ \forall\, i$, por lo tanto los elementos de la diagonal de D son todos positivos y se puede definir $D^{1/2} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ y $\Sigma^{1/2} = TD^{1/2}T^T$.

Ahora bien, sucede que

$$\begin{aligned}\Sigma^{1/2}\Sigma^{1/2} &= (TD^{1/2}T^T)(TD^{1/2}T^T)\\ &= TD^{1/2}T^TTD^{1/2}T^T\\ &= TD^{1/2}I_nD^{1/2}T^T\\ &= TDT^T\\ &= \Sigma.\end{aligned}$$

pues $T^TT = I_n$.

Obsérvese que $\Sigma^{1/2}$ también es simétrica e invertible.

■

Teorema A.0.3 *Si Σ es una matriz de $n \times n$ d.p. y C una matriz de $p \times n$ de rango p, entonces $C\Sigma C^T$ es d.p.*

Demostración:

$\underline{x}^TC\Sigma C^T\underline{x} = \underline{y}^T\Sigma\underline{y} \geq 0$ con igualdad $\iff \underline{y} = \underline{0} \iff C^T\underline{x} = 0 \iff \underline{x} = \underline{0}$ (dado que las columnas de C^T son linealmente independientes pues C es de rango p). Por lo tanto, $\underline{x}^TC\Sigma C^T\underline{x} > 0 \ \forall\, \underline{x} \neq \underline{0}$.

■

Teorema A.0.4 *Sea $X \in \mathbb{R}^{n\times p}$ con $p < n$, y sean sus columnas linealmente independientes, es decir, $rango(X) = p$, entonces X^TX es p.d. y, por lo tanto, invertible.*

Demostración:

Dado que $X \in \mathbb{R}^{n\times p}$, sucede que $X^TX \in \mathbb{R}^{p\times p}$. Ahora bien, para cualquier vector $\underline{v} \in \mathbb{R}^{p\times 1}$ se tiene

$$\underline{v}^T(X^TX)\underline{v} = (X\underline{v})^T(X\underline{v}) = ||X\underline{v}||^2.$$

Obsérvese que $X\underline{v}$ es una combinación lineal de las columnas de X y los coeficientes de la combinación lineal son los componentes del vector $\underline{v}$. Así que $X\underline{v} = \underline{0}$ sólo si $v = \underline{0}$, pues por hipótesis las columnas de X son linealmente independientes. Así que

$$\underline{v}^T(X^TX)\underline{v} = (X\underline{v})^T(X\underline{v}) = ||X\underline{v}||^2 > 0,$$

$\forall\ \underline{v} \neq \underline{0}$ y por la Definición A.0.1 X^TX es d.p.

Por último, por el Teorema A.0.1 se tiene que $(X^TX)^{-1}\ \exists$.

■

Apéndice B

Interpretación geométrica del estimador $\underline{\hat{\beta}}$

El estimador de $\underline{\beta}$, es decir, $\underline{\hat{\beta}}$ también se puede obtener mediante argumentos geométricos. Para ello considérese la notación matricial dada en la sección 1.1.1.

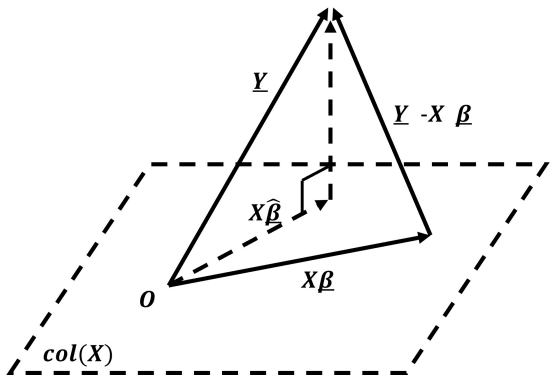

Figura B.1: Representación geométrica de la estimación del vector $\underline{\beta}$ [13].

El razonamiento es el siguiente: del supuesto (V) sabemos que las columnas de la matriz $X \in \mathbb{R}^{n\times p}$ son linealmente independientes, entonces cada columna puede pensarse como un vector y, en consecuencia, las p columnas de X son una base, es decir, generan un espacio vectorial en $\mathbb{R}^{n\times 1}$, digamos $col(X) = \{\underline{v} \mid \underline{v} = X\underline{a}\ \forall\ \underline{a} \in \mathbb{R}^{n\times 1}\}$.

Dado que $X\underline{\beta}$ es

$$X_0\beta_0 + X_1\beta_1 + \cdots + X_k\beta_k,$$

con $X_i \in \mathbb{R}^{n\times 1}$ la i-ésima columna de X y $\beta_i \in \mathbb{R}$ el i-ésimo elemento del vector $\underline{\beta}$ con $i = 0, 1, \ldots, k$, entonces $X\underline{\beta}$ es, de hecho, una combinación lineal de las columnas de X y, por lo tanto, pertenece al espacio vectorial $col(X)$, esto es $X\underline{\beta} \in col(X)$.

Ahora bien, se busca es un vector $\underline{\hat{\beta}} \in \mathbb{R}^{n\times 1}$ tal que $||\underline{Y} - X\underline{\hat{\beta}}||^2$ (el cuadrado de la distancia entre $\underline{Y}$ y $X\underline{\hat{\beta}}$) sea mínimo, es decir, se busca un vector $X\underline{\hat{\beta}}$ del espacio vectorial $col(X)$ tal que la distancia al vector $\underline{Y}$ sea mínima y esto se logra si $\underline{Y} - X\underline{\hat{\beta}}$ es ortogonal (o perpendicular en el caso de dos y tres dimensiones) al espacio $col(X)$, es decir, si

$\underline{Y} - X\underline{\hat{\beta}} \perp X$, en la Figura B.1 se ilustra para el caso de 3 dimensiones. Entonces,

$$\begin{aligned} X^T(\underline{Y} - X\underline{\hat{\beta}}) &= 0 \qquad \textit{por ortogonalidad} \\ X^T\underline{Y} - X^TX\underline{\hat{\beta}} &= 0 \\ X^TX\underline{\hat{\beta}} &= X^T\underline{Y} \\ \underline{\hat{\beta}} &= (X^TX)^{-1}X^T\underline{Y}. \end{aligned}$$

Y se obtiene la misma expresión dada en (2.7).

Ejemplo. Dos dimensiones: Sean $\underline{Y} = \begin{pmatrix} 5 \\ -1 \end{pmatrix}$, $X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, y $\underline{\beta} = \beta_0 \in \mathbb{R}$, entonces el modelo es:

$$\mathbb{E}(\underline{Y}) = X\underline{\beta} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \beta_0.$$

Observando la Figura B.2, el espacio vectorial generado por X es la línea roja pues es el conjunto $col(X) = \left\{ r \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} r \\ r \end{pmatrix} \middle| r \in \mathbb{R} \right\}$, así que $X\underline{\hat{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_0 \end{pmatrix}$ efectivamente es un vector sobre la línea. Luego, $\underline{Y}$ es el vector en azul.

Se busca escoger $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_0 \end{pmatrix}$ cuya distancia al vector $\begin{pmatrix} 5 \\ -1 \end{pmatrix}$ sea mínima, y resulta ser aquel que genera un segmento perpendicular a la línea roja; dado que ésta tiene pendiente igual a 1, un segmento perpendicular tiene pendiente -1, entonces:

$$\frac{-1 - \hat{\beta}_0}{5 - \hat{\beta}_0} = -1 \qquad \therefore\ \hat{\beta}_0 = 2,$$

así que el vector[1] $X\underline{\hat{\beta}} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ efectivamente pertenece a $col(X)$ y minimiza la distancia $||\underline{Y} - X\underline{\beta}||^2$.

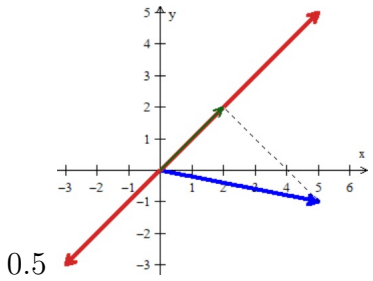

Figura B.2: En rojo se muestra el espacio generado por las columnas de X, en azul el vector $\underline{Y}$ y en verde el vector $\underline{\hat{\beta}}$.

[1]Nótese que el valor de $\hat{\beta}_0$ resulta el mismo si se ocupa la fórmula (2.7).

Apéndice C

Tabla: familia de distribuciones elípticas

Tabla C.1: Distribuciones multivariadas contorneadas elípticamente.

Distribución	Notación	Función generadora	Referencia
Tipo Kotz	$K_n(r, s, N, \underline{\mu}, \Sigma)$	$g(u) = u^{N-1}exp(-ru^s)$, $con\ r, s > 0,\ 2N + n > 2$	[4.3.1]
Normal	$N_n(\underline{\mu}, \Sigma)$	$g(u) = exp(-u/2)$	[4.3.2]
Pearson tipo VII	$MPVII_n(N, m, \underline{\mu}, \Sigma)$	$g(u) = \left(1 + \frac{u}{m}\right)^{-N}$, $con\ N > \frac{n}{2},\ m > 0$	[4.3.3]
t-Student	$t_n(m, \underline{\mu}, \Sigma)$	$g(u) = \left(1 + \frac{u}{m}\right)^{-\frac{m+n}{2}}$, $con\ u \geq 0$	[4.3.4]
Cauchy	$C_n(\underline{\mu}, \Sigma)$	$g(u) = (1 + u)^{-(n+1)/2},\ con\ u \geq 0$	[4.3.5]
Pearson tipo II	$MPII_n(m, \underline{\mu}, \Sigma)$	$g(u) = (1 - u)^m$, $con\ 0 \leq u \leq 1,\ m > -1$	[4.3.6]
Logística	$L_n(\underline{\mu}, \Sigma)$	$g(u) = \frac{e^{-u}}{(1 + e^{-u})^2},\ con\ u \geq 0$	[4.3.7]
Bessel	$Bessel_n(\alpha, \beta, \underline{\mu}, \Sigma)$	$g(u) = \left(\frac{u^{1/2}}{\beta}\right)^{\alpha} k_\alpha \left(\frac{u^{1/2}}{\beta}\right)$, $con\ \alpha > -\frac{n}{2},\ \beta > 0$	[4.3.8]
Laplace	$Laplace_n(\sigma, \underline{\mu}, \Sigma)$	$g(u) = k_0 \left(\frac{\sqrt{2}u^{1/2}}{\sigma}\right)$, $con\ \sigma > 0$	[4.3.9]
Exponencial Potencia	$EP_n(\alpha, \underline{\mu}, \Sigma)$	$g(u) = e^{-u^\alpha/2}$, $con\ u \geq 0$	[4.3.10]

Bibliografía

[1] Theodore Anderson. *An Introduction to Multivariate Statistical Analysis*. Inglés. 3.ª ed. Hoboken, NJ, Estados Unidos: Wiley, 2003.

[2] Andrius Buteikis. *4.4 Restricted Least Squares | Practical Econometrics and Data Science*. 7 de ago. de 2018. URL: http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-4-Multiple-RLS.html (visitado 03-2022).

[3] Comprehensive R Archive Network (CRAN). *CRAN - Package ExtDist*. 7 de dic. de 2020. URL: https://cran.r-project.org/web/packages/ExtDist/ExtDist.pdf (visitado 03-2022).

[4] Comprehensive R Archive Network (CRAN). *CRAN - Package PearsonDS*. 29 de mar. de 2022. URL: https://cran.r-project.org/web/packages/PearsonDS/PearsonDS.pdf (visitado 18-04-2022).

[5] José A. Díaz-García, Manuel Galea Rojas y Víctor Leiva-Sánchez. «Influence Diagnostics for Elliptical Multivariate Linear Regression Models». En: *Communications in Statistics - Theory and Methods* 32.3 (2003), págs. 625-641. DOI: 10.1081/STA-120018555. URL: https://doi.org/10.1081/STA-120018555.

[6] Michael F. Driscoll. «An Improved Result Relating Quadratic Forms and Chi-Square Distributions». En: *The American Statistician* 53.3 (1999), págs. 273-275. ISSN: 00031305. URL: http://www.jstor.org/stable/2686110.

[7] Kai-Tai Fang, Samuel Kotz y Kai Wang Ng. *Symmetric Multivariate and Related Distributions*. Inglés. 1.ª ed. Abingdon, Reino Unido: Taylor Francis, 1990.

[8] Manuel Galea, Marco Riquelme y Gilberto A. Paula. «DIAGNOSTIC METHODS IN ELLIPTICAL LINEAR REGRESSION MODELS». En: *Brazilian Journal of Probability and Statistics* 14.2 (2000), págs. 167-184. ISSN: 01030752, 23176199. URL: http://www.jstor.org/stable/43600976.

[9] Arjun Gupta, Tamas Varga y Taras Bodnar. *Elliptically Contoured Models in Statistics and Portfolio Theory*. Inglés. 2.ª ed. New York, Estados Unidos: Springer Publishing, 2013. DOI: 10.1007/978-1-4614-8154-6.

[10] B.M.Golam Kibria y M.Safiul Haq. «Predictive Inference for the Elliptical Linear Model». En: *Journal of Multivariate Analysis* 68.2 (1999), págs. 235-249. ISSN: 0047-259X. DOI: https://doi.org/10.1006/jmva.1998.1792. URL: https://www.sciencedirect.com/science/article/pii/S0047259X98917924.

[11] Douglas Montgomery, Elizabeth Peck y Geoffrey Vining. *Introduction to Linear Regression Analysis*. Inglés. 5.ª ed. Hoboken, NJ, Estados Unidos: Wiley, 2012.

[12] Shayle Searle. *Linear Models*. Inglés. Hoboken, NJ, Estados Unidos: Wiley, 1997.

[13] George Seber y Alan Lee. *Linear Regression Analysis*. Inglés. 2.ª ed. Hoboken, NJ, Estados Unidos: Wiley, 2003. URL: https://books.google.com.mx/books?id=X2Y6OkXl8ysC&lpg=PP1&dq=seber&pg=PP1#v=onepage&q=seber&f=false.

[14] Brajendra C. Sutradhar. «On the Characteristic Function of Multivariate Student t-Distribution». En: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 14.4 (1986), págs. 329-337. ISSN: 03195724. URL: http://www.jstor.org/stable/3315191 (visitado 18-04-2022).