UNIVERSIDAD NACIONAL DE COLOMBIA

Modelado de cuantiles marginales en presencia de datos faltantes mediante la clase de modelos de regresión con distribución normal/independiente multivariada

Jose Antonio Escobar Arias

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2024

Modelado de cuantiles marginales en presencia de datos faltantes mediante la clase de modelos de regresión con distribución normal/independiente multivariada

Jose Antonio Escobar Arias

Tesis de grado presentada como requisito parcial para optar al título de:
Magister en Ciencias Estadística

Director:
Mauricio Alejandro Mazo Lopera
Doctor en Ciencias - Estadística

Línea:
Modelado de Cuantiles

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2024

Las cosas más grandiosas -grandes pensamientos descubrimientos, inventos- generalmente han sido nutridos en la privación, a menudos ponderadas por la tristeza y finalmente establecidas con dificultad.

Samuel Smiles

Hay una cualidad que uno debe poseer para ganar, y esa es la determinación del propósito, el conocimiento de lo que uno quiere y un ardiente deseo de lograrlo.

Napoleon Hill

La estadística es un problema de datos faltantes.

Roderick J.A. Little

Agradecimientos

Quiero expresar mi profundo agradecimiento a mis padres por el incansable esfuerzo que realizaron para contribuir a mi formación y convertirme en la persona que soy hoy. Además, quiero agradecer a todas las personas que me brindaron su apoyo durante esta etapa de mi vida.

A la Escuela de Estadística de la Universidad Nacional de Colombia sede Medellín y a sus profesores, les estoy agradecido por su invaluable contribución a mi educación. En particular, deseo destacar y reconocer el excepcional trabajo y dedicación de mi tutor, el profesor Mauricio Alejandro Mazo Lopera. Su excepcional trabajo y dedicación han sido fundamentales para mi desarrollo académico. La orientación y el apoyo brindados por él no solo han enriquecido mis conocimientos, sino que también han inspirado mi crecimiento personal. También quiero expresar mi gratitud al profesor Juan Carlos Correa Morales, cuyo enfoque pedagógico me guió en numerosas situaciones, permitiéndome comprender y superar desafíos académicos de manera efectiva.

No puedo dejar de reconocer la contribución del profesor Raúl Alejandro Morán Vásquez de la Universidad de Antioquia. Sus aportes significativos han dejado una marca indeleble en mi desarrollo académico y personal. Su influencia ha sido un pilar crucial en mi formación, y estoy agradecido por la oportunidad de haber aprendido de su experiencia y conocimientos.

Resumen

En este trabajo de investigación, se propone el desarrollo de un modelo de regresión lineal con respuesta multivariada asociado a la clase de distribuciones normal/independiente multivariadas. El objetivo principal es lograr el modelado de cuantiles marginales bajo la presencia de datos faltantes, teniendo en cuenta la asociación entre las variables del vector de respuesta. Se emplea un enfoque Bayesiano, aprovechando las herramientas que este ofrece, como también algoritmos (que serán descritos posteriormente) para llevar a cabo el proceso de imputación y aproximación de distribuciones posteriores. La validez del modelo se evalúa mediante estudios de simulación, que confirman el desempeño satisfactorio en el proceso de estimación de los parámetros. Además, se presenta una aplicación práctica del modelo a un conjunto de datos reales, proporcionando así una validación adicional de su utilidad y aplicabilidad en contextos empíricos.

Palabras clave: Algoritmo de Aumento de Datos Monótonos (MDA Algorithm), distribución normal/independiente multivariada, distribución log-normal/independiente multivariada, modelado de cuantiles, datos faltantes, regresión lineal multivariada

Abstract

Modeling of marginal quantiles in the presence of missing data using the class of regression models with normal/independent multivariate distribution

In this research work, we propose the development of a multivariate linear regression model associated with the class of normal/independent multivariate distributions. The primary objective is to achieve modeling of marginal quantiles in the presence of missing data, considering the association among variables in the response vector. A Bayesian approach is employed, leveraging the tools offered by this approach, including algorithms (which will be described later) for imputation and posterior distribution calculations. The model's validity is assessed through simulation studies, confirming the satisfactory performance of parameters estimation. Additionally, a practical application of the model to a real dataset is presented, providing further validation of its utility and applicability in empirical contexts.

Keywords: Monotone Data Augmentation Algorithm (MDA Algorithm), multivariate normal/independent distribution, multivariate log-normal/independent distribution, quantile modeling, missing data, multivariate linear regression

Contenido

Lista de Figuras

Lista de Tablas

1 Introducción

Los conjuntos de datos multivariados continuos positivos son frecuentemente encontrados en el ámbito práctico. Es ampliamente reconocido que estos datos, al ser continuos y positivos, tienden a exhibir asimetría positiva y ocasionalmente contienen observaciones atípicas, como se ha señalado en estudios anteriores (Ferrari y Fumes, 2017).

A pesar de estas características distintivas, el análisis estadístico de tales conjuntos de datos a menudo se sustenta en los supuestos de la distribución normal multivariada (Morán-Vásquez, Mazo-Lopera, y Ferrari, 2021). Este enfoque puede resultar limitado, ya que tiende a pasar por alto las particularidades inherentes a este tipo de datos (Morán-Vásquez y Ferrari, 2019). Una metodología alternativa para modelar datos positivos multivariados involucra la aplicación de la transformación Box-Cox (Box y Cox, 1964) a cada componente del vector de observaciones, sin embargo, esta suposición implica una deficiencia teórica (Morán-Vásquez y Ferrari, 2019).

Ferrari y Fumes (2017), propusieron la clase de distribuciones Box-Cox simétricas, que resulta útil para modelizar datos sesgados positivos, posiblemente con colas pesadas, en el ámbito univariado. Morán-Vásquez y Ferrari (2019), extendieron la clase de distribuciones simétricas Box-Cox al ámbito multivariado, a la cual llamaron clase de distribuciones Box-Cox elípticas. Una característica distintiva y práctica de la clase Box-Cox elíptica radica en la interpretabilidad de sus parámetros, que se vinculan con cuantiles y dispersiones relativas de las distribuciones marginales y las asociaciones entre pares de variables (Morán-Vásquez y Ferrari, 2019).

Una subclase significativa dentro de las distribuciones Box-Cox elípticas (Morán-Vásquez y Ferrari, 2019), también considerada como una subclase de las distribuciones log-elípticas, utilizadas para la modelización de conjuntos de datos multivariados positivos, es la clase de distribuciones log-normal/independiente (nombre que se le atribuye, debido a que resulta de combinar dos variables aleatorias independientes; una de estas variables sigue una distribución log-normal, mientras que la otra es una variable aleatoria positiva) multivariadas (Morán-Vásquez y cols., 2021). Según Morán-Vásquez y cols. (2021), esta subclase es especialmente atractiva para la modelización estadística robusta, ya que ofrece diversas distribuciones con colas pesadas y soporte positivo, además, presentan propiedades teóricas específicas que no se cumplen en toda la clase log-elíptica. La clase de distribuciones log-normal/independiente multivariadas es apropiada para modelar datos positivos multiva-

riados correlacionados que son sesgados y posiblemente de cola pesada, la cual, tiene como algunos de sus miembros: la distribución log-normal multivariada, la distribución log-t multivariada y la distribución log-slash multivariada, entre otras; además, otra característica atractiva es la fácil interpretación de sus parámetros en términos del vector de variables de interés y su relación con los cuantiles de las distribuciones marginales, lo que las hace particularmente atractivas para la aplicación en modelos de regresión (Morán-Vásquez y cols., 2021).

En las últimas décadas, el modelado de cuantiles, ha cobrado creciente relevancia. Diversos investigadores han dirigido sus esfuerzos hacia propuestas de metodologías para abordar la regresión cuantílica, teniendo como referencia inicial la propuesta presentada por Koenker y Bassett Jr (1978). El planteamiento propuesto por dichos autores, extiende la noción de cuantiles ordinarios en un modelo de localización a una clase más general de modelos lineales en los que los cuantiles condicionales tienen una forma lineal (Buchinsky, 1998).

Este ámbito ha despertado un notable interés, a pesar de que la media poblacional de una variable respuesta proporciona una medida de tendencia central importante, gracias a su capacidad para hacer frente a diversas dificultades comúnmente encontradas en el análisis estadístico, como: valores atípicos, heterocedasticidad, entre otros (Han, Kong, Zhao, y Zhou, 2019). Este tipo de modelos pertenecen a una familia de modelos robustos (Koenker, 2005). Buchinsky (1998), da algunas características útiles de la regresión cuantílica: (a) los modelos pueden utilizarse para caracterizar toda la distribución condicional de una variable dependiente dado un conjunto de variables regresoras, (b) cuando el término de error no es normal, los estimadores de regresión cuantílica pueden ser más eficientes que los de mínimos cuadrados, (c) las soluciones potencialmente diferentes en distintos cuantiles pueden interpretarse como diferencias en la respuesta de la variable dependiente a cambios en los regresores en distintos puntos de la distribución condicional de la variable dependiente (evidenciar el efecto del conjunto de covariables sobre cada cuantil). Además, varios cuantiles transmiten una representación más completa de la distribución condicional de la variable dependiente que un único valor (la media) derivada de un enfoque tradicional, por ejemplo, mínimos cuadrados (Hunter y Lange, 2000).

La regresión cuantílica suele ser abordada principalmente a través de enfoques no paramétricos, los cuales llevan a un problema de minimización (Koenker y Bassett Jr, 1978) basado en desviaciones absolutas ponderadas. Esto significa que no es necesario hacer suposiciones acerca de la distribución de la variable de respuesta. No obstante, algunos investigadores han optado por el enfoque clásico (Sánchez, Lachos, y Labra, 2013; Tian, Tian, y Zhu, 2014), el cual hace suposiciones distribucionales con respecto a la variable de respuesta. Desde una perspectiva Bayesiana (Wichitaksorn, Choy, y Gerlach, 2014; Yu y Moyeed, 2001), también se ha explorado este tema.

Sin embargo, la mayor parte de las propuestas actuales se centran en el ámbito univariado (Galarza Morales, Lachos Davila, Barbosa Cabral, y Castro Cepero, 2017; Sánchez y cols., 2013; Wichitaksorn y cols., 2014; Yu y Moyeed, 2001), en el cual el propósito es modelar cada cuantil de una única variable dependiente sin tener en cuenta el posible efecto o relación con otras variables dependientes. Esto implica, que para cada cuantil que deseemos ajustar, será necesario contar con una cantidad equivalente de parámetros para llevar a cabo dichos ajustes. Estás metodologías producen estimaciones robustas, sobre todo cuando hay presencia de datos atípicos y los errores no son normales (Hunter y Lange, 2000). A pesar de las ventajas en términos de robustez, el enfoque univariado presenta una limitación evidente: cuando se trabaja con múltiples variables dependientes (p variables), es necesario realizar ajustes individuales para cada cuantil, lo que implica un número considerable de ajustes en proporción al total de variables, además, al estimar m cuantiles, la cantidad total de ajustes aumenta a $m \times p$, sin contar con la otra desventaja de que esos ajustes individuales no consideran la relación entre las variables dependientes, a diferencia de lo que logra el enfoque multivariado en el modelado de cuantiles (Morán-Vásquez y cols., 2021).

Es por lo anterior, que se propone el enfoque de modelado de cuantiles multivariados que cuenta con la ventaja distintiva de considerar la relación existente entre las variables respuesta, lo que nos da una ganancia adicional respecto a la regresión cuantílica univariada (Morán-Vásquez y cols., 2021). Esto permite capturar información sobre diferentes aspectos de la distribución conjunta de las variables respuesta. Sin embargo, pese a que varios autores (Chakraborty, 2003; Morán-Vásquez y cols., 2021; Morán-Vásquez, Roldán-Correa, y Nagar, 2023; Petrella y Raponi, 2019; Wei, 2008) han explorado este enfoque desde las perspectivas: clásica, Bayesiana, no paramétrica y semi-paramétrica; muy pocos han considerado un problema ampliamente prevalente en el ámbito estadístico: la presencia de datos faltantes.

El estudio de los datos faltantes, defínanse estos como, aquellos valores no observados que serían significativos para el análisis si se observaran (R. J. A. Little y Rubin, 1987), ha sido de mucha importancia en el campo estadístico y, se han abordado desde diferentes perspectivas, dado que estos se presentan con mucha regularidad. Es común encontrar la presencia de datos faltantes en conjuntos de datos multivariados, debido a múltiples causas, entre las cuales están: mala digitación de registros, no respuesta en encuestas, pérdida de registros, etc. Aunque no jugará un papel relevante en este trabajo, es importante tener en cuenta que no todo dato que no es observado se puede catalogar como dato faltante, dado que, esta "no respuesta" puede ser otro punto muestral (R. J. A. Little y Rubin, 1987).

Para este trabajo, resulta útil diferenciar dos conceptos muy importantes: patrones de datos faltantes y mecanismos de datos faltantes (o mecanismos). Los patrones de los datos faltantes describen cuáles valores son faltantes y cuáles valores son observados en el conjunto de datos, mientras que el mecanismo de los datos faltantes hace referencia a la relación entre los valores faltantes y los valores de las variables en la matriz de datos (R. J. A. Little y Rubin, 1987).

Existen diversas metodologías para tratar los datos faltantes. Estas normalmente están orientadas a los diversos patrones que este tipo de datos pueden presentar, por ejemplo, en el contexto de datos multivariados, datos monótonos. Sin embargo, hay que tener en cuenta que la naturaleza del mecanismo influye enormemente en el rendimiento de las técnicas estadísticas que tratan los datos faltantes (R. J. A. Little y Rubin, 1987; Verhasselt, Flórez, Van Keilegom, y Molenberghs, 2019). R. J. A. Little y Rubin (1987), mencionan que hay tres importantes mecanismos de datos faltantes, que son: datos faltantes completamente aleatorios (MCAR, por sus siglas en inglés), datos faltantes aleatorios (MAR, por sus siglas en inglés), y datos faltantes no aleatorios (MNAR, por sus siglas en inglés). El supuesto más comúnmente utilizado es el MAR (cuando el mecanismo es independiente de las observaciones no observadas condicionadas a la observada). Este supuesto es importante porque: (a) es una condición suficiente para que las inferencias de probabilidad pura y Bayesianas sean válidas sin modelar el mecanismo de los datos faltantes, y (b) la distribución predictiva de los valores faltantes dados los valores observados para cada unidad es independiente del patrón (R. J. A. Little y Rubin, 1987).

En el contexto de la regresión cuantílica bajo la presencia de datos faltantes, existen dos enfoques ampliamente utilizados (Han y cols., 2019; Kleinke, Fritsch, Stemmler, Reinecke, y Lösel, 2021; C. Wang, Tian, y Tang, 2022) para tratar este tipo de datos: el enfoque de imputación y el enfoque de ponderación de probabilidad inversa (IPW, por sus siglas en inglés). Por ejemplo, Han y cols. (2019) plantearon realizar la regresión cuantílica bajo la presencia de datos faltantes mediante la combinación de los dos enfoques, imputación e IPW, para tratar estos tipos de datos, desde una perspectiva paramétrica. Sin embargo, la mayoría de las metodologías en este campo se han centrado predominantemente en el análisis univariado. En contraste, se ha prestado una atención limitada al abordaje de datos faltantes en el contexto multivariado. En numerosos estudios, la tendencia ha sido dirigirse hacia el análisis de conjuntos de datos completos (Howarth, Ben Saad, y Heraganahally, 2023; Roth y cols., 2022), ignorando las observaciones con datos faltantes. No obstante, estos suelen dar lugar a un sesgo sustancial y/o socava la eficacia del estudio, especialmente cuando la proporción de datos faltantes es elevada (Han y cols., 2019).

Por lo tanto, teniendo en cuenta que la metodología del modelado de cuantiles marginales mediante modelos de regresión lineal con respuesta multivariada, positiva y continua, en presencia de datos faltantes tiene pocas propuestas, en esta investigación se pretende desarrollar un modelo de regresión lineal con este tipo de respuesta en presencia de dichos datos y bajo el enfoque de estimación Bayesiana haciendo uso de la clase de distribuciones normal/independiente multivariadas para la estimación de los parámetros de los modelos log-normal/independiente multivariados, con el fin de modelar cuantiles marginales, teniendo en cuenta la asociación entre las variables del vector de respuesta.

Para alcanzar el objetivo establecido, la estructura del trabajo se organiza de la siguiente

manera: en el capítulo 2 se lleva a cabo un estudio exhaustivo de la clase de distribuciones normal/independiente (nombre que se le atribuye, debido a que resulta de combinar dos variables aleatorias independientes; una de estas variables sigue una distribución normal, mientras que la otra es una variable aleatoria positiva) multivariadas, incluyendo la exploración de algunas de sus propiedades más significativas.

En el capítulo 3, se aborda el análisis de los modelos de regresión lineal asociados a la clase de distribuciones normal/independiente multivariadas. Este análisis incluye la consideración del impacto de los datos faltantes, la tipología de dichos datos, así como la exploración de los modelos de regresión lineal en el contexto de la presencia de datos faltantes. Además, se examina un algoritmo diseñado para la estimación de los parámetros del modelo propuesto, el cual posibilita simultáneamente el manejo efectivo de los datos faltantes.

En el capítulo 4, se profundiza en el proceso de modelado de cuantiles marginales a través de la relación existente entre la clase de modelos de regresión lineal log-normal/independiente multivariados y la clase de modelos de regresión lineal normal/independiente multivariados. Este análisis se centra en la exploración y comprensión de cómo estos modelos capturan las relaciones entre variables, destacando su aplicabilidad específica en la estimación de cuantiles marginales.

En los capítulos 5 y 6, se presentan los estudios de simulación, orientado a la estimación de los parámetros de los modelos log-normal/independiente multivariados, haciendo uso de los modelos normal/independiente multivariados, que respaldan la metodología propuesta, así como la aplicación de dicha metodología a un conjunto de datos reales. El código está disponible en `https://github.com/joseescobara/MDA-algorithm`. Finalmente, el capítulo 7 contiene las conclusiones derivadas de los hallazgos obtenidos, junto con algunas recomendaciones pertinentes para futuras investigaciones.

2 Clase de Distribuciones Normal/independiente Multivariadas

Las distribuciones paramétricas desempeñan un papel importante en el análisis y modelado estadístico, siendo la distribución normal ampliamente utilizada debido a sus atractivas características, como la simplicidad, la trazabilidad y diversas propiedades matemáticas (Lee y McLachlan, 2014). Liu (1995) recalca que, la distribución normal multivariada ha sido un modelo estadístico popular en el análisis estadístico, especialmente en el caso de conjuntos de datos que involucran variables continuas. No obstante, en la práctica, los datos reales tienden a desviarse de las suposiciones requeridas por esta distribución, presentando características tales como valores atípicos y colas pesadas. La distribución normal, como se ha observado en estudios previos (K. Lange y Sinsheimer, 1993; Liu, 1996; Rosa, Padovani, y Gianola, 2003), tiende a ser influenciada por valores atípicos o conjuntos de datos de colas pesadas. Es por esto que, en la práctica, el análisis podría no proporcionar inferencias robustas cuando el supuesto de normalidad es cuestionable (Lachos, Bandyopadhyay, y Dey, 2011). Una alternativa para abordar estos problemas, como los datos atípicos, es la detección y eliminación de los mismos, sin embargo, esta práctica conlleva el riesgo de generar inferencias inválidas (Chen y Luo, 2016). Como alternativa, muchos investigadores prefieren emplear transformaciones de los datos con el objetivo de lograr una distribución más cercana a la normalidad. Un ejemplo común es la transformación de Box-Cox (Box y Cox, 1964). A través de esta transformación, se busca lograr la normalidad o una aproximación aceptable de la misma, respaldada por resultados empíricos razonables. Sin embargo, es crucial tener en cuenta algunas consideraciones: (1) la transformación proporciona información limitada sobre un posible esquema subyacente de generación de datos, (2) la transformación por componentes no siempre garantiza la normalidad conjunta, (3) los parámetros pueden perder interpretabilidad en una escala transformada, y (4) las transformaciones no son universales y tienden a variar según el conjunto de datos específico (Lachos y cols., 2011). Por lo tanto, en la práctica se opta por buscar alternativas robustas que puedan tratar estos problemas, evitando transformaciones de los datos. En la práctica, una opción es la clase de distribuciones normal de mezcla de escalas simétricas (SMN, por sus siglas en inglés) (Garay, Bolfarine, Lachos, y Cabral, 2015; Lachos y cols., 2011; Lee y McLachlan, 2014). Estas distribuciones representan una alternativa robusta en presencia de no normalidad en el conjunto de datos (Abanto-Valle, Lachos, y Ghosh, 2012). El uso de la clase de distribuciones SMN está

motivado por las siguientes consideraciones: (1) mantienen su robustez en presencia de no normalidad, y (2) todas estas distribuciones forman una clase que presenta colas más pesadas que la distribución normal, lo que las hace aptas para realizar inferencias robustas (Abanto-Valle y cols., 2012).

Dentro de la familia de las distribuciones SMN, destaca una subclase importante: las distribuciones normal/independiente multivariadas. Estas distribuciones se han utilizado constantemente para el tratamiento de problemas asociados a valores atípicos y colas pesadas, con el propósito de obtener inferencias robustas (Chen y Luo, 2016).

La clase de distribuciones normal/independiente multivariadas hace referencia a una clase de distribuciones normales en la cual la matriz de covarianza está ponderada por una variable de escala (una función positiva) con una distribución a priori específica (Lee y McLachlan, 2014). Estas distribuciones, tienen la característica de regular sus colas, constituyendo una alternativa robusta a la distribución normal cuando se analizan datos que contienen observaciones atípicas (Rosa y cols., 2003).

Según Lee y McLachlan (2014), en términos generales, una distribución de mezcla de escalas puede obtenerse "*mezclando*" o "*promediando*" una densidad base sobre una distribución de escala, en la cual su densidad puede expresarse en la forma siguiente:

$$f(\boldsymbol{y}) = \int_0^\infty g(\boldsymbol{y} \mid k(w))\ dH(w), \tag{2-1}$$

donde $g(\boldsymbol{y} \mid k(w))$ es la densidad condicional (densidad base) de un vector aleatorio dado $k(w)$, y $k(\cdot)$ es una función positiva de una variable de escala W con función de distribución $H(w)$ (la cual se conoce como función de distribución de mezcla).

La clase de distribuciones normal/independiente es un caso especial de las distribuciones de mezclas, y pueden obtenerse (Andrews y Mallows, 1974; Lee y McLachlan, 2014; Liu, 1996; Morán-Vásquez y cols., 2021; Rosa y cols., 2003) al considerar la distribución normal como la densidad base, con $k(w) = 1/w$, es decir, $\phi_p(\boldsymbol{y} \mid \boldsymbol{\mu}, w^{-1}\boldsymbol{\Psi})$; donde $\phi_p(\cdot)$ denota la densidad normal p-multivariada con vector de media $\boldsymbol{\mu}$ y matriz de covarianza $w^{-1}\boldsymbol{\Psi}$ (donde, por generalidad $\boldsymbol{\Psi}$ está ponderada por una función positiva de W, $k(w)$).

Sea $\boldsymbol{Y}$ un vector aleatorio p-dimensional con su distribución perteneciente a la clase normal/independiente multivariada. La función de densidad de probabilidad de $\boldsymbol{Y}$ puede expresarse como el siguiente modelo de mezcla (K. Lange y Sinsheimer, 1993; Morán-Vásquez y cols., 2021; Rosa y cols., 2003):

$$f(\boldsymbol{y}) = \int_0^\infty \phi_p(\boldsymbol{y} \mid \boldsymbol{\mu}, w^{-1}\boldsymbol{\Psi})\ dH(w). \tag{2-2}$$

Esta clase de modelos son unimodales y simétricos (Lee y McLachlan, 2014). La variable de escala W puede ser discreta o continua. En el caso discreto (Lee y McLachlan, 2014), las distribuciones normal/independiente multivariadas tiene la siguiente forma:

$$\sum_{i=0}^{\infty} \phi_p(\boldsymbol{y} \mid \boldsymbol{\mu}, w^{-1}\boldsymbol{\Psi})\ h\ (w). \tag{2-3}$$

Una representación estocástica de las distribuciones normal/independiente multivariadas muy utilizada en la literatura (K. Lange y Sinsheimer, 1993; Lee y McLachlan, 2014; Liu, 1996) es la siguiente:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{Z}/\sqrt{W}, \tag{2-4}$$

donde $\boldsymbol{\mu} \in \mathbb{R}^p$ es el vector de localización, $\boldsymbol{Z} \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Psi})$ e independiente de la variable aleatoria W. Otra forma equivalente de representar estas distribuciones en términos de una estructura jerárquica (Morán-Vásquez y cols., 2021) está dada por:

$$\begin{aligned} \boldsymbol{Y} \mid W = w &\sim \mathrm{N}_p(\boldsymbol{\mu}, w^{-1}\boldsymbol{\Psi}) \\ W &\sim H(w \mid \boldsymbol{\nu}), \end{aligned} \tag{2-5}$$

donde $H\ (w \mid \boldsymbol{\nu})$ es la función de distribución condicionada al vector $\boldsymbol{\nu} \in \mathbb{R}^q$. Cada miembro perteneciente a la clase de distribuciones normal/independiente multivariadas es inducida por la distribución de W (Morán-Vásquez y cols., 2021).

Bajo las suposiciones anteriores, algunas propiedades de la clase de distribuciones normal/independiente multivariadas, según K. Lange y Sinsheimer (1993), son las siguientes:

- $\boldsymbol{Y}$ tiene función de densidad positiva dada por:

$$\int_0^{\infty} \frac{w^{p/2}}{(2\pi)^{p/2} \mid \boldsymbol{\Psi} \mid^{1/2}} e^{\frac{-w\delta^2}{2}}\ dH(w),$$

  donde $\delta^2 = \delta^2(\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{\Psi}) = (\boldsymbol{Y} - \boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$ ($\delta^2(\boldsymbol{Y}; \boldsymbol{\mu}, \boldsymbol{\Psi})$ indica que δ^2 depende de la variable o vector aleatorio $\boldsymbol{Y}$, así como de los parámetros correspondientes $\boldsymbol{\mu}$ y $\boldsymbol{\Psi}$ que caracterizan la distribución de $\boldsymbol{Y}$), es la distancia de Mahalanobis.

- Esta densidad es infinitamente diferenciable en $\boldsymbol{Y}$, $\boldsymbol{\mu}$, y $\boldsymbol{\Psi}$ excepto posiblemente donde $\boldsymbol{Y} = \boldsymbol{\mu}$.

- Si la esperanza $E(\boldsymbol{Y})$ existe, entonces $E(\boldsymbol{Y}) = \boldsymbol{\mu}$.

- Si la covarianza $\mathrm{Cov}(\boldsymbol{Y})$ existe, entonces $\mathrm{Cov}(\boldsymbol{Y}) = E(W^{-1})\ \boldsymbol{\Psi}$.

- Cualquier subvector de $\boldsymbol{Y}$ tiene densidad marginal de la misma forma general que la densidad de $\boldsymbol{Y}$, lo que significa que la familia se preserva bajo marginalización.
- La regresión de cualquier subvector de $\boldsymbol{Y}$ sobre su subvector complementario es lineal.
- La función característica de $\boldsymbol{Y}$ es:

$$E\left(e^{i\boldsymbol{\theta}^t\boldsymbol{Y}}\right) = e^{i\boldsymbol{\theta}^t\boldsymbol{\mu}} \int_0^\infty e^{-1/(2w)\boldsymbol{\theta}^t\boldsymbol{\Psi}\boldsymbol{\theta}}\, dH(w).$$

Debido a la forma de la densidad de $\boldsymbol{Y}$, la distancia de Mahalanobis, $\delta^2 = (\boldsymbol{Y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})$, tiene un rol fundamental, de hecho, δ^2 es extremadamente útil para evaluar la bondad de ajuste y detectar datos atípicos (K. Lange y Sinsheimer, 1993).

La función de distribución de δ^2 está dada por (K. Lange y Sinsheimer, 1993):

$$P(\delta^2 \leq s) = \frac{s^{p/2}}{2^{p/2}\Gamma(p/2)} \int_0^\infty [1-H(w)] w^{p/2-1} e^{-ws/2}\, dw. \tag{2-6}$$

En consecuencia, (2-6) es infinitamente diferenciable, excepto posiblemente en $s = 0$.

Dos distribuciones ampliamente utilizadas que pertenecen a la clase normal/independiente multivariada (K. Lange y Sinsheimer, 1993; R. J. Little, 1988; Liu, 1995; Morán-Vásquez y cols., 2021; Rosa y cols., 2003) son la distribución t-Student (o simplemente t) multivariada y la distribución slash multivariada. Estas distribuciones desempeñan un papel crucial en el modelado estadístico y son particularmente relevantes en contextos donde la robustez y la capacidad de manejar datos atípicos son consideraciones importantes. Otros miembros pertenecientes a esta clase son las distribuciones: normal contaminada multivariada, Pearson tipo VII multivariada, Laplace multivariada, entre otras (K. Lange y Sinsheimer, 1993).

2.1. Distribución t Multivariada

El supuesto de normalidad puede ser fácilmente violado en la práctica, por lo que, se ha visto la necesidad de crear metodologías robustas para suplir la necesidad de la no normalidad en los datos. Como se mencionó anteriormente, la clase de distribuciones normal/independiente multivariadas (K. Lange y Sinsheimer, 1993) ofrecen una alternativa robusta cuando los datos no presentan normalidad. La distribución t multivariada, integrante de esta clase, es ampliamente utilizada en el análisis estadístico como una opción robusta frente a conjuntos de datos que contienen valores atípicos o presentan colas pesadas, en contraste a la distribución normal multivariada (Lee y McLachlan, 2014). De acuerdo con Liu (1994), la distribución t multivariada puede ser empleada para desarrollar procedimientos robustos que permitan

obtener inferencias estadísticas válidas sobre los parámetros de localización, la matriz de dispersión y, en consecuencia, sus funciones, tales como los coeficientes de regresión.

El creciente interés en el estudio de la distribución t multivariada se ha visto impulsado principalmente por las siguientes razones: (1) Esta distribución representa una generalización de la conocida distribución t de Student univariada, la cual desempeña un papel fundamental en la inferencia estadística; la variedad de estructuras posibles es amplia, cada una con características especiales que resultan relevantes en diversas aplicaciones, (2) La aplicación de la distribución t multivariada se presenta como un enfoque altamente prometedor en el análisis multivariado, ya que proporciona una alternativa más realista para modelar datos del mundo real, particularmente debido a la naturaleza más adecuada de sus colas, y (3) Durante las últimas dos o tres décadas, la distribución t multivariada ha desempeñado un papel crucial en el ámbito del análisis Bayesiano de datos multivariados (Kotz y Nadarajah, 2004).

La forma original de la distribución t multivariada en el análisis estadístico no es fácil de tratar, sin embargo, su representación en términos de un modelo de mezcla de escalas la vuelve atractiva (Lee y McLachlan, 2014). En este contexto, la distribución t multivariada puede concebirse como una normal multivariada con una escala faltante distribuida gamma, considerada como peso, que se integra a partir de la distribución conjunta de la normal condicional y la distribución marginal del peso (Liu, 1994). Según Boris Choy y Chan (2008), expresar una distribución simétrica en forma de mezcla de escalas facilita la aplicación de algoritmos Bayesianos de Monte Carlo por cadenas de Markov (MCMC) eficientes en la implementación de modelos estadísticos complejos.

La distribución t fue originalmente propuesta por Student (1908). Posteriormente, diversos autores formalizaron ciertos resultados de esta distribución, y otros, como Kibria y Joarder (2006), Lin (1972) y Dunnett y Sobel (1954), la generalizaron al ámbito multivariado. La forma más común de presentar la función de densidad para un vector p-variado, $\boldsymbol{Y} = (y_1, y_2, \ldots, y_p)'$, que sigue una distribución t multivariada se describe de la siguiente manera (Kotz y Nadarajah, 2004; Liu, 1994):

$$f(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Psi}, \nu) = \frac{\mid \boldsymbol{\Psi} \mid^{-1/2} \Gamma(\frac{v+p}{2})}{(\nu\pi)^{p/2}\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{v+p}{2}}, \tag{2-7}$$

donde $\boldsymbol{\mu} \in \mathbb{R}^p$ denota el vector de localización, $\boldsymbol{\Psi} > 0$ (definida positiva) representa la matriz de dispersión, y ν corresponde a los grados de libertad.

Cada miembro de esta familia se caracteriza por un parámetro de grados de libertad ν (K. Lange y Sinsheimer, 1993). Este parámetro juega un papel crucial en el análisis robusto al trabajar con la distribución t, ya que proporciona una dimensión conveniente para la

inferencia estadística robusta, con un aumento moderado en la complejidad computacional para muchos modelos (K. L. Lange, Little, y Taylor, 1989). Además, la distribución t reduce automáticamente la influencia de los valores atípicos en la inferencia estadística (Boris Choy y Chan, 2008).

Los grados de libertad son especialmente importantes para modular el comportamiento de la probabilidad de cola en la distribución t multivariada (K. L. Lange y cols., 1989). De esta manera, pueden considerarse como un parámetro de ajuste para la robustez; a menor valor de ν, mayor será la probabilidad de cola (Liu, 1994). En consecuencia, la distribución normal multivariada es obtenida cuando $\nu \to \infty$ (K. Lange y Sinsheimer, 1993).

La distribución t multivariada se puede representar de manera conveniente (K. Lange y Sinsheimer, 1993; Liu, 1996; Morán-Vásquez y cols., 2021; Rosa y cols., 2003) mediante la siguiente forma jerárquica:

$$\begin{aligned} \boldsymbol{Y} \mid W = w &\sim \mathrm{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Psi}/w\right) \\ h(w \mid \nu) &\sim \Gamma\left(\nu/2, \nu/2\right), \end{aligned} \tag{2-8}$$

donde $\boldsymbol{\mu} \in \mathbb{R}^p$ denota el vector de localización, $\boldsymbol{\Psi} > 0$ (definida positiva) representa la matriz de dispersión, y w es la variable de escala.

Podemos demostrar, por lo tanto, que $\boldsymbol{Y} \sim \mathrm{t}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$. Esta afirmación se evidencia claramente a través del siguiente proceso de marginalización, del modelo de mezcla (2-8):

$$\begin{aligned} f(\boldsymbol{y} \mid w) &= \frac{w^{p/2}}{(2\pi)^{p/2} \mid \boldsymbol{\Psi} \mid^{1/2}} \exp\left\{-\frac{w(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right\} \\ h(w \mid \nu) &= \frac{(\frac{\nu}{2})^{\frac{\nu}{2}} w^{\frac{\nu}{2}-1} \exp\{-\frac{\nu w}{2}\}}{\Gamma(\frac{\nu}{2})}. \end{aligned}$$

Marginalizando, obtenemos lo siguiente:

$$\begin{aligned} f(\boldsymbol{y}) &= \int_0^\infty \frac{w^{p/2}}{(2\pi)^{p/2} \mid \boldsymbol{\Psi} \mid^{1/2}} \exp\left\{-\frac{w(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right\} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}} w^{\frac{\nu}{2}-1} \exp\{-\frac{\nu w}{2}\}}{\Gamma(\frac{\nu}{2})}\, dw \\ &= \frac{(\frac{v}{2})^{\frac{v}{2}} \mid \boldsymbol{\Psi} \mid^{-1/2}}{(2\pi)^{p/2}\Gamma\left(\frac{v}{2}\right)} \int_0^\infty w^{\frac{v}{2}-1} w^{p/2} \exp\left\{-\frac{w(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right\} \exp\left\{-\frac{\nu w}{2}\right\}\, dw \\ &= \frac{(\frac{v}{2})^{\frac{v}{2}} \mid \boldsymbol{\Psi} \mid^{-1/2}}{(2\pi)^{p/2}\Gamma\left(\frac{v}{2}\right)} \int_0^\infty w^{\frac{v+p}{2}-1} \exp\left\{-w\left[\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right]\right\}\, dw \\ &= \frac{(\frac{v}{2})^{\frac{v}{2}} \mid \boldsymbol{\Psi} \mid^{-1/2}}{(2\pi)^{p/2}\Gamma\left(\frac{v}{2}\right)} \frac{\Gamma(\frac{v+p}{2})}{\Gamma(\frac{v+p}{2})} \frac{\left(\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right)^{\frac{v+p}{2}}}{\left(\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right)^{\frac{v+p}{2}}} \times A, \end{aligned}$$

donde,

$$A = \int_0^\infty w^{\frac{v+p}{2}-1} \exp\left\{-w\left[\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right]\right\} dw$$

Así,

$$\begin{aligned} f(\boldsymbol{y}) &= \frac{(\frac{v}{2})^{\frac{v}{2}} \mid \boldsymbol{\Psi} \mid^{-1/2}}{(2\pi)^{p/2}\Gamma\left(\frac{v}{2}\right)} \frac{\Gamma(\frac{v+p}{2})}{\left(\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right)^{\frac{v+p}{2}}} \int_0^\infty \frac{w^{\frac{v+p}{2}-1} \exp\left\{-w\left[\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right]\right\}}{\left(\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right)^{-\frac{v+p}{2}} \Gamma(\frac{v+p}{2})} dw \\ &= \frac{(\frac{v}{2})^{\frac{v}{2}} \mid \boldsymbol{\Psi} \mid^{-1/2}}{(2\pi)^{p/2}\Gamma\left(\frac{v}{2}\right)} \frac{\Gamma(\frac{v+p}{2})}{\left(\frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2} + \frac{v}{2}\right)^{\frac{v+p}{2}}} \\ &= \frac{\mid \boldsymbol{\Psi} \mid^{-1/2} \Gamma(\frac{v+p}{2})}{(\nu\pi)^{p/2}\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{v+p}{2}}, \end{aligned}$$

lo cual claramente muestra que $\boldsymbol{Y} \sim \mathrm{t}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$. La representación jerárquica (2-8) resultará de gran ayuda para la aplicación efectiva del modelo bajo esta distribución.

Entre las propiedades fundamentales de la distribución t multivariada (K. Lange y Sinsheimer, 1993; Liu, 1994), se encuentran:

1. Su valor esperado es $E\ (\boldsymbol{Y}) = \boldsymbol{\mu}$, para $\nu > 1$.

2. Su matriz de covarianzas está dada por:

$$\mathrm{Cov}(\boldsymbol{Y}) = E\ (W^{-1})\boldsymbol{\Psi} = \frac{\nu}{\nu-2}\boldsymbol{\Psi},$$

   para $\nu > 2$. Este resultado se deriva del cálculo de los momentos recíprocos finitos de la variable de escala W para la distribución t multivariada, los cuales se definen como:

$$E(W^{-m}) = \frac{(w/2)^m \Gamma((w/2)-m)}{\Gamma(w/2)},$$

   para $m < w/2$ (K. Lange y Sinsheimer, 1993).

3. Si $\boldsymbol{Y} \sim \mathrm{t}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$ y sea $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Y}$ una transformación lineal, donde $\boldsymbol{A}$ es una matrix $(q \times p)$ no singular y $q \leq p$, entonces

$$\boldsymbol{X} \sim \mathrm{t}_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}', \nu).$$

4. Si $\boldsymbol{Y} \sim \mathrm{t}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$, la distribución de cualquier componente k $(1 \leq k \leq p)$ de $\boldsymbol{Y}$ también sigue una distribución t univariada. En este caso, los parámetros de localización y dispersión son las correspondientes componentes de $\boldsymbol{\mu}$ y $\boldsymbol{\Psi}$, es decir, $Y_k \sim \mathrm{t}_1(\mu_k, \sigma^2_{kk}, \nu)$, donde ν representa los grados de libertad. En consecuencia, la distribución t mantiene la familia bajo marginalización (Morán-Vásquez y cols., 2021).

5. Sin perder generalidad, al realizar una partición del vector $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)'$ con dimensiones p_1 y p_2, respectivamente, la distribución condicional de $\boldsymbol{Y}_1$ dado $\boldsymbol{Y}_2$ es:

$$\boldsymbol{Y}_1 \mid \boldsymbol{Y}_2 \sim \mathrm{t}_{p_1}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Psi}_{1|2}, \nu + p_2),$$

donde

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}(\boldsymbol{Y}_2 - \boldsymbol{\mu}_2)$$

y

$$\boldsymbol{\Psi}_{1|2} = (\boldsymbol{\Psi}_{11} - \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21})\frac{\nu + (\boldsymbol{Y}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Psi}_{22}^{-1}(\boldsymbol{Y}_2 - \boldsymbol{\mu}_2)}{\nu + p_2}.$$

6. La distribución de la distancia de Mahalanobis dividida sobre p (número de variables respuesta) está dada por:

$$\frac{\delta^2}{p} = \frac{(\boldsymbol{Y} - \boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})}{p} \sim \mathrm{F}_{p,v}.$$

Para obtener información adicional sobre otras propiedades de la distribución t multivariada, se pueden consultar en Liu (1994), Lee y McLachlan (2014), y Kotz y Nadarajah (2004).

2.2. Distribución Slash Multivariada

Otra distribución simétrica con colas pesadas, ampliamente utilizada en estudios robustos, es la distribución slash (Genç, 2007). La versión univariada estándar de esta distribución fue primeramente propuesta por Rogers y Tukey (1972) como la distribución del cociente de dos variables aleatorias independientes: una distribuida normal estándar y la otra distribuida uniforme en el intervalo $(0, 1)$. Similar a la distribución t, la distribución slash es comúnmente empleada como una alternativa robusta a la distribución normal (Arslan y Genç, 2009). La forma general de la distribución slash univariada estándar se define como el cociente $Z/U^{1/q}$ para $q > 0$, siendo el parámetro de forma, donde las variables siguen la descripción previamente mencionada (Arslan y Genç, 2009). Cuando $q = 1$, obtenemos la distribución slash estándar o canónica, y su función de densidad es simétrica, con colas más pesadas que la normal, y puede expresarse en términos de la densidad Gaussiana estándar (Kafadar, 2004), $\phi(y)$, como:

$$\begin{aligned} f(y) &= \frac{1 - e^{-y^2/2}}{\sqrt{2\pi}y^2} \\ &= \begin{cases} [\phi(0) - \phi(y)], & y \neq 0 \\ \phi(0)/2, & y = 0. \end{cases} \end{aligned} \tag{2-9}$$

La distribución slash suele ser utilizada en estudios de simulación, donde se contrasta una situación extrema de colas pesadas versus el caso Gaussiano y se compara el desempeño de dichos modelos estadísticos (Kafadar, 2004). La distribución slash de localización y escala se obtiene mediante la multiplicación de la escala y el cambio de ubicación de una variable aleatoria slash estándar (J. Wang y Genton, 2006).

La versión multivariada de esta distribución slash fue propuesta por K. Lange y Sinsheimer (1993), donde la variable de mezcla de escala tiene una distribución Beta $(\nu, 1)$, $\nu > 0$. Esta distribución tiene las colas más pesadas que la normal y es comúnmente utilizada en el análisis estadístico robusto (Arslan y Genç, 2009). Sin embargo, múltiples autores (Arslan y Genç, 2009; Gómez, Quintana, y Torres, 2007; Reyes, Gallardo, Bolfarine, y Gómez, 2019) se han esforzado por generalizar esta distribución para modelar datos que presentan alta curtosis, dando lugar a nuevas familias distribucionales que presentan colas más pesadas que la distribución slash multivariada usual. Por ejemplo, Reyes y Iriarte (2023) propusieron una versión modificada de la distribución slash mediante ajustes en la distribución de la variable de mezcla, considerando una distribución Birnbaun-Saunders. No obstante, nos enfocaremos en la representación original de la distribución slash propuesta por K. Lange y Sinsheimer (1993). Es importante señalar que la distribución normal multivariada es obtenida cuando $\nu \to \infty$ (Morán-Vásquez y cols., 2021).

La distribución slash multivariada puede ser representada mediante la siguiente estructura jerárquica:

$$\begin{aligned} \boldsymbol{Y} \mid W = w &\sim \mathrm{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Psi}/w\right) \\ h(w \mid \nu) &\sim \text{Beta}\,(\nu, 1), \end{aligned} \tag{2-10}$$

donde $\boldsymbol{\mu} \in \mathbb{R}^p$ denota el vector de localización, $\boldsymbol{\Psi} > 0$ (definida positiva) representa la matriz de dispersión, y w es la variable de escala.

De (2-10) se puede derivar que $\boldsymbol{Y}$ tiene distribución slash p-variada, lo cual se puede denotar como $\boldsymbol{Y} \sim \mathrm{SL}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$. Esta afirmación se evidencia claramente a través del siguiente proceso de marginalización, del modelo de mezcla (2-10):

$$\begin{aligned} f(\boldsymbol{y} \mid w) &= \frac{w^{p/2}}{(2\pi)^{p/2} \mid \boldsymbol{\Psi} \mid^{1/2}} \exp\left\{-\frac{w(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right\}, \\ h(w \mid \nu) &= \nu w^{\nu-1}. \end{aligned}$$

Marginalizando:

$$\begin{aligned} f(\boldsymbol{y}) &= \int_0^1 \frac{w^{p/2}}{(2\pi)^{p/2} \mid \boldsymbol{\Psi} \mid^{1/2}} \exp\left\{-\frac{w(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right\} \nu w^{\nu-1}\, dw \\ &= \frac{\nu}{(2\pi)^{p/2} \mid \boldsymbol{\Psi} \mid^{1/2}} \int_0^1 w^{p/2+\nu-1} \exp\left\{-\frac{w(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}\right\} dw \\ &= \frac{\nu}{(2\pi)^{p/2} \mid \boldsymbol{\Psi} \mid^{1/2}} \int_0^1 w^{p/2+\nu-1} \exp\left\{-\frac{w\delta^2}{2}\right\} dw \\ &= \begin{cases} \frac{\nu|\boldsymbol{\Psi}|^{-1/2}2^{\nu+\frac{p}{2}}\Gamma_{0-1}(\nu+\frac{p}{2},\delta^2/2)}{(2\pi)^{p/2}(\delta^2)^{\nu+\frac{p}{2}}}, & \boldsymbol{Y} \neq \boldsymbol{\mu} \\ \frac{|\boldsymbol{\Psi}|^{-1/2}}{(2\pi)^{p/2}}\frac{2\nu}{2\nu+p}, & \boldsymbol{Y} = \boldsymbol{\mu}, \end{cases} \end{aligned}$$

la cual representa la función densidad de probabilidad de la slash multivariada y está expresada en términos de la función gamma incompleta en el intervalo $(0, 1)$ y que denotamos por Γ_{0-1} (Arslan, 2008; K. Lange y Sinsheimer, 1993). La solución a esta función puede ser evaluada de diferentes maneras. Por ejemplo, la conocida serie de potencias (K. Lange y Sinsheimer, 1993),

$$G(\beta, s) = e^{-s} \sum_{i=0}^{\infty} \frac{s^i}{\prod_{j=0}^{j}(\beta + j)},$$

la cual converge rápidamente para s de orden β o menor. Para s grande comparado a β, K. Lange y Sinsheimer (1993) proponen la siguiente aproximación asintótica,

$$G(\beta, s) = s^{-\beta} \int_0^s z^{\beta-1} e^{-z}\, dz.$$

No obstante, la representación jerárquica (2-10) será de suma utilidad para la implementación de nuestro modelo, facilitándonos una estimación más eficiente de las características poblacionales.

Algunas de sus propiedades (K. Lange y Sinsheimer, 1993; J. Wang y Genton, 2006) son:

1. Su valor esperado es $E\,(\boldsymbol{Y}) = \boldsymbol{\mu}$.

2. Su matriz de covarianzas está dada por:

$$\text{Cov}\,(\boldsymbol{Y}) = E(W^{-1})\boldsymbol{\Psi} = \frac{\nu}{\nu-1}\boldsymbol{\Psi},$$

para $\nu > 1$. Este resultado se deduce del hecho de que los momentos recíprocos finitos de la variable de escala W para la distribución slash multivariada están dados por:

$$E(W^{-m}) = \frac{\nu}{\nu-m},$$

para $m < \nu$.

3. Si $\boldsymbol{Y} \sim \mathrm{SL}_p\ (\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$, entonces la transformación lineal $\boldsymbol{Y} = \boldsymbol{b} + A\boldsymbol{Y} \sim \mathrm{SL}_p(\boldsymbol{b} + A\boldsymbol{\mu}, \boldsymbol{A\Psi A'}, \nu)$. En otras palabras, la distribución slash es invariante bajo transformaciones lineales.

4. Si $\boldsymbol{Y} \sim \mathrm{SL}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$, la distribución de cualquier componente k $(1 \leq k \leq p)$ de $\boldsymbol{Y}$ también sigue una distribución slash univariada. En este contexto, las correspondientes componentes de $\boldsymbol{\mu}$ y $\boldsymbol{\Psi}$ actúan como los parámetros de localización y dispersión, respectivamente, mientras que ν se presenta como el parámetro de cola $(Y_k \sim \mathrm{SL}_1(\mu_k, \sigma^2_{kk}, \nu))$. Es decir, esta distribución mantiene la familia bajo el proceso de marginalización.

5. Sin perder generalidad, al realizar una partición del vector $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)'$ con dimensiones p_1 y p_2, respectivamente, la distribución condicional de $\boldsymbol{Y}_1$ dado $\boldsymbol{Y}_2$ sigue siendo slash.

6. La función de distribución de la distancia de Mahalanobis $\delta^2 = (\boldsymbol{Y} - \boldsymbol{\mu})'\boldsymbol{\Psi^{-1}}(\boldsymbol{Y} - \boldsymbol{\mu})$ está dada por:

$$P(\delta^2 \leq s) = P(\chi^2_p \leq s) - \frac{2^\nu \Gamma((p/2) + \nu)}{s^\nu \Gamma(p/2)} P(\chi^2_{p+2\nu} \leq s).$$

En el siguiente capítulo se exponen los modelos de regresión lineal asociados a la clase de distribuciones normal/independiente multivariadas y se conectará con la metodología utilizada para "lidiar" con el problema de datos faltantes.

3 Modelos de Regresión Lineal Asociados a la Clase de Distribuciones Normal/independiente Multivariadas con Datos Faltantes Monótonos

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. Las aplicaciones de la regresión son numerosas y se dan en casi todos los campos, como la ingeniería, las ciencias físicas y químicas, la economía, la gestión, las ciencias biológicas y de la vida y las ciencias sociales, de hecho, el análisis de regresión puede ser la técnica estadística más utilizada (Montgomery, Peck, y Vining, 2021).

En general, el modelo de regresión lineal más comúnmente utilizado para evaluar las relaciones entre variables implica una variable respuesta o dependiente (regresión univariada) y una variable independiente (regresión simple) o un conjunto de variables independientes (regresión múltiple). Sin embargo, al considerar un conjunto de p variables respuesta (regresión multivariada), se logra una ventaja adicional al examinar las relaciones entre ellas, lo que conduce a estimaciones más precisas. En este enfoque, se explora y modela la interdependencia entre múltiples variables de respuesta simultáneamente, proporcionando una perspectiva más completa de la relación subyacente en los datos.

El modelo de probabilidad más ampliamente empleado para modelar un vector de variables continuas con respuesta multivariada con un conjunto de variables explicativas, es la distribución normal multivariada; no obstante, estos modelos de regresión presentan diversas limitaciones, siendo su sensibilidad a datos atípicos una de las más destacadas (Morán-Vásquez y cols., 2021).

Según Morán-Vásquez y cols. (2021), una alternativa robusta para abordar la presencia de datos atípicos son los modelos de regresión lineal elípticos (Arellano-Valle, Galea-Rojas, y Zuazola, 2000), específicamente la subclase de modelos de regresión lineal normal/independiente multivariados (K. Lange y Sinsheimer, 1993; Liu, 1996). Esta subclase engloba modelos de regresión lineal basados en distribuciones elípticas con colas más pesadas que la distribución normal. Ejemplos de tales modelos incluyen los de regresión lineal multivariados t y slash.

Sea $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ vectores aleatorios independientes que representan observaciones de $\boldsymbol{Y} \in \mathbb{R}^p$ sobre n individuos, donde $\boldsymbol{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$, para $i = 1, \ldots, n$. La componente Y_{ik} representa la respuesta del i-ésimo individuo para la k-ésima variable. Los componentes de $\boldsymbol{Y}_i$ son posiblemente correlacionados. La clase de modelos de regresión lineal normal/independiente multivariados (K. Lange y Sinsheimer, 1993; Liu, 1996) es definido como:

$$\begin{cases} \boldsymbol{Y}_i \overset{\text{ind}}{\sim} \text{NI}_p(\boldsymbol{\mu}_i, \boldsymbol{\Psi}, H), \\ \boldsymbol{\mu}_i = \boldsymbol{\beta}' \boldsymbol{X}_i, \end{cases} \tag{3-1}$$

para $i = 1, \ldots, n$, con $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ir})'$ siendo la i-ésima fila de la matriz modelo $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]'$. El vector $\boldsymbol{X}_i$ contiene los valores del i-ésimo individuo medido sobre r variables explicativas. Por lo tanto, x_{ij} representa la observación del i-ésimo individuo sobre la j-ésima variable explicativa, con $x_{i1} = 1$. $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})' \in \mathbb{R}^p$ representa el vector de localización para el i-ésimo individuo o equivalentemente, del vector $\boldsymbol{Y}_i$. $\boldsymbol{\Psi}(p \times p) > 0$ (definida positiva) es la matriz de dispersión y $\boldsymbol{\beta}$ es la matriz de coeficientes $(r \times p)$.

Una forma equivalente de representar el modelo (3-1) (Liu, 1996) es mediante la siguiente estructura jerárquica:

$$\begin{aligned} \boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{w} &\overset{\text{ind}}{\sim} \text{N}_p(\boldsymbol{\beta}' \boldsymbol{X}_i, \boldsymbol{\Psi}/w_i), \\ w_i \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\nu} &\overset{\text{iid}}{\sim} H(w \mid \boldsymbol{\nu}), \end{aligned} \tag{3-2}$$

para $i = 1, \ldots, n$ y $w_i > 0$.

Cada miembro de la clase de modelos de regresión lineal normal/independiente multivariados (K. Lange y Sinsheimer, 1993; Liu, 1996) está determinado por la función de distribución H en (3-1), proporcionando varias alternativas para el modelado estadístico a través de la regresión lineal multivariada basada en distribuciones elípticas. Por ejemplo, el modelo de regresión lineal normal multivariado (H es la función de distribución de un vector aleatorio $\boldsymbol{W}$ con distribución degenerada en $w = 1$), el modelo de regresión lineal multivariado t (H es la función de distribución de un vector aleatorio $\boldsymbol{W}$ con función de densidad de probabilidad $\Gamma(\nu/2, \nu/2)$), el modelo de regresión lineal multivariado slash (H es la función de distribución de un vector aleatorio $\boldsymbol{W}$ con función de densidad Beta$(\nu, 1)$). Otros miembros son los modelos de regresión lineal normal contaminado multivariado, Pearson tipo VII multivariado, Laplace multivariado, entre otros.

Desde el enfoque Bayesiano, la forma más simple de llevar a cabo la inferencia Bayesiana en el caso de datos completos, es aplicando una familia de distribuciones a priori para $\boldsymbol{\beta}$ y $\boldsymbol{\Psi}$ que es conjugada a la función de verosimilitud (Schafer, 1997). Cuando $\boldsymbol{\beta}$ y $\boldsymbol{\Psi}$ son desconocidos, la clase conjugada más natural para datos multivariados normal/independiente es la familia

normal Wishart inversa (Gelman, Carlin, Stern, y Rubin, 1995; Schafer, 1997). No obstante, cuando la información a priori no es fuerte, comúnmente se recurre a distribuciones a priori no informativas. Siguiendo la distribución a priori propuesta por Liu (1996) para $\boldsymbol{\beta}, \boldsymbol{\Psi}$ y $\boldsymbol{\nu}$, cuando el conjunto de datos se distribuyen normal/independiente multivariado, y asumiendo que $\boldsymbol{\beta}, \boldsymbol{\Psi}$ y $\boldsymbol{\nu}$ son independientes a priori, está dada por:

$$P(\boldsymbol{\beta}, \boldsymbol{\Psi}) \propto |\boldsymbol{\Psi}|^{-(\frac{m+1}{2})} \exp\left\{-\frac{1}{2}\text{tr}\boldsymbol{\Psi}^{-1}\boldsymbol{A}\right\}, \tag{3-3}$$

donde m es un escalar y $\boldsymbol{A}$ es una matriz definida no negativa $(p \times p)$. Si $m = p$ y $\boldsymbol{A} = \boldsymbol{0}$, entonces la distribución a priori (3-3) se convierte en la distribución a priori no informativa o a priori de Jeffreys (Box y Tiao, 1973), que se utiliza comúnmente en estadística aplicada; si $m = -1$ y $\boldsymbol{A} = \boldsymbol{0}$, entonces la a priori anterior (3-3) es plana; es decir, $P(\boldsymbol{\Psi}) \propto$ constante (Liu, 1996).

Las distribuciones a priori para el vector $\boldsymbol{\nu}$, en los 2 modelos considerados, se discutirán en la sección (3.2).

3.1. Datos Faltantes Monótonos

Las metodologías estadísticas estándares están orientadas principalmente al análisis de conjuntos de datos rectangulares, donde normalmente las filas de estos conjuntos de datos representan las unidades, casos u observaciones, dependiendo sea el caso y, las columnas representan las características o variables medidas sobre las unidades. Infortunadamente, por cualquier razón, raramente los datos suelen observarse de manera completa. La presencia de los datos faltantes y los problemas que plantea tanto para el análisis como para la inferencia han dado lugar a una importante literatura estadística que se remonta a la década de 1950 (Carpenter y cols., 2023).

Los datos faltantes son definidos (Carpenter y cols., 2023; R. J. A. Little y Rubin, 1987) como aquellos valores no observados que serían significativos para el análisis si se observaran; en otras palabras, un valor que falta oculta un valor significativo. Sin embargo, hay que tener en cuenta que no en todas las ocasiones la definición anterior aplica, según R. J. A. Little y Rubin (1987), no todo dato no observado se puede considerar como faltante.

Desde una perspectiva general, como señala Carpenter y cols. (2023), la presencia de este tipo de datos puede dar lugar a dos problemas fundamentales: pérdida de eficiencia y sesgo. En primer lugar, la pérdida de eficacia, o de información, se convierte en una consecuencia inevitable. En segundo lugar, al estimar parámetros de interés como la media o la varianza, la falta de datos en una pequeña proporción de observaciones puede ejercer un impacto desproporcionado en la estimación resultante.

Existe una amplia distinción entre dos tipos de datos faltantes: datos faltantes intencionados y datos faltantes no intencionados. Los datos que faltan intencionadamente son planificados por el recolector de datos, mientras que los datos faltantes no intencionados son imprevistos y escapan al control del recolector de datos, como indica Van Buuren (2018).

Otra distinción relevante (Carpenter y cols., 2023; R. J. A. Little y Rubin, 1987; Van Buuren, 2018), es entre la falta de respuesta por ítem y la falta de respuesta por unidad. La falta de respuesta por ítem se refiere a la situación en la que el encuestado omite uno o más ítems de la encuesta. En cambio, la falta de respuesta por unidad ocurre cuando el encuestado se niega a participar, resultando en la ausencia de todos los datos de resultados asociados a ese encuestado.

En este marco, conviene distinguir entre el patrón de los datos faltantes y el mecanismo de los datos faltantes. En primer lugar, el mecanismo de los datos faltantes, es aquel proceso que genera estos datos, en otras palabras, es un modelo o proceso probabilístico que da lugar a valores faltantes y su relación con las variables de estudio o análisis (Beale y Little, 1975). Es muy importante comprender los mecanismos de los datos que faltan para juzgar la idoneidad de un procedimiento de análisis que utilice los datos observados; es decir, conocer las razones por las que los datos están incompletos es un primer paso hacia la solución (Van Buuren, 2018).

La idea de asignarle un proceso a los datos faltantes fue propuesta en el artículo semilla Rubin (1976), donde clasificó los mecanismos en tres categorías diferentes, en la cual la base de su estudio fue asignarles a todos los puntos alguna probabilidad de faltar. Hay tres tipos de mecanismos (Beale y Little, 1975; Carpenter y cols., 2023; Raghunathan, 2015; Rubin, 1976; Van Buuren, 2018) que gobiernan los datos faltantes: faltantes completamente aleatorios (MCAR, por sus siglas en inglés), faltantes aleatorios (MAR, por sus siglas en inglés) y faltantes no aleatorios (MNAR, por sus siglas en inglés). La distinción de Rubin (1976) es importante para entender por qué algunos métodos funcionan y otros no (Van Buuren, 2018). Los datos faltantes son MCAR si la probabilidad de que falte es la misma para todos los casos, esto implica que las causas de los datos que faltan no están relacionadas con los datos (Van Buuren, 2018). Cuando el mecanismo tiene esta característica el análisis de caso completo es válido, sin embargo, es difícil establecer condiciones generales para la validez del análisis de caso completo, además, el análisis de caso completo suele tener errores de muestreo más grandes debido al menor tamaño de la muestra, aunque las estimaciones puntuales sean insesgadas (Raghunathan, 2015). Aunque el enfoque MCAR es conveniente, a menudo resulta poco realista para los datos en cuestión (R. J. A. Little y Rubin, 1987).

El supuesto MCAR es muy fuerte (R. J. A. Little y Rubin, 1987), dado que, a menudo, varios factores o características de los sujetos pueden influir en la decisión de responder a las preguntas de la encuesta. Un supuesto más débil, según Van Buuren (2018), es el supuesto

MAR. Si la probabilidad de que falte un dato es la misma sólo dentro de los grupos definidos por los datos observados, entonces los datos faltantes son MAR (Raghunathan, 2015). así, la probabilidad de que falte un dato depende solo de la información observada, incluidos los factores de diseño (Carpenter y cols., 2023). La ventaja de este supuesto es que si los datos faltantes son MAR esto es una condición suficiente para que las inferencias de probabilidad pura y Bayesianas sean válidas sin modelar el mecanismo, además, la distribución predictiva de los valores que faltan dados los valores observados para cada unidad es independiente del patrón; esta distribución predictiva es entonces la base de los métodos de imputación, y el supuesto MAR permite estimar esta distribución predictiva de los datos faltantes a partir de los datos observados (R. J. A. Little y Rubin, 1987). El supuesto MAR, aunque a veces poco realista, puede ser una mejor aproximación a la realidad que el supuesto MCAR (R. J. A. Little y Rubin, 1987). En consecuencia, MNAR significa que la probabilidad de que falte algún dato varía por razones que desconocemos (Van Buuren, 2018).

En segundo lugar, el patrón de los faltantes es aquel que describe qué valores faltan y cuáles están observados en la matriz de datos (R. J. A. Little y Rubin, 1987). Es muy importante investigar los patrones de los datos que faltan antes de embarcarse en un análisis formal, esto puede arrojar información vital que de otro modo se pasaría por alto, e incluso puede permitir rastrear los datos faltantes (Carpenter y cols., 2023). Una ventaja de identificar el patrón de los datos faltantes es que podría aprovecharse en la especificación del modelo o dividiendo el problema de estimación en tareas modulares más sencillas; el segundo uso del patrón es comprender la limitación de los datos o identificar los parámetros que no pueden estimarse (Raghunathan, 2015). Por lo tanto, resulta beneficioso ordenar las filas y columnas de los datos de acuerdo a los datos faltantes para ver si surge un patrón ordenado. Según R. J. A. Little y Rubin (1987), algunos patrones de datos faltantes que pueden presentarse son: Univariado, monótono, general, etc.

El patrón monótono, como menciona Liu (1994), es un caso muy importante de los datos incompletos. Un conjunto de datos con valores faltantes tiene un comportamiento monótono si, (Carpenter y cols., 2023; R. J. A. Little y Rubin, 1987; Raghunathan, 2015; Schafer, 1997), siempre que falte un elemento y_{ij}, también faltará y_{ik} para todo $k > j$, como se observa en la Tabla (**3-1**).

Otra definición de datos monótonos, muy similar y en la que nos basaremos, es proporcionada por Liu (1995). Según este autor, un conjunto de datos incompletos con n observaciones y p variables presenta un patrón monótono si puede ser ordenado de tal manera que la j-ésima variable está al menos tan observada como la $(j-1)$-ésima variable para $j = 2, \ldots, p$, con p posibles patrones. La Tabla (**3-2**) ilustra los p posibles patrones de un conjunto de datos incompleto monótono, representados de la siguiente manera (Liu, 1995):

$$\boldsymbol{Y}_{\text{MP}} = \{(y_{i,k}^{(k)}, ..., y_{i,p}^{(k)}) : i = 1, ..., n_k; k = 1, ..., p\}, \tag{3-4}$$

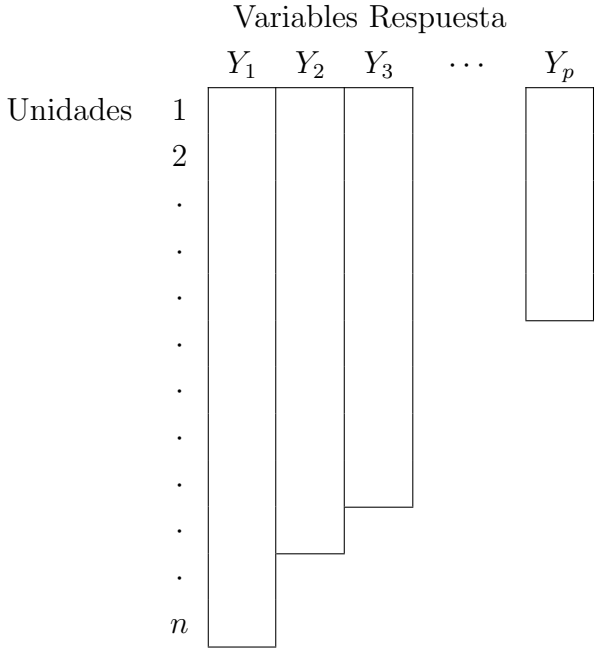


Tabla 3-1: Tomada de Schafer (1997)

Variables Respuesta (el segundo subíndice)

	$y_{1,1}^{(1)}$	$y_{1,2}^{(1)}$ $\cdots$	$y_{1,p}^{(1)}$	
	$\vdots$	$\vdots$ $\cdots$	$\vdots$	Patrón 1
	$y_{n_1,1}^{(1)}$	$y_{n_1,2}^{(1)}$ $\cdots$	$y_{n_1,p}^{(1)}$	(El superíndice)
		$y_{1,2}^{(2)}$ $\cdots$	$y_{1,p}^{(2)}$	
		$\vdots$ $\cdots$	$\vdots$	Posible patrón 2
$n = \sum_{k=1}^{p} n_k$Unidades		$y_{n_2,2}^{(2)}$ $\cdots$	$y_{n_2,p}^{(2)}$	
(El primer subíndice)		$\vdots$	$\vdots$	$\vdots$
			$y_{1,p}^{(p)}$	
			$\vdots$	Posible patrón p
			$y_{n_p,p}^{(p)}$	

Tabla 3-2: Tomada de Liu (1995)

donde $n_k \geq 0$ para $k = 1, ..., p, \sum_{k=1}^{p} n_k = n$ y el superíndice indexa el patrón.

Cuando en (3-4) se incluyen covariables (con estas totalmente observadas), los p posibles patrones de un conjunto de datos incompleto monótono pueden ser representados como se indica en Liu (1996):

$$(\boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}) = \{(y_{i,k}^{(k)}, ..., y_{i,p}^{(k)}, x_{i,1}^{(k)}, ..., x_{i,r}^{(k)}) : i = 1, ..., n_k; k = 1, ..., p\}, \tag{3-5}$$

donde en (3-5) $(\boldsymbol{Y}_{\text{MP}}, \boldsymbol{X})$ representa el conjunto de datos incompleto con patrón monótono (posibles valores faltantes en las variables respuesta, y todas las variables explicativas totalmente observadas). Además, $n_k \geq 0$ para $k = 1, \ldots, p, \sum_{k=1}^{p} n_k = n$ y el superíndice indexa el patrón (un conjunto de datos completos es un caso especial de (3-5) cuando $n_1 = n$ y $n_i = 0$ para $i = 2, \ldots, p$). Esto se ilustra en la Tabla (**3-3**).

	Variables Respuesta				Variables Predictoras		
	$y_{1,1}^{(1)}$	$y_{1,2}^{(1)}$	$\cdots$	$y_{1,p}^{(1)}$	$x_{1,1}^{(1)}$	$\cdots$	$x_{1,r}^{(1)}$
Patrón 1	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
(El superíndice)	$y_{n_1,1}^{(1)}$	$y_{n_1,2}^{(1)}$	$\cdots$	$y_{n_1,p}^{(1)}$	$x_{n_1,1}^{(1)}$	$\cdots$	$x_{n_1,r}^{(1)}$
		$y_{1,2}^{(2)}$	$\cdots$	$y_{1,p}^{(2)}$	$x_{1,1}^{(2)}$	$\cdots$	$x_{1,r}^{(2)}$
Posible patrón 2		$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
		$y_{n_2,2}^{(2)}$	$\cdots$	$y_{n_2,p}^{(2)}$	$x_{n_2,1}^{(2)}$	$\cdots$	$x_{n_2,r}^{(2)}$
			$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
				$y_{1,p}^{(p)}$	$x_{1,1}^{(p)}$	$\cdots$	$x_{1,r}^{(p)}$
Posible patrón p				$\vdots$	$\vdots$	$\cdots$	$\vdots$
				$y_{n_p,p}^{(p)}$	$x_{n_p,1}^{(p)}$	$\cdots$	$x_{n_p,r}^{(p)}$

$n = \sum_{k=1}^{p} n_k$Unidades

(El primer subíndice)

Tabla 3-3: Tomada de Liu (1996)

Los patrones monótonos, donde los datos faltantes univariados es un caso especial, suelen presentarse en conjuntos de datos longitudinales o de medidas repetidas, esto se debe a que, si un sujeto abandona el estudio en un período de tiempo determinado, es probable que falten sus datos en todos los periodos subsiguientes (Carpenter y cols., 2023; Schafer, 1997). Según Schafer (1997), otra situación que conduce a un comportamiento monótono se presenta frecuentemente en el muestreo doble, donde los investigadores intentan medir ciertas variables para todas las unidades de una muestra y, posteriormente, miden variables adicionales solo para una submuestra. Si no hubiera valores faltantes, excepto aquellos que faltan por diseño, los datos serían perfectamente monótonos, sin embargo, en la práctica, suelen surgir faltas adicionales no planificadas que desvían ligeramente el patrón general de la monotonicidad, en este caso, el conjunto de datos no monótono puede convertirse en monótono o casi monótono al reorganizar las variables según sus porcentajes de faltantes (Schafer, 1997). Por lo tanto, si un patrón de datos incompletos no es monótono, las variables pueden ordenarse de manera conveniente según el porcentaje de datos faltantes, lo que también puede resultar en ahorros computacionales significativos (Van Buuren, 2018).

3.2. Clase de Modelos de Regresión Lineal Normal/independiente Multivariados con Datos Faltantes Monótonos

Cuando $\boldsymbol{Y}$ presenta valores faltantes, procederemos bajo la suposición de que el mecanismo de los datos faltantes se puede ignorar (supuesto MAR), según lo definido por R. J. A. Little y Rubin (1987). El modelo dado en las ecuaciones (3-1), (3-2), para datos con patrón monótono, $(\boldsymbol{Y}_{\mathrm{MP}}, \boldsymbol{X})$, puede representarse como sigue (Liu, 1996):

$$\begin{aligned} \boldsymbol{Y}_{i,[k:p]}^{(k)} \mid (\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{w}) &\overset{\text{ind}}{\sim} \mathrm{N}_{p-k+1}((\boldsymbol{\beta}^{(k)})' \boldsymbol{X}_i^{(k)}, \boldsymbol{\Psi}^{(k)}/w_i^{(k)}), \\ w_i^{(k)} \mid \nu &\overset{\text{iid}}{\sim} H(w \mid \nu), \end{aligned} \tag{3-6}$$

para $i = 1, \ldots, n_k$; $k = 1, \ldots, p$, donde $\boldsymbol{Y}_{i,[k:p]}^{(k)}$ representa al i-ésimo individuo perteneciente al patrón k-ésimo con $[k : p]$ variables observadas. Los pesos faltantes están representados por $\boldsymbol{w} = \{w_i^{(k)} : i = 1, \ldots, n_k; k = 1, \ldots, p\}$. Además, $\boldsymbol{\beta}^{(k)}$ se refiere a las últimas $p - k + 1$ columnas de $\boldsymbol{\beta}$, y $\boldsymbol{\Psi}^{(k)}$ es la submatriz inferior derecha $(p - k + 1) \times (p - k + 1)$ de la matriz $\boldsymbol{\Psi}$.

3.2.1. Modelo de Regresión Lineal t Multivariado con Datos Faltantes Monótonos

La distribución t proporciona una extensión útil de la distribución normal para el modelado estadístico de conjuntos de datos que involucran errores con colas más pesadas que la normal (K. L. Lange y cols., 1989). La estimación por máxima verosimilitud de la distribución t multivariada, especialmente cuando los grados de libertad son desconocidos, ha sido un tema de interés en el desarrollo del algoritmo de Expectation-Maximization (EM) (Liu, 1997). Diversos autores (K. L. Lange y cols., 1989; R. J. Little, 1988; Liu, 1997; Rubin, 1983) han abordado este problema desde varias perspectivas metodológicas basadas en la verosimilitud. Han aplicado el modelo t en diversos contextos, como regresión lineal univariada, estimación robusta del vector de medias y la matriz de covarianzas, regresión lineal multivariada con datos completos, y en presencia de datos faltantes, entre otros enfoques. Desde la perspectiva Bayesiana, Liu (1994) propuso una metodología para llevar a cabo la estimación de los parámetros de la distribución t multivariada. Este enfoque es aplicable tanto cuando se incluyen covariables (con estas completamente observadas) como cuando no se incorporan, incluso en presencia de posibles datos faltantes en las variables respuesta. Esta metodología resulta particularmente útil cuando los datos siguen un patrón monótono y el mecanismo de los datos faltantes puede ser ignorado (bajo el supuesto MAR). La implementación de esta metodología se realiza mediante el algoritmo aumento de datos monótonos (MDA, por sus siglas en inglés).

Siguiendo la propuesta de Liu (1994) y asumiendo que el mecanismo de los datos faltantes es ignorable y monótono, representado por $(\boldsymbol{Y}_{\mathrm{MP}}, \boldsymbol{X})$, se establece el siguiente modelo para los datos en (3-1), bajo la suposición de que los errores son independientes e idénticamente distribuidos t multivariados:

$$\boldsymbol{Y}_{i,[k:p]}^{(k)} \equiv (y_{i,k}^{(k)}, \ldots, y_{i,p}^{(k)})' \mid (\boldsymbol{\beta}, \boldsymbol{\Psi}, \nu, \boldsymbol{X}_i^{(k)}) \overset{\text{ind}}{\sim} \mathrm{t}_{p-k+1}((\boldsymbol{\beta}^{(k)})' \boldsymbol{X}_i^{(k)}, \boldsymbol{\Psi}^{(k)}, \nu), \tag{3-7}$$

donde, $\boldsymbol{\beta}^{(k)}$ es la matriz de coeficientes de la regresión con las $p-k+1$ columnas de $\boldsymbol{\beta}$ de $\boldsymbol{Y}$ sobre $\boldsymbol{X}$, $\boldsymbol{\Psi}^{(k)}$, es la submatriz inferior derecha de $\boldsymbol{\Psi}$, y ν son los grados de libertad.

Una forma equivalente de representar el modelo (3-7) (Liu, 1994), aplicable a los datos en la ecuación (3-2), bajo la presencia de datos faltantes monótonos, se presenta de la siguiente manera:

$$\begin{gathered} \boldsymbol{Y}_{i,[k:p]}^{(k)} \mid (\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{w}) \overset{\text{ind}}{\sim} \mathrm{N}_{p-k+1}((\boldsymbol{\beta}^{(k)})' \boldsymbol{X}_i^{(k)}, \boldsymbol{\Psi}^{(k)}/w_i^{(k)}), \\ w_i^{(k)} \mid \nu \overset{\text{iid}}{\sim} \Gamma\ (\nu/2, \nu/2), \end{gathered} \tag{3-8}$$

para $i = 1, \ldots, n_k$ y $k = 1, \ldots, p$, donde $\boldsymbol{w} = \{w_i^{(k)} : i = 1, \ldots, n_k; k = 1, \ldots, p\}$ representan los pesos faltantes.

Para la distribución a priori de los parámetros $\boldsymbol{\theta} = (\boldsymbol{\beta}', \text{vec}(\boldsymbol{\Psi})', \nu)'$ (donde vec es el operador de vectorización), se asume (Liu, 1996) que $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$ y ν son independientes con $P(\boldsymbol{\Psi})$ dado en la ecuación (3-3), $P(\boldsymbol{\beta}) \propto$ constante. En cuanto a la distribución a priori para los grados de libertad ν para el modelo t multivariado, Liu (1995) explora diversas alternativas. En esta instancia, seleccionaremos la propuesta de Liu (1996), que establece $P(\nu) \propto \nu^{-2}$.

3.2.2. Modelo de Regresión Lineal Slash Multivariado con Datos Faltantes Monótonos

La distribución slash, al igual que la distribución t, se presenta como otra alternativa a la distribución normal en el ámbito del modelado estadístico robusto. Esta distribución se distingue por sus colas más pesadas en comparación con la distribución normal, lo que la convierte en una elección apropiada para modelar conjuntos de datos que presentan errores con colas más pesadas de lo que se esperaría bajo la suposición de normalidad (Arslan y Genç, 2009).

Desde la perspectiva Bayesiana; diversos autores (De la Cruz, 2014; Garay y cols., 2015; Rosa y cols., 2003) han abordado problemas específicos, como modelos de regresión lineal mixtos, modelos de regresión lineal multivariados, modelos de regresión no lineal multivariados, entre otros, asumiendo el uso del modelo slash.

Liu (1996) propuso una metodología para realizar la estimación de los parámetros en el contexto del modelo de regresión lineal multivariado con errores independientes e idénticamente distribuidos slash multivariados, considerando además la presencia de datos faltantes con un patrón monótono, con el mecanismo de estos ignorable. Este enfoque se desarrolla desde la perspectiva Bayesiana y utiliza el algoritmo MDA para su implementación.

Siguiendo la metodología propuesta por Liu (1996) y asumiendo que el mecanismo de los datos faltantes es ignorable y monótono, representado por $(\boldsymbol{Y}_{\text{MP}}, \boldsymbol{X})$, se establece el siguiente modelo para los datos en (3-1), bajo la hipótesis de que los errores son independientes e idénticamente distribuidos slash multivariados:

$$\boldsymbol{Y}_{i,[k:p]}^{(k)} \equiv (y_{i,k}^{(k)}, \ldots, y_{i,p}^{(k)})' \mid (\boldsymbol{\beta}, \boldsymbol{\Psi}, \nu, \boldsymbol{X}_i^{(k)}) \overset{\text{ind}}{\sim} \text{SL}_{p-k+1}((\boldsymbol{\beta}^{(k)})' \boldsymbol{X}_i^{(k)}, \boldsymbol{\Psi}^{(k)}, \nu), \tag{3-9}$$

donde, $\boldsymbol{\beta}^{(k)}$ denota la matriz de coeficientes de la regresión, considerando las últimas $p-k+1$ columnas de $\boldsymbol{\beta}$, de $\boldsymbol{Y}$ sobre $\boldsymbol{X}$, $\boldsymbol{\Psi}^{(k)}$, representa la submatriz inferior derecha de $\boldsymbol{\Psi}$, y ν, como indicado por Morán-Vásquez y cols. (2021), representa el parámetro de cola de la distribución slash multivariada.

Una forma equivalente de representar el modelo (3-9) (Liu, 1996), para los datos en la ecuación (3-2), bajo la presencia de datos faltantes monótonos, es el siguiente:

$$\begin{aligned} \boldsymbol{Y}_{i,[k:p]}^{(k)} \mid (\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{X}, \boldsymbol{w}) &\overset{\text{ind}}{\sim} \mathrm{N}_{p-k+1}((\boldsymbol{\beta}^{(k)})' \boldsymbol{X}_i^{(k)}, \boldsymbol{\Psi}^{(k)}/w_i^{(k)}), \\ w_i^{(k)} \mid \nu &\overset{\text{iid}}{\sim} \text{Beta } (\nu, 1), \end{aligned} \tag{3-10}$$

para $i = 1, \ldots, n_k$ y $k = 1, \ldots, p$, donde $\boldsymbol{w} = \{w_i^{(k)} : i = 1, \ldots, n_k; k = 1, \ldots, p\}$ representan los pesos faltantes.

Para la distribución a priori de los parámetros $\boldsymbol{\theta} = (\boldsymbol{\beta}', \text{vec}(\boldsymbol{\Psi})', \nu)'$, se asume, según Liu (1996), que $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$ y ν son independientes con $P(\boldsymbol{\Psi})$ dado en la ecuación (3-3), $P(\boldsymbol{\beta}) \propto$ constante. Además, Liu (1996) sugiere utilizar una distribución a priori para el parámetro de cola ν una $\Gamma\ (a, b)$, la cual es conjugada con respecto a $\boldsymbol{w}$, con valores positivos pequeños de a y b (a siendo mucho mayor que $b, b \ll a$).

3.3. Algoritmo de Aumento de Datos Monótonos (MDA) para Obtener Extracciones de los Parámetros a partir de su Distribución Posterior

Numerosos algoritmos estadísticos que incorporan la técnica de aumento de datos han ganado amplia aceptación en diversas disciplinas como las ciencias biológicas, médicas, físicas, sociales, ingeniería, entre otras (Van Dyk y Meng, 2010). Estos algoritmos abarcan tanto versiones determinísticas, como el algoritmo EM (Dempster, Laird, y Rubin, 1977) y sus numerosas extensiones, como versiones estocásticas, entre las que se incluye el algoritmo de aumento de datos (DA, por sus siglas en inglés) (Schafer, 1997; Van Dyk y Meng, 2010).

La popularización de esta técnica para construir algoritmos determinísticos se le atribuye a Dempster y cols. (1977), quien introdujo el algoritmo EM en su artículo semilla. No obstante, el término *aumento de datos* se originó con el algoritmo de aumento de datos propuesto por Tanner y Wong (1987), que ofrece una ilustración destacada de esta técnica en un entorno de simulación (Van Dyk y Meng, 2001). La estrategia fundamental de estos algoritmos radica en simplificar un problema complejo descomponiéndolo en una secuencia iterativa de problemas más manejables.

La idea del aumento de datos se desarrolló como un enfoque para abordar problemas de datos faltantes, especialmente desde la perspectiva Bayesiana, donde el objetivo principal es calcular la distribución posterior de los parámetros de interés (Tanner y Wong, 1987). El DA es un algoritmo iterativo pertenecientes a los métodos de Cadenas de Markov Monte Carlo, el cual su objetivo es aproximar cierta distribución posterior de los parámetros de interés por medio de una cadena de Markov (Schafer, 1997).

En muchos problemas con datos incompletos, la distribución posterior observada $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}})$, donde $\boldsymbol{Y}_{\text{obs}}$ represeta los datos observados, resulta intratable y difícil de resumir o simular de manera directa (Schafer, 1997). La idea central del DA es abordar esta dificultad. Según Tanner y Wong (1987), los datos observados $\boldsymbol{Y}_{\text{obs}}$ son aumentados por una cantidad asumida para las observaciones faltantes ($\boldsymbol{Y}_{\text{mis}}$), donde la distribución $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$ es más fácil de manejar. Sin embargo, el objetivo es obtener la distribución posterior deseada $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}})$, la cual ya se ha mencionado que es complicada de tratar directamente. Si se logra generar valores de $\boldsymbol{Y}_{\text{mis}}$ desde su distribución predictiva $P(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}})$, entonces es posible aproximar $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}})$. Este proceso se realiza de manera iterativa, generando muestras $\boldsymbol{Y}_{\text{mis}}$ desde $P(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}})$ y $\boldsymbol{\theta}$ desde la distribución posterior aumentada $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$ (Tanner y Wong, 1987).

Asumiendo que se cumplen las condiciones necesarias para la convergencia (Tanner y Wong, 1987), el resultado obtenido mediante el DA es una secuencia $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{Y}_{\text{mis}}^{(t)}) : t = 1, 2, \dots\}$ con distribución estacionaria $P(\boldsymbol{\theta}, \boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{Y}_{\text{obs}})$ (Schafer, 1997). En consecuencia, la secuencia generada, al ser una cadena de Markov, presenta la propiedad de ser una muestra dependiente. No obstante, para un número suficientemente grande de iteraciones, $\boldsymbol{\theta}^{(t)}$ puede considerarse como una extracción aleatoria de $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}})$ (Schafer, 1997).

En consecuencia, para implementar el algoritmo, se debe realizar el muestreo desde dos distribuciones: $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$ y $P(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{\theta}, \boldsymbol{Y}_{\text{obs}})$ (Tanner y Wong, 1987). Esto da lugar al siguiente algoritmo iterativo (Tanner y Wong, 1987), que consta de dos pasos, I-step (paso de imputación) y P-step (paso de simulación posterior), donde la iteración t-ésima está dada por:

- **I- step**. Simular $\boldsymbol{Y}_{\text{mis}}^{(t)}$ desde su distribución predictiva posterior

$$\boldsymbol{Y}_{\text{mis}}^{(t)} \sim P(\boldsymbol{Y}_{\text{mis}} \mid \boldsymbol{\theta}, \boldsymbol{Y}_{obs}),$$

- **P -step**. Extraer $\boldsymbol{\theta}^{(t+1)}$ desde la distribución posterior aumentada

$$\boldsymbol{\theta}^{(t+1)} \sim P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}^{(t)}).$$

Se puede inferir del proceso iterativo anterior que el objetivo del I-step es imputar los datos faltantes suficientes para llevar el P-step a una simulación posterior de datos completos (Schafer, 1997). Li (1988) observó que no es necesario imputar todos los datos faltantes presentes en el conjunto de datos en cada I-step, de manera que el P-step sea una simulación posterior de datos completos. Su observación se centró en que imputar solo una parte selectiva de los datos faltantes puede ser suficiente para mantener la estructura o patrón de los datos originales, evitando así una imputación innecesaria de valores que no contribuyen significativamente a la calidad de los resultados. Basándose en esta observación, Rubin y Schafer (1990) propusieron un método eficiente llamado Aumento de Datos Monótonos

(MDA), que aprovecha la estructura de datos faltantes monótonos, especialmente para datos normales multivariados. Con MDA, ya no es necesario imputar el conjunto completo de los datos faltantes en el I-step para llevar el P-step a una simulación de datos completos (Schafer, 1997).

Como se mencionó anteriormente, según Schafer (1997), en la práctica, normalmente el conjunto de datos faltantes no tiene una estructura monótona, pero se puede llevar a una estructura cercana a la monótona, incluyendo algunos datos faltantes que destruyen dicho patrón monótono. Sea $\boldsymbol{Y}_{\text{mis}}$ el conjunto de los datos faltantes. Este conjunto puede ser particionado en $(\boldsymbol{Y}_{\text{MP,mis}}, \boldsymbol{Y}_{\text{mis}^*})$, donde $\boldsymbol{Y}_{\text{MP,mis}}$ representa el conjunto de datos faltantes que destruyen el patrón monótono, y $\boldsymbol{Y}_{\text{mis}^*}$ representa el conjunto de datos faltantes que no es necesario imputar, dado que solo estamos interesados en completar el patrón monótono. Por lo tanto, la idea del MDA es modificar el I-step del DA ordinario, imputando solo el conjunto de datos faltantes que destruyen el patrón monótono ($\boldsymbol{Y}_{\text{MP,mis}}$). El algoritmo iterativo MDA procede en los siguientes dos pasos (Liu, 1996; Rubin y Schafer, 1990; Schafer, 1997):

- **I step**. Simular $\boldsymbol{Y}_{\text{MP,mis}}$ desde su distribución predictiva posterior

$$\boldsymbol{Y}_{\text{MP,mis}} \sim P(\boldsymbol{Y}_{\text{MP,mis}} \mid \boldsymbol{\theta}, \boldsymbol{Y}_{\text{obs}}),$$

- **P step**. Extraer $\boldsymbol{\theta}$ desde la distribución posterior aumentada

$$\boldsymbol{\theta} \sim P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{MP,mis}}).$$

El MDA presenta dos grandes ventajas computacionales sobre el DA ordinario (Liu, 1994; Schafer, 1997). La primera es que las extracciones secuenciales obtenidas por el MDA están menos correlacionadas que las obtenidas por el DA, ya que en la simulación de los parámetros intervienen menos valores faltantes. Por lo tanto, las extracciones aleatorias por MDA son más eficaces para la estimación mediante la técnica de Monte Carlo que las obtenidas por DA (Liu, 1994). La segunda ventaja es que el MDA tiene una tasa de convergencia más rápida, es decir, alcanza la estacionariedad aproximada en menos iteraciones (Liu, 1994). Según Schafer (1997), la convergencia del DA está gobernada por la cantidad de información contenida en $\boldsymbol{Y}_{\text{mis}}$ en relación a $\boldsymbol{Y}_{\text{obs}}$, mientras que con el MDA, la convergencia está gobernada por la cantidad de información contenida en $\boldsymbol{Y}_{\text{MP,mis}}$ en relación a $\boldsymbol{Y}_{\text{obs}}$. Con un patrón monótono perfecto, el MDA simula los parámetros directamente a partir de su distribución posterior marginal; en consecuencia, con un patrón monótono, el MDA no requiere un muestreo iterativo y alcanza la convergencia en una iteración (Liu, 1994). Cuando el conjunto de datos no tiene un patrón monótono o no está lejos de este, en primer lugar, el MDA crea un patrón monótono que contiene todos los valores observados y algunos valores faltantes que destruyen el patrón monótono (Liu, 1994). La distribución $P(\boldsymbol{\theta} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{MP,mis}})$ es entonces casi independiente de $\boldsymbol{Y}_{\text{MP,mis}}$, dado que la proporción es relativamente pequeña

(Schafer, 1997). A continuación, se aplica el algoritmo iterativo DA a estos datos faltantes monótonos construidos, y solo serán necesarios unos pocos pasos del MDA para lograr una estacionariedad aproximada (Liu, 1994; Schafer, 1997).

Para identificar los $\boldsymbol{Y}_{\text{MP,mis}}$ es útil ordenar las filas y columnas de la matriz de datos. En muchos casos, según Schafer (1997), no existe un conjunto único de $\boldsymbol{Y}_{\text{MP,mis}}$ que complete un patrón monótono. Para eficiencia computacional, es ventajoso que $\boldsymbol{Y}_{\text{MP,mis}}$ sea pequeño por dos razones: (1) para reducir el número de extracciones aleatorias en cada I-step y (2) $\boldsymbol{Y}_{\text{MP,mis}}$ debe contener la menor información posible sobre los parámetros desconocidos, para reducir el número de pasos necesarios para alcanzar la estacionariedad aproximada (Schafer, 1997). Según (Schafer, 1997), encontrar un conjunto $\boldsymbol{Y}_{\text{MP,mis}}$ para maximizar la eficiencia del algoritmo es un problema difícil, por lo que, sugiere el enfoque de simplemente ordenar las columnas de $\boldsymbol{Y}_{\text{obs}}$ según sus proporciones de observaciones faltantes.

3.3.1. Implementación del Algoritmo MDA para Extraer Muestras de los Parámetros de los Modelos de Regresión Lineal con Distribución Normal/independiente Multivariadas

En aras de la claridad, siguiendo a Liu (1996), vamos a explicar el MDA, para tomar extracciones de los parámetros de los modelos desde su distribución posterior, paso a paso:

1. **Patrón Monótono Perfecto con Pesos Conocidos**: Cuando el conjunto de datos observados $\boldsymbol{Y}_{\text{obs}}$ tiene un patrón monótono perfecto $\boldsymbol{Y}_{\text{MP}}$ y los pesos conocidos $\boldsymbol{w}$.

2. **Pesos Conocidos pero Sin Patrón Monótono**: Cuando los pesos $\boldsymbol{w}$ son conocidos pero el conjunto de datos observados $\boldsymbol{Y}_{\text{obs}}$ no tiene un patrón monótono.

3. **Pesos Desconocidos pero Hiperparámetros Conocidos**: Cuando los pesos $\boldsymbol{w}$ son desconocidos, pero los hiperparámetros son conocidos $\boldsymbol{\nu}$.

4. **Pesos e Hiperparámetros Desconocidos**: Cuando los pesos $\boldsymbol{w}$ son desconocidos y los hiperparámetros $\boldsymbol{\nu}$ también son desconocidos.

Algoritmo MDA con Pesos Conocidos y Datos Observados con Patrón Monótono Perfecto

Con datos observados que presentan un patrón monótono perfecto y pesos conocidos, el algoritmo MDA no es iterativo (Liu, 1994, 1996; Schafer, 1997). Por lo tanto, la simulación posterior de los parámetros $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ puede llevarse a cabo por medio de la ecuación:

$$P(\boldsymbol{\beta}, \boldsymbol{\Psi} \mid \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}, \boldsymbol{w}) = P(\boldsymbol{\Psi} \mid \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}, \boldsymbol{w}) \times P(\boldsymbol{\beta} \mid \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}, \boldsymbol{w}).$$

Se han propuesto diferentes técnicas para tomar extracciones de los parámetros $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ desde su distribución posterior cuando el conjunto de datos tiene patrón monótono. Por ejemplo, el enfoque de la factorización de la verosimilitud propuesto por Rubin y Schafer (1990), el cual reduce el problema de inferencia a una secuencia de regresiones de datos completos sobre subconjuntos de las filas de la matriz de datos (Schafer, 1997). Otro algoritmo propuesto por Tang (2015), es un método de muestreo de la distribución Wishart bajo la presencia de datos faltantes monótonos, representando la matriz aleatoria distribuida Wishart como una función de vectores aleatorios multivariados independientes normal-gamma.

Para nuestros propósitos, nos basaremos en el enfoque propuesto por Liu (1996), el cual extiende los resultados de Liu (1993) y Liu (1995). Liu (1996), presenta una forma fácil de simular la matriz de dispersión y la matriz de coeficientes de los modelos de regresión lineal asociado a la clase de distribuciones normal/independiente multivariadas a través de una extensión de la descomposición de Bartlett (Bartlett, 1934), cuando el conjunto de datos tiene patrón monótono y el mecanismo de los datos faltantes es ignorable.

En consecuencia, para tomar extracciones desde $P(\boldsymbol{\Psi} \mid \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}, \boldsymbol{w})$ y $P(\boldsymbol{\beta} \mid \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}, \boldsymbol{w})$, se tiene el Teorema 1 presentado por Liu (1996), que está definido de la siguiente forma:

Teorema 1. Para $k = 1, \ldots, p$ (los p posibles patrones), sea

$$\hat{\boldsymbol{\beta}}^{(k)} = \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i^{(j)} \boldsymbol{X}_i^{(j)} (\boldsymbol{X}_i^{(j)})' \right)^{-1} \times \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i^{(j)} \boldsymbol{X}_i^{(j)} (\boldsymbol{Y}_{i,[k:p]}^{(j)})' \right)$$

$$= (\boldsymbol{X}' \boldsymbol{\Lambda}_k \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{\Lambda}_k \boldsymbol{Y}_{[k:p]},$$

donde $\boldsymbol{\Lambda}_k = \text{diag}\, \{w_1^{(1)}, \ldots, w_{n_1}^{(1)}, \ldots, w_1^{(k)}, \ldots, w_{n_k}^{(k)}, 0, \ldots, 0\}$ y $\boldsymbol{Y}_{[k:p]}$ es la submuestra de las últimas $p - k + 1$ columnas de $\boldsymbol{Y}$. Por lo tanto, $\hat{\boldsymbol{\beta}}^{(k)}$ son las estimaciones ponderadas por mínimos cuadrados de los coeficientes de regresión de $\boldsymbol{Y}_{[k:p]}$ sobre $\boldsymbol{X}$ basados en la muestra $\{(\boldsymbol{Y}_{i,[k:p]}^{(j)}, \boldsymbol{X}_i^{(j)}) : i = 1, \ldots, n_j; j = 1, \ldots, k\}$. Sea $\boldsymbol{S}_k$ la correspondiente k-ésima suma total ponderada de cuadrados residuales y productos cruzados, es decir,

$$\boldsymbol{S}_k = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_i^{(j)} \left(\boldsymbol{Y}_{i,[k:p]}^{(j)} - (\hat{\boldsymbol{\beta}}^{(k)})' \boldsymbol{X}_i^{(j)} \right) \times \left(\boldsymbol{Y}_{i,[k:p]}^{(j)} - (\hat{\boldsymbol{\beta}}^{(k)})' \boldsymbol{X}_i^{(j)} \right)',$$

y

$$\boldsymbol{B}_k = \boldsymbol{A}_k + \boldsymbol{S}_k,$$

donde $\boldsymbol{A}_k$ es la submatriz triangular inferior $(p - k + 1) \times (p - k + 1)$ de la matriz definida no negativa $\boldsymbol{A}$ dada en (3-3). Suponga que $\boldsymbol{B}_1$ es definida positiva ($\boldsymbol{B}_1 > 0$), y así para $k = 1, \ldots, p$, $\boldsymbol{B}_k$ es definida positiva y tiene factorización de Cholesky $\boldsymbol{B}_k^{-1} = \boldsymbol{L}_k \boldsymbol{L}_k'$, donde $\boldsymbol{L}_k$ es una matriz triangular inferior. Sea $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p)$ una matriz triangular inferior $p \times p$ con su parte trinagular inferior dada por las columnas $\boldsymbol{L}_1 \boldsymbol{u}_1, \boldsymbol{L}_2 \boldsymbol{u}_2, \ldots, \boldsymbol{L}_p \boldsymbol{u}_p$, donde

$\boldsymbol{u}_k = (u_{k,k}, \ldots, u_{p,k})'$ con $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ satisfaciendo (a) cada $u_{i,j}$ son independientes para $1 \leq j \leq i \leq p$, (b) cada $u_{i,j} \sim N(0,1)$ para $1 \leq j < i \leq p$, y (c) $u_{j,j}^2 \sim \chi^2_{n_1+n_2+\cdots+n_j-j+(m-p-q+1)}$ para $j = 1, \ldots p$. Si $n_1 + n_2 + \cdots + n_k > k + (p - m - 1 + q)$ para $k = 1, \ldots, p$, entonces, dado $\boldsymbol{w}, \boldsymbol{\nu}$, y $\boldsymbol{Y}_{\text{MP}}$ la distribución posterior condicional de $\boldsymbol{\Psi}^{-1}$ es distribuida como $\boldsymbol{H}\boldsymbol{H}'$.

Por lo tanto, basados en el Teorema 1 de Liu (1996) descrito anteriormente, para tomar una extracción de $\boldsymbol{\Psi}$ desde su distribución posterior $P(\boldsymbol{\Psi} \mid \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}, \boldsymbol{w})$, solo es generar una muestra $\boldsymbol{\Psi} = (\boldsymbol{H}\boldsymbol{H}')^{-1}$.

Otro hecho importante que se desprende del teorema anteriormente descrito, presentado por Liu (1996), es el Corolario 2, el cual muestra cómo extraer de forma fácil una muestra desde la distribución posterior de la matriz de coeficientes del modelo $\boldsymbol{\beta}$, $P(\boldsymbol{\beta} \mid \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}, \boldsymbol{w})$. Siguiendo a Liu (1996), este está dado de la siguiente manera:

Corolario 2. Sea $\boldsymbol{Z} = (\boldsymbol{Z_1}, \ldots, \boldsymbol{Z}_p)$ una matriz aleatoria $(r \times p)$ cuyos elementos están distribuidos independientes normal estándar. Entonces, la distribución condicional de $\boldsymbol{\beta}$ dado $\boldsymbol{\Psi}, \boldsymbol{w}$ y $(\boldsymbol{Y}_{\text{MP}}, \boldsymbol{X})$, queda distribuida de la siguiente manera:

$$(\boldsymbol{G}_1\boldsymbol{Z}_1, \ldots, \boldsymbol{G}_p\boldsymbol{Z}_p)\boldsymbol{H}^{-1} + (\hat{\boldsymbol{\beta}}^{(1)}\boldsymbol{h}_1, \ldots, \hat{\boldsymbol{\beta}}^{(p)}\boldsymbol{h}_p)\boldsymbol{H}^{-1},$$

donde $\boldsymbol{G}_k\boldsymbol{G}_k{}'$ proviene de la factorización de Cholesky de la matriz definida positiva $(\boldsymbol{X}'\boldsymbol{\Lambda}_k\boldsymbol{X})^{-1}$, con $\boldsymbol{G}_k$ siendo una matriz triangular inferior, para $k = 1, \ldots, p$, y donde $\boldsymbol{H}$ proviene de $\boldsymbol{H}\boldsymbol{H}'$.

Algoritmo MDA con Pesos Conocidos y Datos Observados con Patrón no Monótono

Cuando el conjunto de datos no tiene un patrón monótono, en primer lugar, el MDA crea un patrón monótono $\boldsymbol{Y}_{\text{MP}}$ que contiene todos los valores observados $\boldsymbol{Y}_{\text{obs}}$ y algunos valores faltantes que destruyen el patrón monótono $\boldsymbol{Y}_{\text{MP,mis}}$ (Liu, 1994, 1996; Schafer, 1997). Así, siguiendo a Liu (1996), $\boldsymbol{Y}_{\text{MP}} = \{\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{MP,mis}}\}$, donde una iteración completa del MDA en esta situación queda dada por los siguientes dos pasos:

- **I-step**. Llenar en los datos faltantes $\boldsymbol{Y}_{\text{MP,mis}}$ con una extracción desde $P(\boldsymbol{Y}_{\text{MP,mis}} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{w})$, el cual es un distribución normal multivariada.
- **P-step**. Simular $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ desde $P(\boldsymbol{\beta}, \boldsymbol{\Psi} \mid \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}_{\text{MP}}, \boldsymbol{w})$, como un MDA con pesos conocidos y datos observados con patrón monótono perfecto.

Dado que cada una de las filas de la matriz de datos observados $\boldsymbol{Y}_{\text{obs}}$, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, son independientes, el I-step puede llevarse a cabo de manera independiente para cada $\boldsymbol{Y}_i^{(k)}$, con $i = 1, \ldots, n_k$ y $k = 1, \ldots, p$. Por lo tanto, esto se reduce a la generación sucesiva de vectores aleatorios normales multivariados para cada fila donde haya datos faltantes que destruyen el patrón monótono. Luego, basándonos en la propiedad de la distribución normal multivariada de que esta mantiene la familia bajo marginalización, el valor faltante de interés a rellenar se

toma del vector aleatorio generado. Por lo tanto, el modelo predictivo para cada observación es:

$$\boldsymbol{Y}_{i,[k:p]}^{(k)} \overset{\text{ind}}{\sim} N_{p-k+1}((\boldsymbol{\beta}^{(k)})'\boldsymbol{X}_i^{(k)}, \boldsymbol{\Psi}^{(k)}/w_i^{(k)}),$$

para $i = 1, \ldots, n_k$ y $k = 1, \ldots, p$. Luego, el $Y_{i,MP,mis}^{(k)}$ de interés se extrae del vector aleatorio generado para dicha observación, y así se repite el proceso hasta completar el patrón monótono.

Algoritmo MDA con Pesos Desconocidos y Hiperparámetros Conocidos

Cuando los pesos son desconocidos y los hiperparámetros son conocidos, siguiendo a Liu (1996), cada iteración del algoritmo MDA consiste de los siguientes dos pasos:

- **I-step**. Imputar los pesos desconocidos $\boldsymbol{w}$ con una extracción desde su distribución posterior $P(\boldsymbol{w} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\nu})$. Luego, extraer los datos faltantes $\boldsymbol{Y}_{\text{MP,mis}}$ (si $\boldsymbol{Y}_{\text{MP,mis}} \neq \varnothing$), como un MDA con pesos conocidos.

- **P-step**. Simular $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ desde $P(\boldsymbol{\beta}, \boldsymbol{\Psi} \mid \boldsymbol{Y}_{\text{MP}}, \boldsymbol{X}_{\text{MP}}, \boldsymbol{w})$, como un MDA con pesos conocidos y datos observados con patrón monótono perfecto.

Para tomar extracciones desde $P(\boldsymbol{w} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\nu})$, usamos el hecho de que dado $\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}$ y $\boldsymbol{\nu}$, $w_i^{(k)} (i = 1, \ldots, n_k; k = 1, \ldots, p)$ son independientes y,

$$P(w_i^{(k)} = w \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\nu}) \propto w^{p_i^{(k)}/2} \exp\{-w(\delta_{i,\text{obs}}^2)^{(k)}/2\} f(w \mid \boldsymbol{\nu}), \tag{3-11}$$

donde $p_i^{(k)}$ es la dimensión de las componentes observadas $\boldsymbol{Y}_{i,\text{obs}}^{(k)}$ de $\boldsymbol{Y}_i^{(k)}$,

$$(\delta_{i,\text{obs}}^2)^{(k)} = (\boldsymbol{Y}_{i,\text{obs}}^{(k)} - \boldsymbol{\mu}_{i,\text{obs}}^{(k)})'(\boldsymbol{\Psi}_{i,\text{obs}}^{(k)})^{-1}(\boldsymbol{Y}_{i,\text{obs}}^{(k)} - \boldsymbol{\mu}_{i,\text{obs}}^{(k)}),$$

donde $\boldsymbol{\mu}_{i,\text{obs}}^{(k)}$ consiste de las componentes de $\boldsymbol{\beta}'\boldsymbol{X}_i^{(k)}$ que corresponden a los componentes observados de $\boldsymbol{Y}_i^{(k)}$, y $\boldsymbol{\Psi}_{i,\text{obs}}^{(k)}$ es la submatriz de $\boldsymbol{\Psi}$ correspondiente a la matriz de dispersión de las componentes observadas de $\boldsymbol{Y}_i^{(k)}$. Por lo tanto, una extracción de $\boldsymbol{w}$ puede ser obtenida tomando n extracciones independientes, cada una desde la distribución unidimensional posterior (3-11).

Claramente, para el modelo t multivariado (Liu, 1996):

$$w_i^{(k)} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \nu \sim \Gamma\ (\nu/2 + p_i^{(k)}/2, \nu/2 + (\delta_{i,\text{obs}}^2)^{(k)}/2). \tag{3-12}$$

La distribución posterior (3-12) se desprende del hecho de que:

$$
\begin{aligned}
P(w_i^{(k)} = w \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \nu) &\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} h(w \mid \nu) \\
&\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} \times \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\
&\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} w^{\frac{\nu}{2}-1} \exp\{-\frac{\nu w}{2}\}}{\Gamma\left(\frac{\nu}{2}\right)} \\
&\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} w^{\frac{\nu}{2}-1} \exp\left\{-\frac{\nu w}{2}\right\} \\
&\propto w^{\frac{p_i^{(k)}}{2}+\frac{\nu}{2}-1} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2} - \frac{\nu w}{2}\right\} \\
&\propto w^{\frac{p_i^{(k)}}{2}+\frac{\nu}{2}-1} \exp\left\{-w\left(\frac{(\delta^2_{i,\text{obs}})^{(k)}}{2} + \frac{\nu}{2}\right)\right\}
\end{aligned}
$$

lo cual es igual a una $\Gamma(p_i^{(k)}/2+\nu/2, (\delta^2_{i,\text{obs}})^{(k)}/2+\nu/2)$, donde $p_i^{(k)}$ corresponde a las variables respuestas que el i-ésimo individuo tiene observada, y $(\delta^2_{i,\text{obs}})^{(k)} = (\boldsymbol{y}_{i,\text{obs}}^{(k)} - \boldsymbol{\mu}_{i,\text{obs}}^{(k)})'(\boldsymbol{\Psi}_{i,\text{obs}}^{(k)})^{-1}(\boldsymbol{y}_{i,\text{obs}}^{(k)} - \boldsymbol{\mu}_{i,\text{obs}}^{(k)})$ es la distancia de Mahalanobis del i-ésimo individuo respecto a su media, con las respectivas unidades observadas.

Para el modelo slash multivariado, la distribución posterior de $\{w_i^{(k)} \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \nu\}$ es modelada como una distribución gama incompleta en el intervalo $(0, 1]$. La función de distribución posterior es proporcional a (Liu, 1996):

$$
w^{(p_i^{(k)}/2)+\nu-1} \exp\{-w(\delta^2_{i,\text{obs}})^{(k)}/2\}, \qquad (0 < w \leq 1,\ \nu > 0). \tag{3-13}
$$

(3-13) se desprende del siguiente proceso:

$$
\begin{aligned}
P(w_i^{(k)} = w \mid \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \nu) &\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} h(w \mid \nu) \\
&\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} \times \text{Beta}\,(\nu, 1) \\
&\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} \nu w^{\nu-1} \\
&\propto w^{\frac{p_i^{(k)}}{2}} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\} w^{\nu-1}
\end{aligned}
$$

$$\propto w^{\frac{p_i^{(k)}}{2}+v-1} \exp\left\{-\frac{w(\delta^2_{i,\text{obs}})^{(k)}}{2}\right\},$$

$(0 < w \leq 1, \nu > 0)$, la cual corresponde a una gamma incompleta (Γ_{0-1}) (Liu, 1996).

En el enfoque propuesto por Liu (1996), se propone la utilización de extracciones aleatorias mediante las técnicas descritas por Schmeiser y Lal (1980). No obstante, en nuestro estudio, hemos optado por emplear el paquete gamlss (Stasinopoulos y cols., 2023), el cual simplificó la creación de una función de distribución gamma truncada en el intervalo $(0, 1)$. Adicionalmente, llevamos a cabo una reparametrización de la función de densidad para adaptarla a nuestros objetivos específicos, asegurando así una extracción adecuada de las muestras.

Algoritmo MDA con Pesos Desconocidos y Hiperparámetros Desconocidos

Según Liu (1994), cuando los hiperparámetros son desconocidos, existen tres posibles enfoques para obtener la estimación de máxima verosimilitud de los mismos mediante el uso de EM (expectation maximization) (Dempster y cols., 1977), ECM (Expectation Conditional Maximization) (Meng y Rubin, 1993) y ECME (Expectation Conditional Maximization Either) (Liu y Rubin, 1995). En consecuencia, se derivan tres versiones del MDA para llevar a cabo la simulación posterior de los hiperparámetros (Liu, 1994). Sin embargo, la versión del MDA correspondiente al algoritmo ECM no será de interés, ya que este requiere la imputación de los valores faltantes $\{\boldsymbol{Y}_{\text{MP,mis}}, \boldsymbol{w}\}$ dos veces en cada iteración completa para visitar cada variable al menos una vez (Liu, 1995).

Por lo tanto, según la propuesta de Liu (1996), existen dos formas sencillas de agrupar las variables $\boldsymbol{w}, \boldsymbol{Y}_{\text{MP,mis}}, \boldsymbol{\beta}, \boldsymbol{\Psi}$, y $\boldsymbol{\nu}$ para utilizar el MDA en el proceso de extracción de parámetros y valores faltantes. La primera forma consiste en particionar estas variables en dos grupos: $\{\boldsymbol{Y}_{\text{MP,mis}}, \boldsymbol{w}\}$ y $\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\nu}$. Dado que $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ y $\boldsymbol{\nu}$ son condicionalmente independientes dado $\{\boldsymbol{Y}_{\text{MP,mis}}, \boldsymbol{w}\}$, el algoritmo MDA correspondiente para este caso puede implementarse en los siguientes dos pasos:

- **I-step**. Imputar $\{\boldsymbol{Y}_{\text{MP,mis}}, \boldsymbol{w}\}$ como un MDA con pesos desconocidos y hiperparámetros conocidos.
- **P-step**. Simular $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ como un MDA con pesos desconocidos y hiperparámetros conocidos y extraer $\boldsymbol{\nu}$ desde su distribución posterior $P(\boldsymbol{\nu} \mid \boldsymbol{w})$.

La versión previamente mencionada del MDA se alinea con el algoritmo EM (Dempster y cols., 1977) utilizado para las estimaciones de máxima verosimilitud de los parámetros. Liu (1996) denominó a esta versión del MDA como la *versión EM*.

La segunda forma implica la partición de las variables en dos grupos: $\{\boldsymbol{\nu}, \boldsymbol{w}, \boldsymbol{Y}_{\text{MP,mis}}\}$ y $(\boldsymbol{\beta}, \boldsymbol{\Psi})$. En este enfoque, se sustituye el paso de la versión EM, que extrae $\boldsymbol{\nu}$ de $P(\boldsymbol{\nu} \mid \boldsymbol{w})$, con el paso que extrae $\boldsymbol{\nu}$ de $P(\boldsymbol{\nu} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Psi})$. Liu (1996) denominó a esta versión del MDA como la *versión ECME* para las estimaciones de máxima verosimilitud de los parámetros. Se destaca que la versión EM del MDA es más sencilla de implementar, pero converge de manera más lenta en comparación con la versión ECME (Liu, 1995).

Liu (1995) discute cómo llevar a cabo el proceso de extracción de los grados de libertad desde su distribución posterior para el modelo t multivariado, la cual está dada por:

$$P(\nu \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Psi}) \propto \exp\{-\log(P(\nu)) + \sum_{k=1}^{p}\sum_{i=1}^{n_k} \log\left(\Gamma\left(\frac{\nu + p_i^{(k)}}{2}\right)\right) - n\log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{n\nu}{2}\log(\nu) - \sum_{k=1}^{p}\sum_{i=1}^{n_k} \frac{\nu + p_i^{(k)}}{2}\log(\nu + (\delta^2_{i,\text{obs}})^{(k)})\}, \quad (3\text{-}14)$$

donde $(\delta^2_{i,\text{obs}})^{(k)} = (\boldsymbol{y}^{(k)}_{i,\text{obs}} - \boldsymbol{\mu}^{(k)}_{i,\text{obs}})'(\boldsymbol{\Psi}^{(k)}_{\boldsymbol{i},\text{obs}})^{-1}(\boldsymbol{y}^{(k)}_{i,\text{obs}} - \boldsymbol{\mu}^{(k)}_{i,\text{obs}})$ es la distancia de Mahalanobis del i-ésimo individuo respecto a vector de localización, considerando las respectivas componentes observadas. El proceso de extracción de los grados de libertad desde su distribución posterior, cuando se trabaja bajo la versión ECME del MDA, utiliza un método que combina el griddy-sampler (Ritter y Tanner, 1992) y el método de aceptación-rechazo, de tal forma que, utilizar la técnica griddy para obtener una función lineal a trozos que domine la densidad verdadera y, a continuación, utilizar el método de aceptación-rechazo para tomar extracciones de la densidad verdadera (Liu, 1995). Sin embargo, para nuestros propósitos, trabajando bajo la versión ECME del MDA, aplicamos el método de la rejilla (la cual fue uniformemente espaciada y mantuvimos el número de puntos constantes a lo largo de las iteraciones) propuesto por Ritter y Tanner (1992), el cual consiste en: (1) evaluar $P(\nu \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Psi})$ en la rejilla de puntos $\nu_1, \ldots, \nu_n$ para obtener $c_1, ..., c_n$, (2) usar $c_1, ..., c_n$ para obtener una aproximación a la función de distribución inversa de $P(\nu \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Psi})$, y (3) muestrear una muestra $U(0, 1)$ y transformar la observación mediante la función de distribución inversa aproximada. Algunos comentarios a tener en cuenta al momento de aplicar el método de la rejilla (Ritter y Tanner, 1992) son: (1) la función $P(\nu \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Psi})$ sólo necesita conocerse hasta una constante de proporcionalidad, ya que la normalización puede obtenerse directamente a partir de $c_1, ..., c_n$, (2) la rejilla $c_1, ..., c_n$ no necesita estar uniformemente espaciada, y (3) el número de puntos de la rejilla no tiene por qué ser constante a lo largo de las iteraciones.

Para la distribución slash multivariada, en línea con la propuesta de Liu (1996), adoptamos una distribución gamma (a, b) como priori para el parámetro de cola ν. Esta elección resulta ser una a priori conjugada con respecto a la verosimilitud de $\boldsymbol{w}$. Según la sugerencia de Liu (1996), se recomienda seleccionar valores pequeños y positivos para a y b

(a siendo mucho mayor que $b, b \ll a$). La distribución posterior de ν dado $\boldsymbol{w}$ está dada por (Liu, 1996):

$$\nu \mid \boldsymbol{w} \sim \Gamma \left(\mathrm{a} + \mathrm{n}, \mathrm{b} - \sum_{k=1}^{p} \sum_{i=1}^{n_k} \log(w_i^{(k)}) \right). \tag{3-15}$$

(3-15) se deriva del siguiente procedimiento:

$$\begin{aligned}
P(\nu \mid \boldsymbol{w}) &\propto \nu^{\mathrm{a}-1} \exp\{-\mathrm{b}\nu\} \prod_{k=1}^{p} \prod_{i=1}^{n_k} \nu (w_i^{(k)})^{\nu-1} \\
&\propto \nu^{\mathrm{a}-1} \exp\{-\mathrm{b}\nu\} \nu^{\mathrm{n}} \prod_{k=1}^{p} \prod_{i=1}^{n_k} (w_i^{(k)})^{\nu-1} \\
&\propto \nu^{\mathrm{a}+\mathrm{n}-1} \exp\{-\mathrm{b}\nu\} \prod_{k=1}^{p} \prod_{i=1}^{n_k} \exp\{\log((w_i^{(k)})^{\nu-1})\} \\
&\propto \nu^{\mathrm{a}+\mathrm{n}-1} \exp\{-\mathrm{b}\nu\} \exp\left\{ \sum_{k=1}^{p} \sum_{i=1}^{n_k} \log((w_i^{(k)})^{\nu-1}) \right\} \\
&\propto \nu^{\mathrm{a}+\mathrm{n}-1} \exp\{-\mathrm{b}\nu\} \exp\left\{ \sum_{k=1}^{p} \sum_{i=1}^{n_k} (\nu-1) \log(w_i^{(k)}) \right\} \\
&\propto \nu^{\mathrm{a}+\mathrm{n}-1} \exp\{-\mathrm{b}\nu\} \exp\left\{ \sum_{k=1}^{p} \sum_{i=1}^{n_k} \nu \log(w_i^{(k)}) \right\} \\
&\propto \nu^{\mathrm{a}+\mathrm{n}-1} \exp\{-\mathrm{b}\nu\} \exp\left\{ \nu \sum_{k=1}^{p} \sum_{i=1}^{n_k} \log(w_i^{(k)}) \right\} \\
&\propto \nu^{\mathrm{a}+\mathrm{n}-1} \exp\left\{ -\mathrm{b}\nu + \nu \sum_{k=1}^{p} \sum_{i=1}^{n_k} \log(w_i^{(k)}) \right\} \\
&\propto \nu^{\mathrm{a}+\mathrm{n}-1} \exp\left\{ -\nu \left(\mathrm{b} - \sum_{k=1}^{p} \sum_{i=1}^{n_k} \log(w_i^{(k)}) \right) \right\}
\end{aligned}$$

la cual corresponde a una $\Gamma\ (\mathrm{a} + \mathrm{n}, \mathrm{b} - \sum_{k=1}^{p} \sum_{i=1}^{n_k} \log(w_i^{(k)}))$.

(3-15) corresponde a la versión EM del algoritmo MDA, el cual, es mucho más sencillo de aplicar (Liu, 1996).

En el siguiente capítulo se presenta la metodología que enlaza los modelos expuestos en este capítulo con la obtención de cuantiles marginales de las variables respuesta.

4 Modelado de Cuantiles a través de la Clase de Modelos de Regresión Lineal Normal/independiente Multivariados

Los conjuntos de datos multivariados continuos positivos son frecuentemente encontrados en el ámbito práctico. Es ampliamente reconocido que estos datos, al ser continuos y positivos, tienden a exhibir asimetría positiva y ocasionalmente contienen observaciones atípicas, como se ha señalado en estudios anteriores (Ferrari y Fumes, 2017).

A pesar de estas características distintivas, el análisis estadístico de tales conjuntos de datos a menudo se sustenta en los supuestos de la distribución normal multivariada (Morán-Vásquez y cols., 2021). Este enfoque puede resultar limitado, ya que tiende a pasar por alto las particularidades inherentes a este tipo de datos (Morán-Vásquez y Ferrari, 2019). Una metodología alternativa para modelar datos positivos multivariados involucra la aplicación de la transformación Box-Cox (Box y Cox, 1964) a cada componente del vector de observaciones (Morán-Vásquez y Ferrari, 2019). No obstante, en este enfoque (Quiroz, Nakamura, y Pérez, 1996) se asume que el vector resultante de las observaciones transformadas sigue una distribución normal multivariada o una distribución elíptica. Esta suposición implica una deficiencia teórica porque el soporte del vector transformado de observaciones no es necesariamente $\mathbb{R}^p$, además, los parámetros del modelo solo son interpretativos en términos de las características de las observaciones transformadas, y no proporcionan información directa sobre las variables originales de interés (Morán-Vásquez y Ferrari, 2019).

Ferrari y Fumes (2017), propusieron la clase de distribuciones simétricas de Box-Cox, que resulta útil para modelizar datos sesgados positivos, posiblemente con colas pesadas, en el ámbito univariado, la cual incluye distribuciones tales como, Box-Cox t, Box-Cox Cole-Green (o Box-Cox normal), Box-Cox exponencial de potencia y la clase de las distribuciones log-simétricas (Vanegas y Paula, 2016) como casos especiales: log-normal, log-student-t, Birnbaum Saunders, Birnbaum-Saunders-t y Birnbaum-Saunders generalizada. Morán-Vásquez y Ferrari (2019), extendieron la clase de distribuciones simétricas Box-Cox al ámbito multivariado, a la cual llamaron clase elíptica de distribuciones Box-Cox. Esta clase de distribuciones proporciona alternativas para modelizar datos multivariados positivos, marginalmente sesgados y posiblemente datos con colas pesadas, la cual, tiene como caso especial la clase de

distribuciones log-elípticas (Morán-Vásquez y Ferrari, 2019). Una característica distintiva y práctica de la clase Box-Cox elíptica radica en la interpretabilidad de sus parámetros, que se vinculan con cuantiles y dispersiones relativas de las distribuciones marginales y las asociaciones entre pares de variables; esta relación entre los parámetros de escala y los cuantiles las hace particularmente atractivas para la aplicación en modelos de regresión (Morán-Vásquez y Ferrari, 2019).

Una subclase significativa dentro de las distribuciones Box-Cox elípticas (Morán-Vásquez y Ferrari, 2019), también considerada como una subclase de las distribuciones log-elípticas, utilizadas para la modelización de conjuntos de datos multivariados positivos, es la clase de distribuciones log-normal/independiente multivariadas (Morán-Vásquez y cols., 2021). Según Morán-Vásquez y cols. (2021), esta subclase es especialmente atractiva para la modelización estadística robusta, ya que ofrece diversas distribuciones con colas pesadas y soporte positivo, además, presentan propiedades teóricas específicas que no se cumplen en toda la clase log-elíptica. La clase de distribuciones log-normal/independiente multivariadas es apropiada para modelar datos positivos multivariados correlacionados que son sesgados y posiblemente de cola pesada, la cual, tiene como algunos de sus miembros: las distribuciones log-normal multivariada, log-t y log-slash, entre otras; además, otra característica atractiva es la fácil interpretación de sus parámetros en términos del vector de variables de interés y su relación con los cuantiles de las distribuciones marginales (Morán-Vásquez y cols., 2021).

4.1. La Clase de Distribuciones Log-Normal/independiente Multivariadas

Un vector aleatorio $\boldsymbol{Z} \in \mathbb{R}_+^p$, conforme a la definición 2 proporcionada por Morán-Vásquez y cols. (2021), tiene una distribución log-normal/independiente con vector de mediana $\boldsymbol{\mu} \in \mathbb{R}_+^p$ y matriz de dispersión $\boldsymbol{\Psi}(p \times p) > 0$ si $\boldsymbol{Z} = \boldsymbol{D_\mu} \boldsymbol{V}^{1/\sqrt{W}}$, donde $\boldsymbol{V} \sim \mathrm{LN}_p(\mathbf{1}, \boldsymbol{\Psi})$ y W son independientes, con W siendo una variable aleatoria positiva con función de distribución acumulada $H(\cdot \mid \boldsymbol{\nu})$ y donde $\boldsymbol{\nu} \in \mathbb{R}^q$ es un vector de parámetros extra inducido por H. Escribimos $\boldsymbol{Z} \sim \mathrm{LNI}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, H)$.

En la anterior definición (Morán-Vásquez y cols., 2021), $\mathbf{1}$ representa el vector con sus componentes todos unos de dimensión $(p \times 1)$, $\boldsymbol{D_\mu}$ representa una matriz diagonal $(p \times p)$, con su diagonal dada por $\mathrm{diag}\{\mu_1, \ldots, \mu_p\}$.

Siguiendo a Morán-Vásquez y cols. (2021), equivalentemente, $\boldsymbol{Z} \sim \mathrm{LNI}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, H)$ si $\boldsymbol{Y} \sim \mathrm{NI}_p(\log(\boldsymbol{\mu}), \boldsymbol{\Psi}, H)$, $\boldsymbol{Y} = \log(\boldsymbol{Z})$, lo cual establece la forma en la cual las distribuciones log-normal/independiente y normal/independiente están relacionadas a través de la transformación logarítmica. Otras representaciones estocásticas (Morán-Vásquez y cols., 2021) equivalentes para $\boldsymbol{Z} \sim \mathrm{LNI}_p\ (\boldsymbol{\mu}, \boldsymbol{\Psi}, H)$ son: $\boldsymbol{Z} \mid W = w \sim \mathrm{LN}_p\ (\boldsymbol{\mu}, w^{-1}\boldsymbol{\Psi})$, $W \sim H\ (w \mid \boldsymbol{\nu})$,

y $\log(\boldsymbol{Z}) \mid W = w \sim \mathrm{N}_p\ (\log(\boldsymbol{\mu}), w^{-1}\boldsymbol{\Psi})$, $W \sim H\ (w \mid \boldsymbol{\nu})$. La función de densidad de probabilidad (Morán-Vásquez y cols., 2021) de $\boldsymbol{Z} \sim \mathrm{LNI}_p\ (\boldsymbol{\mu}, \boldsymbol{\Psi}, H)$ es dada por:

$$\mathrm{LNI}_p(\boldsymbol{z} \mid \boldsymbol{\mu}, \boldsymbol{\Psi}, \boldsymbol{\nu}) = \int_0^\infty \mathrm{LN}_p(\boldsymbol{z} \mid \boldsymbol{\mu}, w^{-1}\boldsymbol{\Psi}) h(w \mid \boldsymbol{\nu})\ dw, \tag{4-1}$$

donde $\mathrm{LN}_p(\boldsymbol{\mu}, \boldsymbol{\Psi})$ es la función de densidad de probabilidad de la distribución log-normal.

Esta subclase tiene miembros que son atractivos para la modelización estadística robusta, ya que tienen colas más pesadas que la distribución log-normal multivariada, por ejemplo, las distribuciones log-t y log-slash multivariadas; además, esta subclase disfruta de varias propiedades que no necesariamente satisfacen todas las distribuciones log-elípticas, como la preservación de las familias bajo marginalización (Morán-Vásquez y cols., 2021). Según Morán-Vásquez y cols. (2021), existe una correspondencia entre las clases de distribuciones multivariadas normal/independiente y log normal/independiente según la distribución de la variable aleatoria W. Así, si consideramos una distribución degenerada en $w = 1$ para W, obtenemos la función de densidad de $\boldsymbol{Z} \sim \mathrm{LN}_p(\boldsymbol{\mu}, \boldsymbol{\Psi})$. Cuando $W \sim \Gamma(\nu/2, \nu/2)$, obtenemos la función de densidad de probabilidad de $\boldsymbol{Z} \in \mathbb{R}^p_+$ con distribución log-t multivariada con vector de mediana $\boldsymbol{\mu} \in \mathbb{R}^p_+$, matriz de dispersión $\boldsymbol{\Psi}(p \times p) > 0$ y parámetro de grados de libertad $\nu > 0$. Si $W \sim \mathrm{Beta}(\nu, 1)$, entonces obtenemos la función de densidad de probabilidad de un vector aleatorio $\boldsymbol{Z} \in \mathbb{R}^p_+$ con distribución log-slash multivariada con vector de mediana $\boldsymbol{\mu} \in \mathbb{R}^p_+$, matriz de dispersión $\boldsymbol{\Psi}(p \times p) > 0$ y parámetro de cola $\nu > 0$. El parámetro ν controla el comportamiento de la cola en las distribuciones multivariadas log-t y log-slash, por lo que, estas distribuciones presentan colas más pesadas en comparación con la distribución log-normal multivariada para valores pequeños de ν, lo que hace que las distribuciones log-t y log-slash, sean adecuadas para modelar datos positivos sesgados multivariados con posibles valores atípicos (Morán-Vásquez y cols., 2021).

Según Morán-Vásquez y cols. (2021), el Teorema 5 de Morán-Vásquez y Ferrari (2019) es válido para la clase de distribuciones log-normal/independiente multivariadas. Este teorema (Morán-Vásquez y Ferrari, 2019) establece que estas distribuciones preservan la familia bajo marginalización, lo cual no ocurre en toda la clase de distribuciones log-elípticas, es decir, que si $\boldsymbol{Z} \sim \mathrm{LNI}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, H)$, entonces $Z_k \sim \mathrm{LNI}_1(\mu_k, \sigma^2_{kk}, H)$, para $k = 1, \ldots, p$, además, a partir del Corolario 2 de Morán-Vásquez y Ferrari (2019), se obtiene que el α-cuantil $z_{k,\alpha}$ de Z_k, $\alpha \in (0, 1)$, satisface $z_{k,\alpha} = \mu_k \exp\{\sqrt{\sigma^2_{kk}} q_\alpha\}$, donde q_α es el α-cuantil de la variable normal/independiente univariada estándar $Q \sim \mathrm{NI}_1(0, 1, H)$, lo que significa que todos los cuantiles de las distribuciones marginales univariadas de $\boldsymbol{Z} \sim \mathrm{LNI}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, H)$ son proporcionales a los respectivos componentes del vector de mediana $\boldsymbol{\mu}$.

4.2. La Clase de Modelos de Regresión Lineal Log-Normal/independiente Multivariados

Siguiendo a Morán-Vásquez y cols. (2021), sean $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ vectores aleatorios que representan observaciones de $\boldsymbol{Z} \in \mathbb{R}_+^p$ sobre n individuos, donde $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})'$, $i = 1, \ldots, n$. El componente Z_{ik} representa la respuesta del i-ésimo individuo para la variable k-ésima. Las componentes de $\boldsymbol{Z}_i$ son posiblemente correlacionadas y los vectores $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ son independientes.

La clase de modelos de regresión lineal log-normal/independiente multivariados se define según Morán-Vásquez y cols. (2021) como:

$$\begin{cases} \boldsymbol{Z}_i \overset{\text{ind}}{\sim} \text{LNI}_p(\boldsymbol{\mu}_i, \boldsymbol{\Psi}, H), \\ \quad \log(\boldsymbol{\mu}_i) = \boldsymbol{\beta}' \boldsymbol{X}_i, \end{cases} \tag{4-2}$$

para $i = 1, \ldots, n$, donde $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ir})'$ es la i-ésima fila de la matriz modelo $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]'$. El vector $\boldsymbol{X}_i$ contiene los valores del i-ésimo individuo para las r variables explicativas $x_1, \ldots, x_r$. Por lo tanto, x_{ij} es el valor observado del i-ésimo individuo en la j-ésima variable explicativa, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})' \in \mathbb{R}_+^p$ es el vector de mediana de $\boldsymbol{Z}_i$, $\boldsymbol{\Psi}(p \times p) > 0$ es la matriz de dispersión y $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ es una matriz de coeficientes de regresión $r \times p$, con $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{rk})'$, $k = 1, \ldots, p$, β_{jk} corresponde a x_{ij}, con $x_{i1} = 1$, $i = 1, \ldots, n$, $j = 1, \ldots, r$.

Los modelos de regresión lineal in (4-2) (Morán-Vásquez y cols., 2021) permiten modelar la relación entre los cuantiles de los componentes del vector aleatorio $\boldsymbol{Z}$ y el conjunto de variables explicativas $x_1, \ldots, x_r$, facilitando la interpretación de los coeficientes de regresión, de hecho, como consecuencia (Morán-Vásquez y cols., 2021) del Corolario 2 de Morán-Vásquez y Ferrari (2019), el α-cuantil $z_{k,\alpha}$ de Z_k, $\alpha \in (0, 1)$, es dado por:

$$z_{k,\alpha} = \exp\left\{ \sum_{j=1}^{r} \beta_{jk} x_j + \sqrt{\sigma_{kk}^2} q_\alpha \right\}, \tag{4-3}$$

para $k = 1, \ldots, p$, donde q_α es el α-cuantil de la variable aleatoria normal/independiente univariada estándar $Q \sim \text{NI}_1(0, 1, H)$. En particular, $z_{k,1/2} = \exp\left\{ \sum_{j=1}^{r} \beta_{jk} x_j \right\}$, esto es, la mediana es afectada solo por las variables explicativas a través de la función exponencial, además, de (4-3) tenemos que si x_j se incrementa en una unidad manteniendo las demás variables explicativas fijas, entonces $z_{k,\alpha}$ es multiplicada por el factor $\exp\{\beta_{jk}\}$ (Morán-Vásquez y cols., 2021).

Así, (Morán-Vásquez y cols., 2021) la estimación $z_{k,\alpha_1}, \ldots, z_{k,\alpha_m}$, $k = 1, \ldots, p$, requiere las estimaciones de $\boldsymbol{\beta}, \boldsymbol{\Psi}$ y $\boldsymbol{\nu}$ obtenidas en un solo ajuste del modelo (4-2), y los cálculos separados de $q_{\alpha_1}, \ldots, q_{\alpha_m}$. La regresión cuantílica lineal univariada no requiere supuestos distribucionales, mientras que el modelo (4-2) supone que el vector de respuesta tiene una distribución que pertenece a la clase de distribuciones log-normal/independiente multivariada (Morán-Vásquez y cols., 2021).

El modelo (4-2) (Morán-Vásquez y cols., 2021) es equivalente a $\log(\boldsymbol{Z}_i) \overset{\text{ind}}{\sim} \text{NI}_p(\boldsymbol{\beta}'\boldsymbol{X}_i, \boldsymbol{\Psi}, H)$, $i = 1, \ldots, n$, el cual puede expresarse como:

$$\begin{aligned} \log(\boldsymbol{Z}_i) \mid W_i = w_i &\overset{\text{ind}}{\sim} \text{N}_p(\boldsymbol{\beta}'\boldsymbol{X}_i, w_i^{-1}\boldsymbol{\Psi}), \\ W_i &\overset{\text{iid}}{\sim} H(w_i \mid \boldsymbol{\nu}), \end{aligned} \tag{4-4}$$

para $i = 1, \ldots, n$. Por lo tanto, los parámetros en (4-2) pueden estimarse mediante el ajuste de un modelo de regresión lineal normal/independiente multivariado (K. Lange y Sinsheimer, 1993; Liu, 1996) utilizando el vector de respuesta $\log(\boldsymbol{Z}) = \boldsymbol{Y}$ (Morán-Vásquez y cols., 2021). En el contexto bayesiano, las especificaciones distribucionales a prior de los parámetros son necesarias para la inferencia a posterior. Este proceso se llevará a cabo de acuerdo con lo discutido en el capítulo anterior.

Cada miembro de la clase de modelos de regresión lineal log-normal/independiente multivariados (Morán-Vásquez y cols., 2021) está determinado por la función de distribución de H en (4-2), proporcionando varias alternativas para el modelado estadístico a través de la regresión lineal multivariada basada en distribuciones sesgadas multivariadas con soporte positivo, por ejemplo, el modelo de regresión lineal log-normal multivariado (H es la función de distribución de un vector aleatorio W con distribución degenerada en $w = 1$), el modelo de regresión lineal multivariado log-t (H es la función de distribución de un vector aleatorio w con función de densidad de probabilidad $\Gamma(\nu/2, \nu/2)$), el modelo de regresión lineal multivariado log-slash (H es la función de distribución de un vector aleatorio W con función de densidad Beta$(\nu, 1)$). Otros miembros son los modelos de regresión lineal multivariados log normal contaminado, log-Pearson tipo VII, log Laplace, entre otros (Morán-Vásquez y cols., 2021). Además, según Morán-Vásquez y cols. (2021), es sencillo mostrar que si $\boldsymbol{Z} \sim \text{LNI}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, H)$, entonces $\delta^2 = \delta^2(\text{Log}(\boldsymbol{Z}); \text{Log}(\boldsymbol{\mu}), \boldsymbol{\Psi})$ tiene la misma función de distribución de (2-6).

En el siguiente capítulo se evalúan los algoritmos MDA, el proceso de estimación de los modelos expuestos en el capítulo 3 y el método de obtención de los cuantiles marginales presentado en este capítulo. Todo lo anterior, obtenido a través de estudios de simulación realizados a partir de diversos escenarios.

5 Estudios de Simulación

En este capítulo, se presentan los resultados del estudio de simulación que se llevó a cabo con el fin de evaluar el proceso de aproximación de la distribución posterior y la estimación de la mediana posterior, bajo la presencia de datos faltantes, de los parámetros del modelo (4-2), haciendo uso del modelo (3-6). Así mismo, se examinó la estimación de los cuartiles en relación con el modelo (4-2), por medio del modelo (3-6). Para ello, se crearon distintos escenarios de simulación, implicando la ejecución de tres situaciones para cada modelo considerado (modelo t multivariado y modelo slash multivariado). Estas simulaciones se llevaron a cabo bajo la premisa de $\boldsymbol{Z}_i \overset{\text{ind}}{\sim} \text{LNI}_3(\boldsymbol{\mu}_i, \boldsymbol{\Psi}, H)$, donde $\log(\mu_{ik}) = \sum_{j=0}^{3} \beta_{jk} x_{ij}$, con $k = 1, 2, 3$, y $i = 1, \ldots, n$. Note que, a diferencia de la notación utilizada en el modelo (3-1), la suma comienza en 0 y no en 1 con el fin de denotar al intercepto de la manera más comúnmente utilizada en la literatura. Las simulaciones fueron programadas y ejecutadas utilizando el lenguaje de programación estadístico R. El código está disponible en `https://github.com/joseescobara/MDA-algorithm`.

5.1. Metodología

Los parámetros reales o verdaderos fueron obtenidos ajustando el modelo slash trivariado y el modelo t trivariado al logaritmo del conjunto de datos de los niños, que son descritos en el capítulo 6. El vector de respuesta $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$ en cada escenario de simulación es una imitación del logaritmo de las variables perímetro braquial (Y_1), peso (Y_2) y talla (Y_3) de los niños. Las covariables se generaron como extracciones aleatorias independientes de diversas distribuciones (cuyos parámetros fueron seleccionados con base en el conjunto de datos reales utilizado en la aplicación descrita en el capítulo 6) y se conservaron constantes durante todas las simulaciones:

- X_0: vector de unos representando la variable "dummy" relacionada con el parámetro constante β_{0k} para $k = 1, 2, 3$.
- X_1 (edad): sigue una distribución gamma con parámetro de forma $\alpha = 2.16585$ y parámetro de escala $\beta = 1/1.308587$. Estos dos parámetros se derivaron de la variable edad en la base de datos real. En este proceso, calculamos el promedio y la varianza de la variable edad. Luego, igualamos el valor esperado y la varianza de la distribución gamma a esos respectivos valores y resolvimos el sistema de ecuaciones, obteniendo así

los valores para α y la tasa. El parámetro de escala β se obtiene al dividir 1 entre la tasa, esta relación surge de la parametrización de la distribución gamma en el lenguaje estadístico R.

- X_2 (sexo; 0 para mujer, 1 para hombre): sigue una distribución Bernoulli con probabilidad de éxito igual a 0.6127168. La probabilidad de éxito se obtuvo dividiendo el número total de hombres de la variable sexo entre el total de la muestra.

- X_3 (tiempo de leche materna; en semanas): sigue una distribución gamma con parámetro de forma $\alpha = 1.814255$ y parámetro de escala $\beta = 1/0.1975243$. El procedimiento empleado para obtener ambos valores fue análogo al utilizado para la variable de edad.

Generamos 1000 conjuntos de datos para las familias de distribuciones t y slash trivariadas, considerando tamaños muestrales de $n = 50$, $n = 100$ y $n = 150$. Con el fin de lograr la máxima similitud entre los conjuntos de datos simulados y el conjunto de datos real, se tuvo en cuenta de la premisa de que el conjunto de datos real contenía varios valores atípicos. Se buscó garantizar la inclusión de datos atípicos en los conjuntos de datos simulados, destacando la robustez de la metodología frente a la presencia de estos valores. Este aspecto se alineó con la intención de demostrar la eficacia de utilizar distribuciones de colas más pesadas que la normal.

Para llevar a cabo esta tarea, se estableció un criterio específico de selección de los conjuntos de datos simulados. Este criterio se basó en la suma de las distancias de Mahalanobis de los vectores de variables respuesta de cada conjunto de datos simulado con respecto a la suma de las distancias de Mahalanobis de los vectores de variables respuesta del conjunto de datos real, proporcional al tamaño muestral del escenario de simulación correspondiente. Este enfoque aseguró la incorporación de datos atípicos, fortaleciendo así la validez y representatividad de los conjuntos generados.

Con el propósito de evaluar el desempeño de la metodología propuesta en presencia de datos faltantes, donde las variables explicativas estaban completamente observadas y las posibles ausencias de datos podrían ocurrir en las variables respuesta, se llevó a cabo el proceso de inclusión de estos datos faltantes en cada conjunto de datos simulados, seleccionados previamente según el criterio mencionado. Este proceso se desarrolló de la siguiente manera:

1. Considerando la proporción de datos faltantes en cada variable respuesta del conjunto de datos real.

2. Atendiendo a los distintos patrones de datos faltantes presentes en el conjunto de datos real: (obs, obs, obs), (mis, obs, obs), (obs, mis, obs), (obs, obs, mis), (mis, mis, obs), (obs, mis, mis), (mis, obs, mis), siendo faltante igual a *mis* y observado igual a *obs*.

3. Manteniendo la simetría del conjunto de datos reales. A pesar de la presencia de algunos datos atípicos en el conjunto real, la mayoría de las distancias de Mahalanobis de los niños (considerando solo las componentes observadas) se situaban muy cerca de su media real, oscilando entre 0.001 y 5.

Por lo tanto, en la selección de los datos a los cuales se les agregaron datos faltantes, se consideraron las distancias de Mahalanobis. Es decir, aquellos datos con distancias de Mahalanobis entre 1 y 15 fueron seleccionados para la inclusión de datos faltantes. Esto se realizó con dos objetivos principales: (1) preservar la integridad de los datos atípicos y (2) reducir las distancias de Mahalanobis para lograr la mayor simetría posible en el conjunto de datos.

Siguiendo a Morán-Vásquez y cols. (2021), adoptamos la desviación absoluta mediana (MAD, por sus siglas en inglés) como la métrica para evaluar el rendimiento de las estimaciones generadas por los modelos empleados. El MAD está definido de la siguiente manera: Sea θ un parámetro escalar y $\hat{\theta}_1, \ldots, \hat{\theta}_N$ sean valores estimados ordenados obtenidos a partir de N muestras Monte Carlo simuladas. La desviación absoluta mediana para $\{\hat{\theta}_1, \ldots, \hat{\theta}_N\}$ es definida como la mediana de $\{\mid \hat{\theta}_1 - M(\hat{\theta}) \mid, \ldots, \mid \hat{\theta}_N - M(\hat{\theta}) \mid\}$, donde $M(\hat{\theta})$ es la mediana de $\{\hat{\theta}_1, \ldots, \hat{\theta}_N\}$.

5.2. Escenarios de Simulación Modelo t Multivariado

A continuación, se presentan los escenarios considerados en la simulación para el modelo t trivariado.

5.2.1. Escenario 1

En este escenario, se generaron 1000 conjuntos de datos simulados, cada uno con un tamaño muestral de 50 observaciones. La selección de estos conjuntos se llevó a cabo de acuerdo el criterio previamente establecido. Específicamente, para el modelo t, se calculó la suma total de las distancias de Mahalanobis de los vectores de variables respuesta, con respecto a las componentes observadas para el conjunto de datos real (como ya se mencionó antes, con tamaño muestral de 173), resultando en una suma total de de distancias Mahalanobis igual a 644. Por lo tanto, se estableció que las muestras seleccionadas (sin datos faltantes) debían tener una suma de distancias de Mahalanobis superior a $(50 \times 644) \ / \ 173 = 186$, respecto a los parámetros reales. Es importante señalar que esta suma aún no incluía la contribución de los datos faltantes.

Posteriormente, a cada muestra seleccionada se le incorporaron los datos faltantes de acuerdo con el procedimiento previamente mencionado. Se realizaron selecciones adicionales, optando por aquellas muestras que presentaban una suma de las distancias de Mahalanobis alrededor de 186 (176 a 186) o superior a 186 después de la inclusión de los datos faltantes.

5.2.2. Escenario 2

En este escenario, se generaron 1000 conjuntos de datos simulados, cada uno con un tamaño muestral de 100 observaciones. La selección de estos conjuntos se hizo de manera análoga a lo mencionado en el escenario 1, a diferencia de que, las muestras seleccionadas (sin datos faltantes) debían tener una suma de distancias de Mahalanobis superior a (100 × 644) / 173 = 372, respecto a los parámetros reales. Es importante señalar que esta suma aún no incluía la contribución de los datos faltantes.

Posteriormente, a cada muestra seleccionada se le incorporaron los datos faltantes de acuerdo con el procedimiento previamente mencionado. Se realizaron selecciones adicionales, optando por aquellas muestras que presentaban una suma de las distancias de Mahalanobis alrededor de 372 (352 a 372) o superior a 372 después de la inclusión de los datos faltantes.

5.2.3. Escenario 3

En este escenario, se generaron 1000 conjuntos de datos simulados, cada uno con un tamaño muestral de 150 observaciones. La selección de estos conjuntos se hizo de manera análoga a lo mencionado en el escenario 1 y 2, a diferencia de que, las muestras seleccionadas (sin datos faltantes) debían tener una suma de distancias de Mahalanobis superior a (150 × 644) / 173 = 558, respecto a los parámetros reales. Es importante señalar que esta suma aún no incluía la contribución de los datos faltantes.

Posteriormente, a cada muestra seleccionada se le incorporaron los datos faltantes de acuerdo con el procedimiento previamente mencionado. Se realizaron selecciones adicionales, optando por aquellas muestras que presentaban una suma de las distancias de Mahalanobis alrededor de 558 (540 a 558) o superior a 558 después de la inclusión de los datos faltantes.

5.3. Escenarios de Simulación Modelo Slash Multivariado

A continuación, se presentan los escenarios considerados en la simulación para el modelo slash trivariado.

5.3.1. Escenario 1

En este escenario, se generaron 1000 conjuntos de datos simulados, cada uno con un tamaño muestral de 50 observaciones. La selección de estos conjuntos se llevó a cabo de acuerdo el criterio previamente establecido. Específicamente, para el modelo slash, se calculó la suma total de las distancias de Mahalanobis de los vectores de variables respuesta, con relación a las componentes observadas para el conjunto de datos real (el cual tenía un tamaño muestral de 173), resultando en una suma total de distancias de Mahalanobis igual a 845. Por lo tanto,

para mantener la proporcionalidad de dicha medida en un base de datos de 50 observaciones, se estableció que las muestras seleccionadas (sin datos faltantes) debían tener una suma de distancias de Mahalanobis superior a $(50 \times 845) / 173 = 244$, respecto a los parámetros reales. Es importante señalar que esta suma aún no incluía el problema de los datos faltantes.

Posteriormente, a cada muestra seleccionada se le incorporaron los datos faltantes de acuerdo con el procedimiento previamente mencionado. Se realizaron selecciones adicionales, optando por aquellas muestras que presentaban una suma de las distancias de Mahalanobis alrededor de 244 (224 a 244) o superior a 244 después de la inclusión de los datos faltantes.

5.3.2. Escenario 2

En este escenario, se generaron 1000 conjuntos de datos simulados, cada uno con un tamaño muestral de 100 observaciones. La selección de estos conjuntos se hizo de manera análoga a lo mencionado en el escenario 1, a diferencia de que, las muestras seleccionadas (sin datos faltantes) debían tener una suma de distancias de Mahalanobis superior a $(100 \times 845) / 173 = 488$, respecto a los parámetros reales. Es importante señalar que esta suma aún no incluía la contribución de los datos faltantes.

Posteriormente, a cada muestra seleccionada se le incorporaron los datos faltantes de acuerdo con el procedimiento previamente mencionado. Se realizaron selecciones adicionales, optando por aquellas muestras que presentaban una suma de las distancias de Mahalanobis alrededor de 488 (469 a 488) o superior a 488 después de la inclusión de los datos faltantes.

5.3.3. Escenario 3

En este escenario, se generaron 1000 conjuntos de datos simulados, cada uno con un tamaño muestral de 150 observaciones. La selección de estos conjuntos se hizo de manera análoga a lo mencionado en el escenario 1 y 2, a diferencia de que, las muestras seleccionadas (sin datos faltantes) debían tener una suma de distancias de Mahalanobis superior a $(150 \times 845) / 173 = 733$, respecto a los parámetros reales. Es importante señalar que esta suma aún no incluía la contribución de los datos faltantes.

Posteriormente, a cada muestra seleccionada se le incorporaron los datos faltantes de acuerdo con el procedimiento previamente mencionado. Se realizaron selecciones adicionales, optando por aquellas muestras que presentaban una suma de las distancias de Mahalanobis alrededor de 733 (713 a 733) o superior a 733 después de la inclusión de los datos faltantes.

5.4. Proceso de Simulación

El proceso de simulación, destinado a obtener la distribución posterior de los parámetros de interés, así como las estimaciones correspondientes y la evaluación de los cuartiles, se ejecutó de la siguiente manera:

1. A cada muestra generada, para cada uno de los escenarios asociados a los modelos utilizados (modelo t y slash), se implementó el proceso iterativo del algoritmo MDA. En este procedimiento, se extrajo una muestra de tamaño 10000 de los parámetros de interés desde su distribución posterior. Se aplicó un período de Burn-in de 1000 iteraciones con el propósito de mitigar el impacto de los valores iniciales, asegurando así que las extracciones pudieran considerarse como muestras independientes.

2. Para cada una de las muestras independientes extraídas de los 1000 conjuntos de datos, se calcularon las estimaciones de interés mediante el cálculo de la mediana posterior. La elección de la mediana como medida central aseguró una evaluación robusta, ya que esta es menos sensible a valores extremos en comparación con otras medidas de tendencia central, como la media. Así, esto proporcionó el cálculo de la distribución posterior de los parámetros de interés.

3. En relación con cada una de las estimaciones obtenidas en los 1000 conjuntos de datos, se aplicó el cálculo de la mediana. Este procedimiento permitió obtener la estimación final para cada escenario de simulación respectivo.

4. Para la estimación de los diferentes cuartiles, se utilizaron las estimativas obtenidas en el paso 2. Para cada una de estas estimativas, se calcularon los cuartiles (25, 50, 75). Denotamos por $z_{k,\alpha,0}$ y $z_{k,\alpha,1}$, donde $k = 1, 2, 3$ y $\alpha \in (0, 1)$, representando el α-cuantil de la distribución marginal de Z_k cuando $x_2 = 0$ y $x_2 = 1$, respectivamente. Las variables edad y tiempo de leche materna se fijaron en su respetiva media.

5. Las estimaciones finales de los cuartiles se obtuvieron al aplicar la mediana a los cuartiles estimados en los 1000 conjuntos de datos.

5.5. Resultados

En esta sección, presentamos los resultados de los seis escenarios del estudio de simulación. Para cada uno de ellos, se aproximaron las distribuciones posteriores de los coeficientes de los modelos, las estimaciones de los parámetros con sus correspondientes MAD, y los cuartiles junto con sus respectivos MAD. Este análisis permitirá una comprensión detallada de las características y variabilidad de los resultados en cada escenario, ofreciendo una visión integral del desempeño de los modelos y sus parámetros, así como la estimación de los cuartiles.

5.5.1. Resultados de los 3 Escenarios de Simulación Modelo t Multivariado

La Figura (**5-1**) ilustra la comparación entre las aproximaciones de las distribuciones posteriores para la muestra de tamaño 50 y la distribución posterior aproximada con la base de datos real del modelo t trivariado. De manera similar, las Figuras (**5-2**) y (**5-3**) exhiben la comparación de las aproximaciones de las distribuciones posteriores para las muestras de tamaño 100 y 150, respectivamente, con la distribución posterior aproximada con la base de datos real.

Se puede evidenciar que, a medida que aumenta el tamaño de la muestra, las aproximaciones de las distribuciones posteriores de los coeficientes del modelo convergen gradualmente hacia la distribución posterior aproximada con la base de datos real. Este comportamiento sugiere una mejora en la precisión de las aproximaciones a medida que se incrementa el tamaño de la muestra.

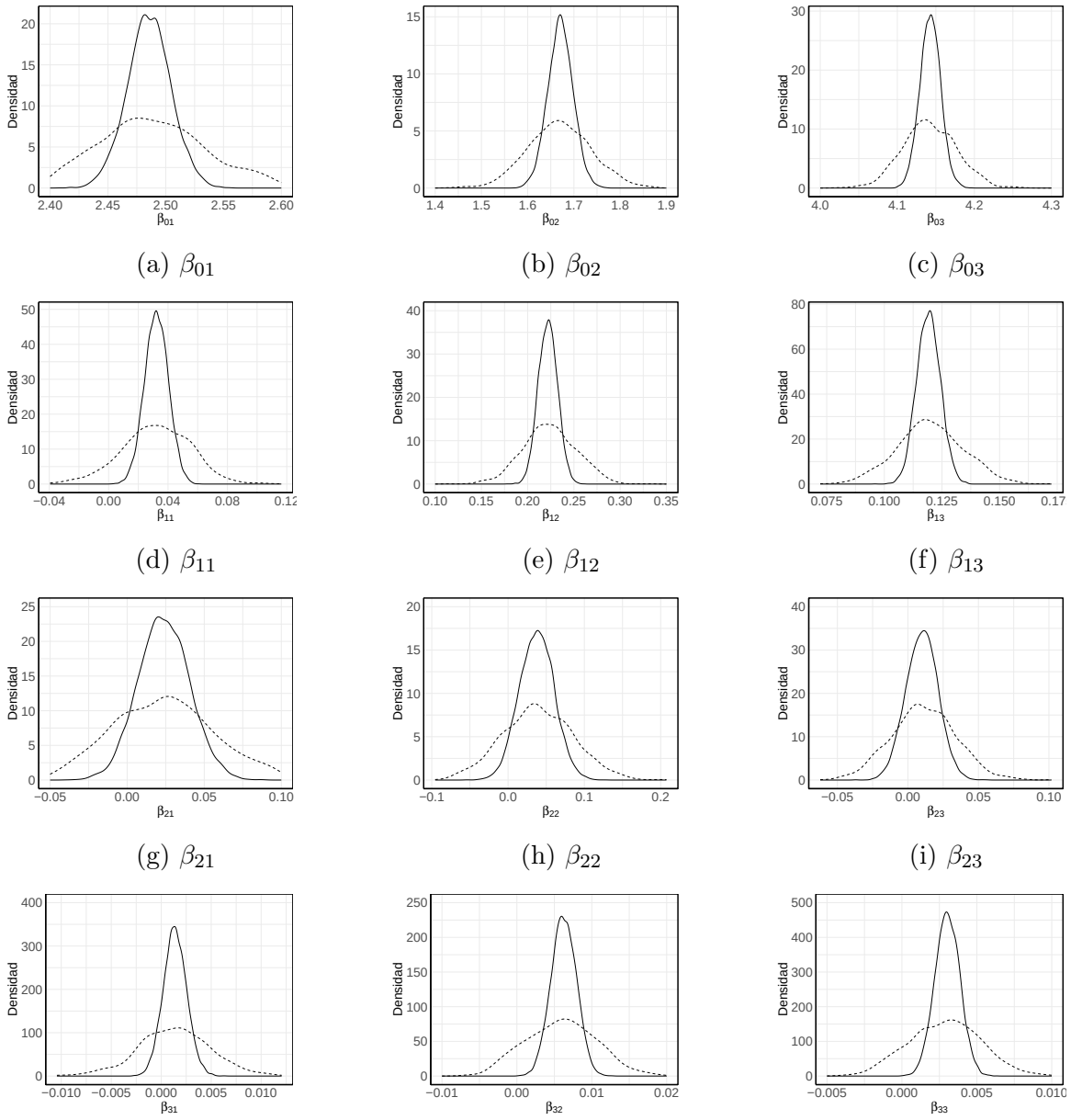
La Tabla (**5-1**) muestra la mediana y la Desviación Absoluta Mediana (MAD) correspondientes a los valores estimados de los parámetros. Es notable el desempeño satisfactorio de los estimadores, ya que las medianas muestran una aproximación cercana a los valores reales o verdaderos de los parámetros, y el MAD disminuye a medida que se incrementa el tamaño de la muestra. Este patrón refleja una mejora en la precisión de las estimaciones a medida que se aumenta la cantidad de datos en consideración.

La Tabla (**5-2**) muestra la mediana y la Desviación Absoluta Mediana (MAD) correspondientes a los cuartiles estimados. Es notable el desempeño satisfactorio de las estimaciones, ya que las medianas muestran una aproximación cercana a los cuartiles reales o verdaderos, y el MAD disminuye a medida que se incrementa el tamaño de la muestra. Este patrón refleja una mejora en la precisión de las estimaciones de los cuartiles a medida que se aumenta la cantidad de datos en consideración.

5.5.2. Resultados de los 3 Escenarios de Simulación Modelo Slash Multivariado

La Figura (**5-4**) ilustra la comparación entre las aproximaciones de las distribuciones posteriores para la muestra de tamaño 50 (línea punteada) y la distribución posterior aproximada con la base de datos real del modelo slash trivariado (línea continua). De manera similar, las Figuras (**5-5**) y (**5-6**) exhiben la comparación de las aproximaciones de las distribuciones posteriores para las muestras de tamaño 100 y 150, respectivamente, con la distribución posterior aproximada con la base de datos real. Se puede evidenciar que, a medida que aumenta el tamaño de la muestra, las aproximaciones de las distribuciones posteriores de los coeficientes del modelo convergen gradualmente hacia la distribución posterior aproximada

con la base de datos real. Este comportamiento sugiere una mejora en la precisión de las aproximaciones a medida que se incrementa el tamaño de la muestra.



(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

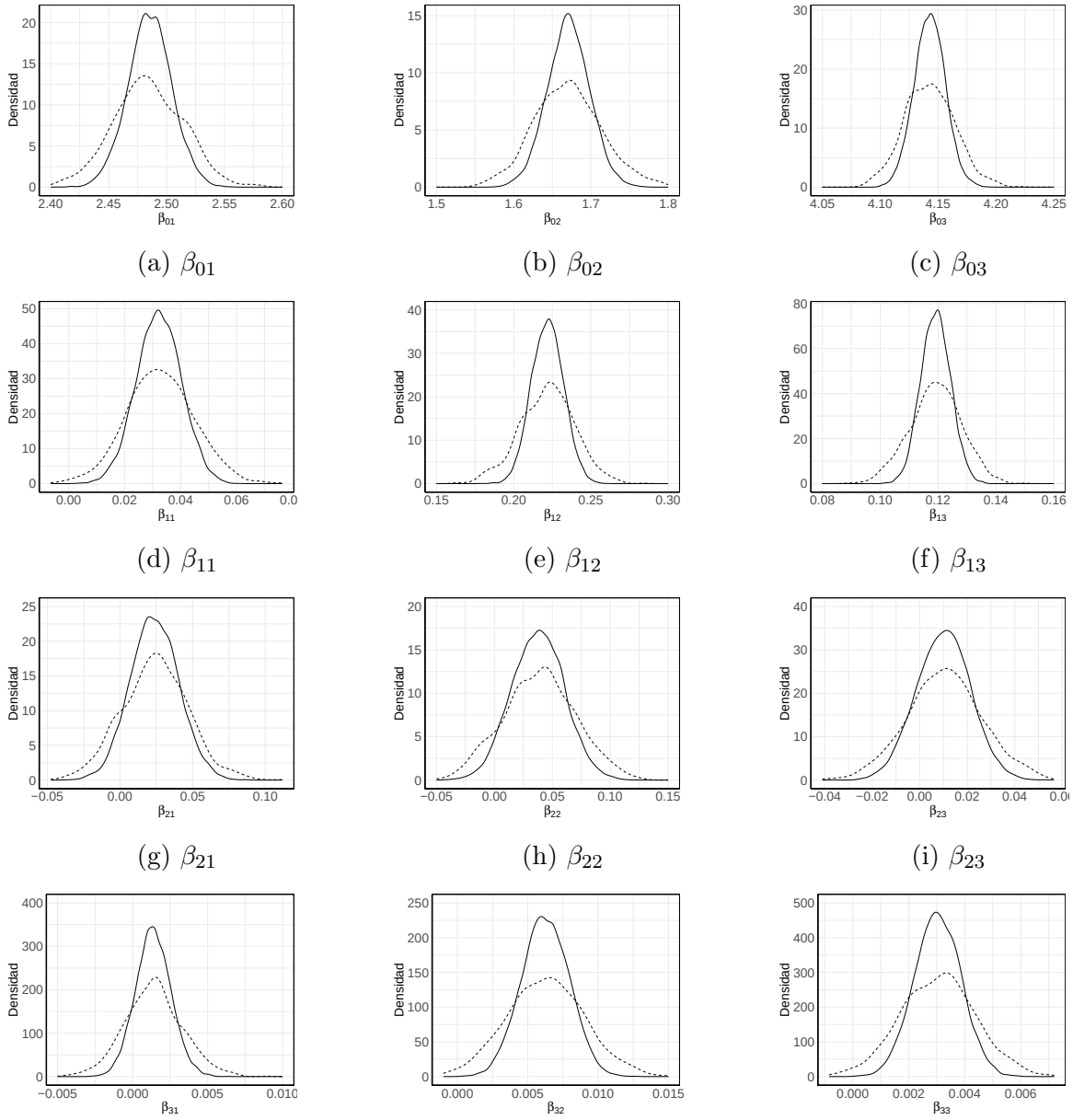
(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 5-1: Comparación de las Distribuciones Posteriores Aproximadas (línea punteada) de los Coeficientes del Modelo t Multivariado (Tamaño de Muestra 50) con las Distribuciones Posteriores Aproximadas de los Coeficientes con la Base de Datos Real (línea continua).

La Tabla (**5-3**) muestra la mediana y la Desviación Absoluta Mediana (MAD) correspon-

dientes a los valores estimados de los parámetros. Es notable el desempeño satisfactorio de los estimadores, ya que las medianas muestran una aproximación cercana a los valores reales o verdaderos de los parámetros, y el MAD disminuye a medida que se incrementa el tamaño de la muestra.



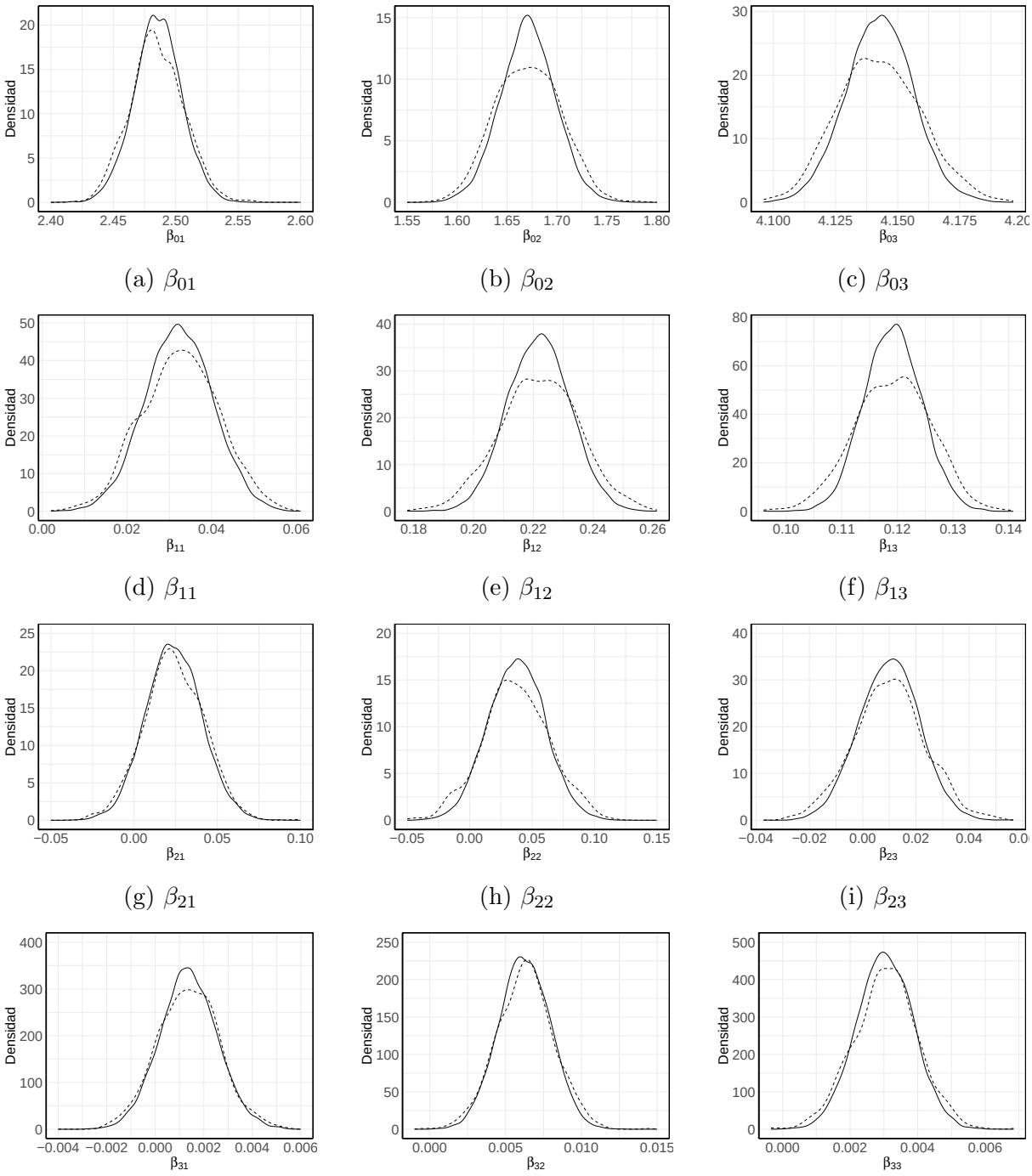
(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 5-2: Comparación de las Distribuciones Posteriores (línea punteada) de los Coeficientes del Modelo t Multivariado (Tamaño de Muestra 100) con las Distribuciones Posteriores Aproximadas de los Coeficientes con la Base de Datos Real (línea continua).



(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 5-3: Comparación de las Distribuciones Posteriores (línea punteada) de los Coeficientes del Modelo t Multivariado (Tamaño de Muestra 150) con las Distribuciones Posteriores Aproximadas de los Coeficientes con la Base de Datos Real (línea continua).

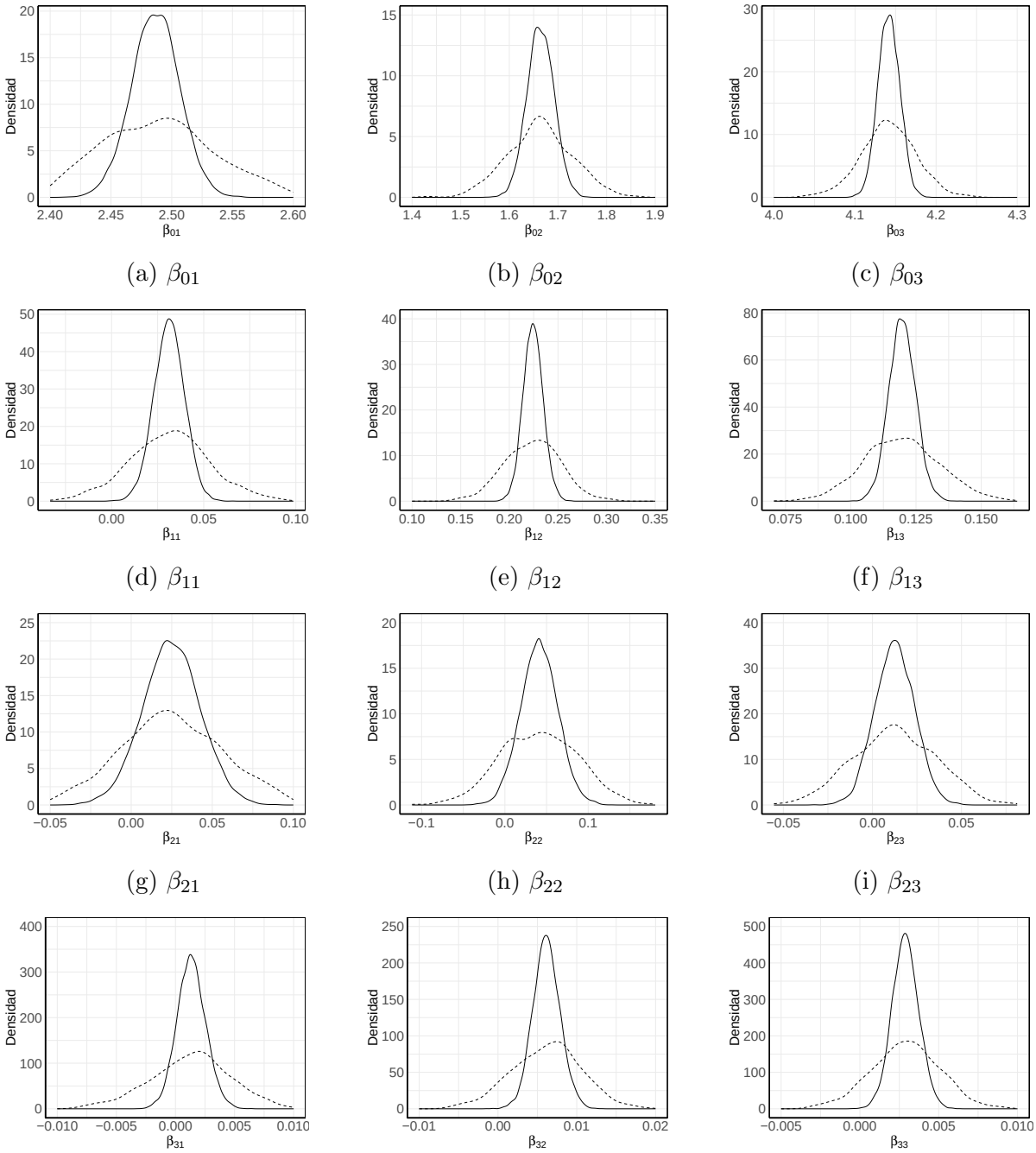
t		$n = 50$		$n = 100$		$n = 150$	
Párametro	**real**	**Mediana**	**MAD**	**Mediana**	**MAD**	**Mediana**	**MAD**
β_{01}	2.4855	2.4839	0.0342	2.4832	0.0206	2.4838	0.0138
β_{11}	0.0323	0.0324	0.0157	0.0325	0.0080	0.0327	0.0061
β_{21}	0.0236	0.0234	0.0239	0.0245	0.0150	0.0233	0.0118
β_{31}	0.0014	0.0014	0.0024	0.0014	0.0012	0.0014	0.0009
β_{02}	1.6709	1.6689	0.0461	1.6698	0.0290	1.6707	0.0230
β_{12}	0.2219	0.2229	0.0194	0.2224	0.0118	0.2223	0.0090
β_{22}	0.0377	0.0384	0.0318	0.0396	0.0209	0.0365	0.0175
β_{32}	0.0062	0.0060	0.0033	0.0063	0.0019	0.0063	0.0012
β_{03}	4.1428	4.1410	0.0236	4.1425	0.0150	4.1423	0.0112
β_{13}	0.1192	0.1192	0.0096	0.1194	0.0058	0.1193	0.0047
β_{23}	0.0101	0.0105	0.0152	0.0105	0.0103	0.0101	0.0085
β_{33}	0.0030	0.0029	0.0016	0.0031	0.0009	0.0031	0.0006
ψ_{11}	0.0066	0.0100	0.0022	0.0075	0.0014	0.0070	0.0011
ψ_{22}	0.0169	0.0243	0.0043	0.0204	0.0031	0.0199	0.0024
ψ_{33}	0.0042	0.0061	0.0011	0.0052	0.0008	0.0049	0.0006
ψ_{12}	0.0027	0.0033	0.0023	0.0029	0.0014	0.0028	0.0010
ψ_{13}	0.0014	0.0017	0.0011	0.0015	0.0007	0.0015	0.0005
ψ_{23}	0.0077	0.0105	0.002	0.0089	0.0014	0.0086	0.0011
ν	5.6875	10.3844	1.3688	6.3750	1.4750	5.5125	0.8844

Tabla 5-1: Mediana y MAD de las estimaciones de los parámetros; modelo de regresión lineal t multivariado.

La Tabla (**5-4**) muestra la mediana y la Desviación Absoluta Mediana (MAD) correspondientes a los cuartiles estimados. Es notable el desempeño satisfactorio de las estimaciones, ya que las medianas muestran una aproximación cercana a los cuartiles reales o verdaderos, y el MAD disminuye a medida que se incrementa el tamaño de la muestra. Este patrón refleja una mejora en la precisión de las estimaciones de los cuartiles a medida que se aumenta la cantidad de datos en consideración.

log-t		$n = 50$		$n = 100$		$n = 150$	
Cuantil real		**Mediana**	**MAD**	**Mediana**	**MAD**	**Mediana**	**MAD**
$z_{1,1/4,0}$	12.098	11.882	0.240	12.008	0.156	12.034	0.120
$z_{1,1/2,0}$	12.828	12.724	0.231	12.781	0.153	12.790	0.110
$z_{1,3/4,0}$	13.602	13.641	0.270	13.603	0.181	13.581	0.137
$z_{2,1/4,0}$	7.410	6.945	0.182	7.172	0.134	7.182	0.104
$z_{2,1/2,0}$	8.138	7.747	0.189	7.947	0.131	7.944	0.112
$z_{2,3/4,0}$	8.937	8.651	0.234	8.803	0.160	8.791	0.136
$z_{3,1/4,0}$	75.310	72.805	0.915	74.056	0.670	74.048	0.569
$z_{3,1/2,0}$	78.928	76.941	0.906	77.961	0.660	77.905	0.543
$z_{3,3/4,0}$	82.719	81.235	1.060	82.108	0.739	81.955	0.629
$z_{1,1/4,1}$	12.387	12.158	0.202	12.300	0.135	12.322	0.103
$z_{1,1/2,1}$	13.134	13.030	0.175	13.095	0.125	13.091	0.093
$z_{1,3/4,1}$	13.926	13.953	0.229	13.932	0.150	13.908	0.118
$z_{2,1/4,1}$	7.695	7.233	0.157	7.455	0.108	7.451	0.097
$z_{2,1/2,1}$	8.451	8.055	0.150	8.262	0.115	8.253	0.093
$z_{2,3/4,1}$	9.281	8.978	0.191	9.163	0.138	9.136	0.114
$z_{3,1/4,1}$	76.077	73.632	0.748	74.829	0.530	74.807	0.484
$z_{3,1/2,1}$	79.731	77.795	0.768	78.792	0.499	78.721	0.434
$z_{3,3/4,1}$	83.561	82.199	0.903	82.952	0.609	82.815	0.512

Tabla 5-2: Mediana y MAD de los cuartiles estimados; modelo de regresión lineal log-t multivariado.



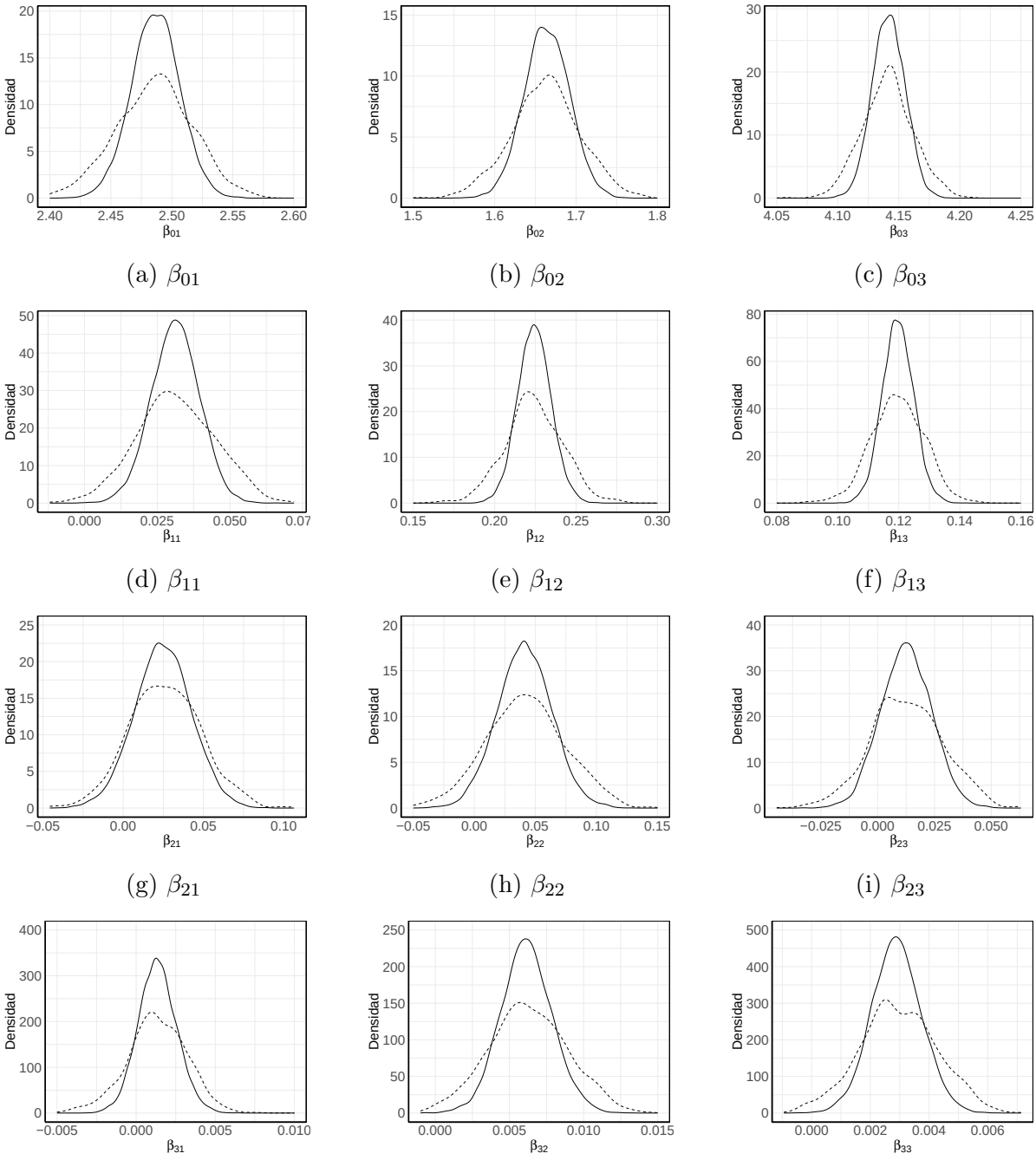
(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 5-4: Comparación de las Distribuciones Posteriores (línea punteada) de los Coeficientes del Modelo Slash Multivariado (Tamaño de Muestra 50) con las Distribuciones Posteriores Aproximadas de los Coeficientes con la Base de Datos Real (línea continua).



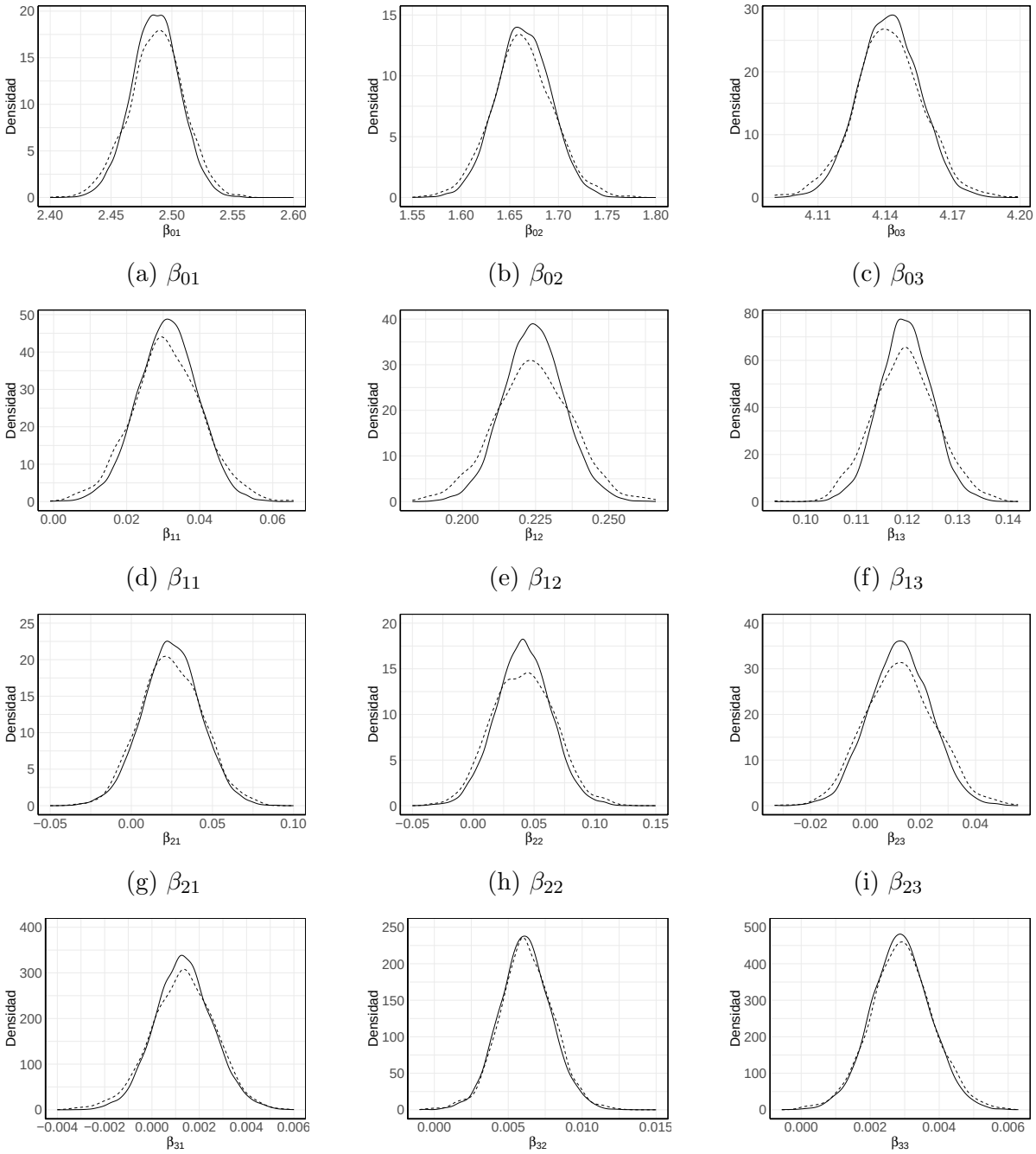
(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 5-5: Comparación de las Distribuciones Posteriores (línea punteada) de los Coeficientes del Modelo Slash Multivariado (Tamaño de Muestra 100) con las Distribuciones Posteriores Aproximadas de los Coeficientes con la Base de Datos Real (línea continua).



(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 5-6: Comparación de las Distribuciones Posteriores (línea punteada) de los Coeficientes del Modelo Slash Multivariado (Tamaño de Muestra 150) con las Distribuciones Posteriores Aproximadas de los Coeficientes con la Base de Datos Real (línea continua).

slash		$n = 50$		$n = 100$		$n = 150$	
Párametro	**real**	**Mediana**	**MAD**	**Mediana**	**MAD**	**Mediana**	**MAD**
β_{01}	2.4867	2.4874	0.0339	2.4868	0.0207	2.4880	0.0144
β_{11}	0.0313	0.0322	0.0140	0.0309	0.0089	0.0308	0.0062
β_{21}	0.0247	0.0243	0.0224	0.0258	0.0153	0.0240	0.0132
β_{31}	0.0013	0.0016	0.0022	0.0013	0.0012	0.0013	0.0009
β_{02}	1.6639	1.6618	0.0439	1.6634	0.0263	1.6623	0.0196
β_{12}	0.2243	0.2249	0.0195	0.2235	0.0113	0.2240	0.0087
β_{22}	0.0411	0.0402	0.0331	0.0417	0.0214	0.0416	0.0177
β_{32}	0.0061	0.0064	0.0029	0.0062	0.0017	0.0062	0.0012
β_{03}	4.1417	4.1411	0.0219	4.1410	0.0135	4.1414	0.0097
β_{13}	0.1197	0.1192	0.0097	0.1197	0.0057	0.1195	0.0043
β_{23}	0.0125	0.0124	0.0164	0.0123	0.0107	0.0127	0.0087
β_{33}	0.0029	0.0029	0.0014	0.0029	0.0009	0.0029	0.0006
ψ_{11}	0.0050	0.0066	0.0019	0.0057	0.0012	0.0052	0.0008
ψ_{22}	0.0128	0.0166	0.0036	0.0152	0.0022	0.0146	0.0018
ψ_{33}	0.0032	0.0042	0.0009	0.0038	0.0006	0.0036	0.0005
ψ_{12}	0.0019	0.0022	0.0014	0.0021	0.0011	0.0021	0.0008
ψ_{13}	0.0010	0.0010	0.0008	0.0011	0.0050	0.0011	0.0004
ψ_{23}	0.0058	0.0072	0.0017	0.0066	0.0010	0.0063	0.0004
ν	2.2912	2.6129	0.4351	2.4672	0.3356	2.3307	0.2392

Tabla 5-3: Mediana y MAD de las estimaciones de los parámetros; modelo de regresión lineal slash multivariado.

log-slash		$n = 50$		$n = 100$		$n = 150$	
Cuantil real		**Mediana**	**MAD**	**Mediana**	**MAD**	**Mediana**	**MAD**
$z_{1,1/4,0}$	12.090	11.900	0.238	12.002	0.156	12.040	0.130
$z_{1,1/2,0}$	12.816	12.720	0.221	12.759	0.142	12.775	0.129
$z_{1,3/4,0}$	13.587	13.580	0.257	13.556	0.167	13.556	0.143
$z_{2,1/4,0}$	7.383	6.962	0.180	7.158	0.137	7.170	0.104
$z_{2,1/2,0}$	8.105	7.723	0.186	7.911	0.137	7.911	0.109
$z_{2,3/4,0}$	8.898	8.573	0.221	8.755	0.158	8.735	0.133
$z_{3,1/4,0}$	75.206	72.989	0.934	73.992	0.683	74.037	0.548
$z_{3,1/2,0}$	78.793	76.883	0.949	77.829	0.678	77.794	0.548
$z_{3,3/4,0}$	82.551	81.052	1.039	81.853	0.755	81.761	0.581
$z_{1,1/4,1}$	12.393	12.223	0.216	12.319	0.137	12.337	0.107
$z_{1,1/2,1}$	13.137	13.038	0.192	13.096	0.122	13.096	0.106
$z_{1,3/4,1}$	13.927	13.918	0.227	13.916	0.145	13.901	0.131
$z_{2,1/4,1}$	7.693	7.253	0.153	7.461	0.112	7.464	0.084
$z_{2,1/2,1}$	8.445	8.050	0.158	8.248	0.109	8.241	0.088
$z_{2,3/4,1}$	9.271	8.932	0.185	9.122	0.141	9.105	0.112
$z_{3,1/4,1}$	76.151	73.837	0.808	74.985	0.551	74.952	0.423
$z_{3,1/2,1}$	79.783	77.819	0.742	78.828	0.543	78.776	0.412
$z_{3,3/4,1}$	83.588	81.974	0.867	82.899	0.653	82.766	0.499

Tabla 5-4: Mediana y MAD de los cuartiles estimados; modelo de regresión lineal log-slash multivariado.

En el siguiente capítulo se ilustran las diferentes metodologías expuestas hasta el momento en este trabajo, considerando una base de datos real de niños menores de 5 años en la comuna de Robledo situada en la ciudad de Medellín.

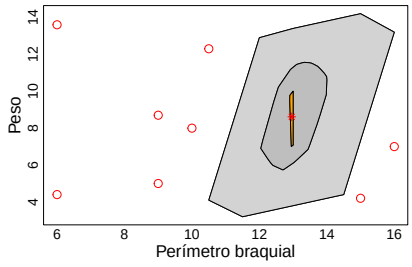
6 Aplicación

Las tablas de crecimiento antropométricas son un componente esencial del instrumental pediátrico. Su valor reside en ayudar a determinar el grado en que se satisfacen las necesidades fisiológicas de crecimiento y desarrollo durante el importante periodo de la infancia (WHO, 2007). Es decir, posibles desviaciones del patrón que se describe en las tablas de crecimiento son prueba de un crecimiento anormal. La Organización Mundial de la Salud (WHO, 2006, 2007) proporciona una variedad de curvas de cuantiles de referencias para describir varias características antropométricas de los niños en función de la edad y el género, tales como, peso, altura, perímetro braquial, entre otras. La construcción de estás curvas fueron basadas en una muestra de 8440 lactantes y niños pequeños (0 a 5 años) sanos alimentados con leche materna de diversos orígenes étnicos y entornos culturales (Brasil, Ghana, India, Noruega, Omán y EE.UU.) (WHO, 2007). En particular, el peso en función de la edad es uno de las tablas de crecimiento más empleadas para controlar los cambios en la salud o el estado nutricional de los niños (Morán-Vásquez, Giraldo-Melo, y Mazo-Lopera, 2023). Sin embargo, las estimaciones (Morán-Vásquez, Roldán-Correa, y Nagar, 2023) de las curvas de cuantiles de referencias construidas por la Organización Mundial de la Salud se obtienen ajustando modelos univariados para cada medida antropométrica por separado, ignorando la asociación entre ellas. Varios autores han aplicado el modelado de cuantiles en el ámbito multivariado (Morán-Vásquez y cols., 2021; Morán-Vásquez, Roldán-Correa, y Nagar, 2023) para la construcción de tablas de crecimiento antropométricas de niños de 0 a 5 años, teniendo en cuenta la asociación entre las parejas de variables.

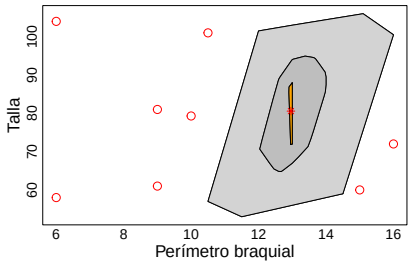
Nosotros usamos la clase de modelos de regresión lineal normal/independiente multivariados bajo la presencia de datos faltantes para llevar a cabo el proceso de estimación de los cuantiles del peso (en kilogramos), la altura (en centímetros), y el perímetro braquial (en centímetros) de niños (0 a 4 años, ya que de 4 a 5 no aparecían en la base de datos finalmente seleccionada), teniendo en cuenta la asociación natural entre las variables, en función del sexo, la edad y el tiempo de lactancia materna. Utilizamos la base de datos de registro de pacientes atendidos en las Instituciones Prestadoras de Servicios de Salud con diagnóstico confirmado de Desnutrición Aguda en menores de 5 años en la ciudad de Medellín, Colombia (se puede encontrar en el sitio web `http://tinyurl.com/23ff85zx`). El objetivo de analizar dicha base de datos es ilustrar cómo nuestra metodología permite caracterizar una población particular a través de la elaboración de cuantiles en función de un conjunto de covariables y teniendo en cuenta las posibles relaciones entre las variables analizadas. Todo lo anterior, teniendo en

cuenta la existencia de datos faltantes en las variables dependientes peso, altura y perímetro braquial. Por practicidad en la aplicación, decidimos seleccionar solo una comuna de la ciudad de Medellín, llamada Robledo y obtuvimos finalmente una muestra de 173 niños (106 hombres y 67 mujeres), tomada entre los años 2016 y 2021. Las variables respuesta son peso (8 datos faltantes), talla (6 datos faltantes) y perímetro braquial (63 datos faltantes), y las variables explicativas (totalmente observadas) son edad (edad; en años), sexo (sexo; 1 = hombre, 0 = mujer), y tiempo de leche materna o lactancia (tiempo lechem; en semanas).

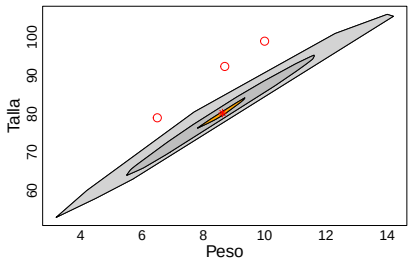
La Figura (**6-1**) presenta los bagplots, una herramienta de visualización útil para examinar la relación conjunta entre pares de variables. En este contexto, los bagplots representan la relación entre las tres variables respuesta: perímetro braquial, peso y talla. Se evidencia una asociación positiva entre estas variables, con la presencia de algunos valores atípicos. Un análisis detenido de los gráficos revela que el par de variables que exhibe la mayor asociación es aquel conformado por la talla y el peso. La identificación de esta fuerte asociación es esencial, ya que la correlación entre las variables respuesta desempeña un papel crucial en la obtención de estimaciones más precisas de los coeficientes del modelo.



(a) Peso vs Perímetro braquial



(b) Talla vs Perímetro braquial



(c) Talla vs Peso

Figura 6-1: Bagplots de las variables respuesta

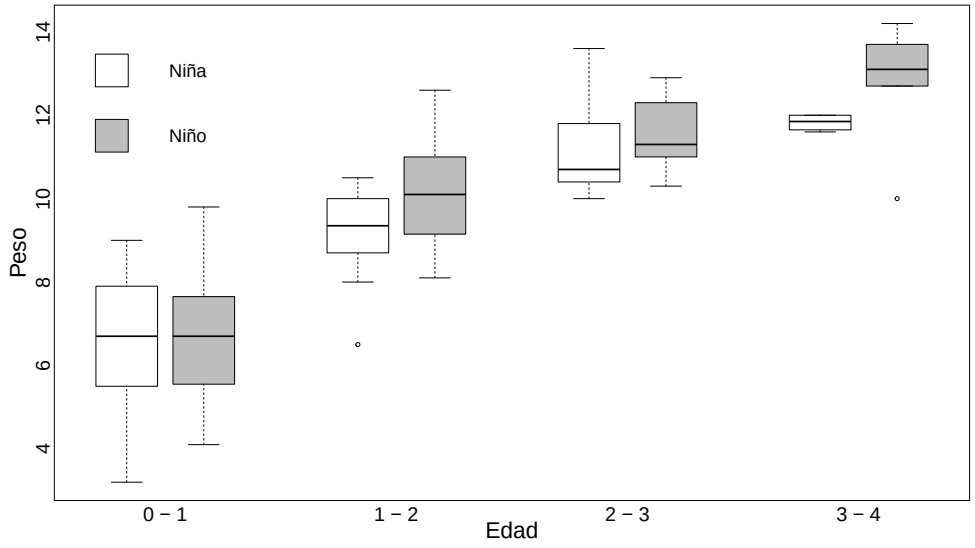


Figura 6-2: Gráficos de caja comparativos del peso de los niños por edad y sexo; datos de los niños

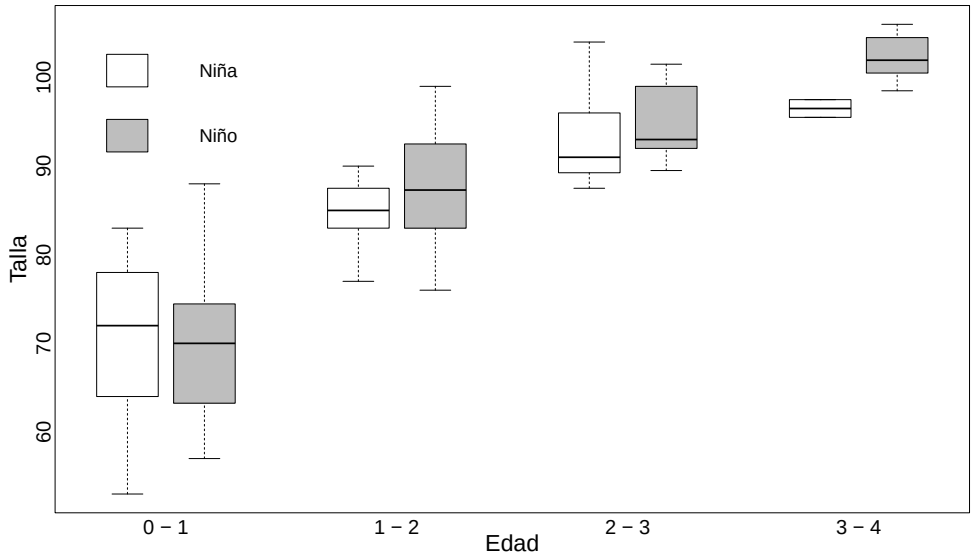


Figura 6-3: Gráficos de caja comparativos de la talla de los niños por edad y sexo; datos de los niños

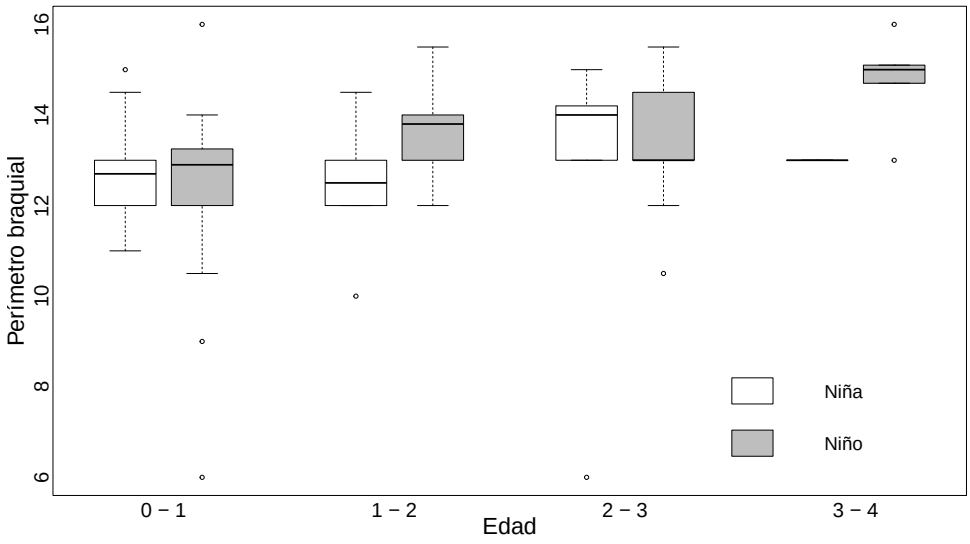
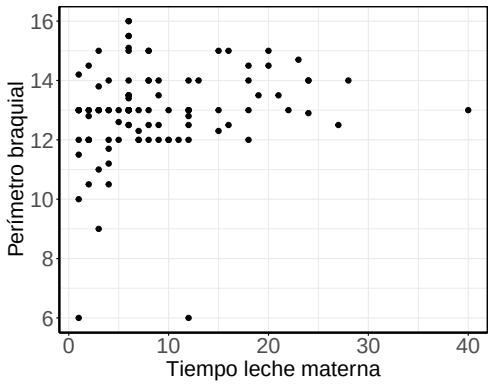
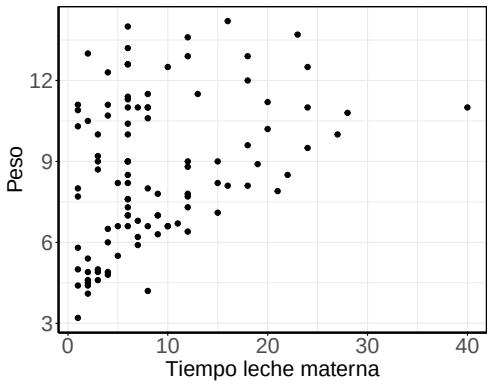


Figura 6-4: Gráficos de caja comparativos del perímetro braquial de los niños por edad y sexo; datos de los niños

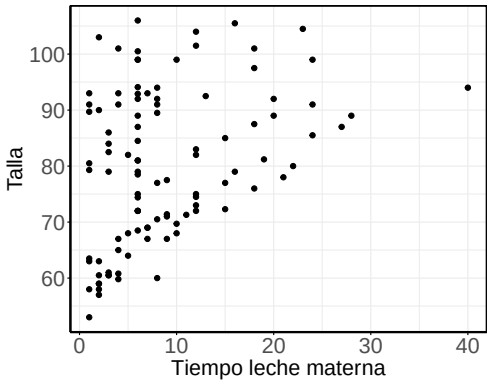
Las Figuras (**6-2**), (**6-3**), y (**6-4**) muestran los gráficos de cajas comparativos del peso, talla y perímetro braquial de los niños, discriminados por sexo y divididos en 4 intervalos de edad. Se observa que los cuantiles empíricos del peso, talla y perímetro braquial se ven influenciados tanto por la edad como por el sexo de manera significativa. En cada intervalo de edad, tanto para niños como para niñas, se percibe una ligera asimetría positiva y la presencia de valores atípicos en los pesos. Así mismo, en cada intervalo de edad, se identifica una ligera asimetría negativa y valores atípicos en los perímetros braquiales de los niños. Estos hallazgos indican que tanto la edad como el sexo son factores determinantes que influyen en las medidas antropométricas, y la presencia de asimetrías y valores atípicos destaca la variabilidad dentro de cada subgrupo, subrayando la importancia de considerar estas diferencias al realizar análisis comparativos.



(a) Perímetro braquial vs Tiempo lm



(b) peso vs Tiempo leche materna



(c) Talla vs Tiempo leche materna

Figura 6-5: Gráficos de dispersión de las variables respuesta vs Tiempo de leche materna

La Figura (**6-5**) presenta gráficos de dispersión que ilustran la relación entre las variables respuesta y la variable explicativa *Tiempo de leche materna*. En la observación de estos gráficos, se evidencia una tendencia lineal y positiva, sugiriendo que un aumento en el tiempo de leche materna está asociado con cambios positivos en cada una de las variables respuesta. Este patrón lineal positivo indica una posible relación beneficiosa entre la duración de la lactancia materna y los valores de las variables respuesta, brindando indicios valiosos sobre la influencia positiva de la variable *Tiempo de leche materna* en el contexto del análisis.

Para investigar estas relaciones, procedimos a ajustar los modelos de regresión lineal normal/independiente multivariados, $\boldsymbol{Y}_i \overset{\text{ind}}{\sim} \text{NI}_3(\text{Log}(\boldsymbol{\mu}_i), \boldsymbol{\Psi}, H)$, $\boldsymbol{Y}_i = \text{Log}(\boldsymbol{Z}_i)$, donde

$$\text{Log}(\mu_{i1}) = \beta_{01} + \beta_{11}\text{Edad}_i + \beta_{21}\text{Sexo}_i + \beta_{31}\text{Tiempo leche materna}_i,$$

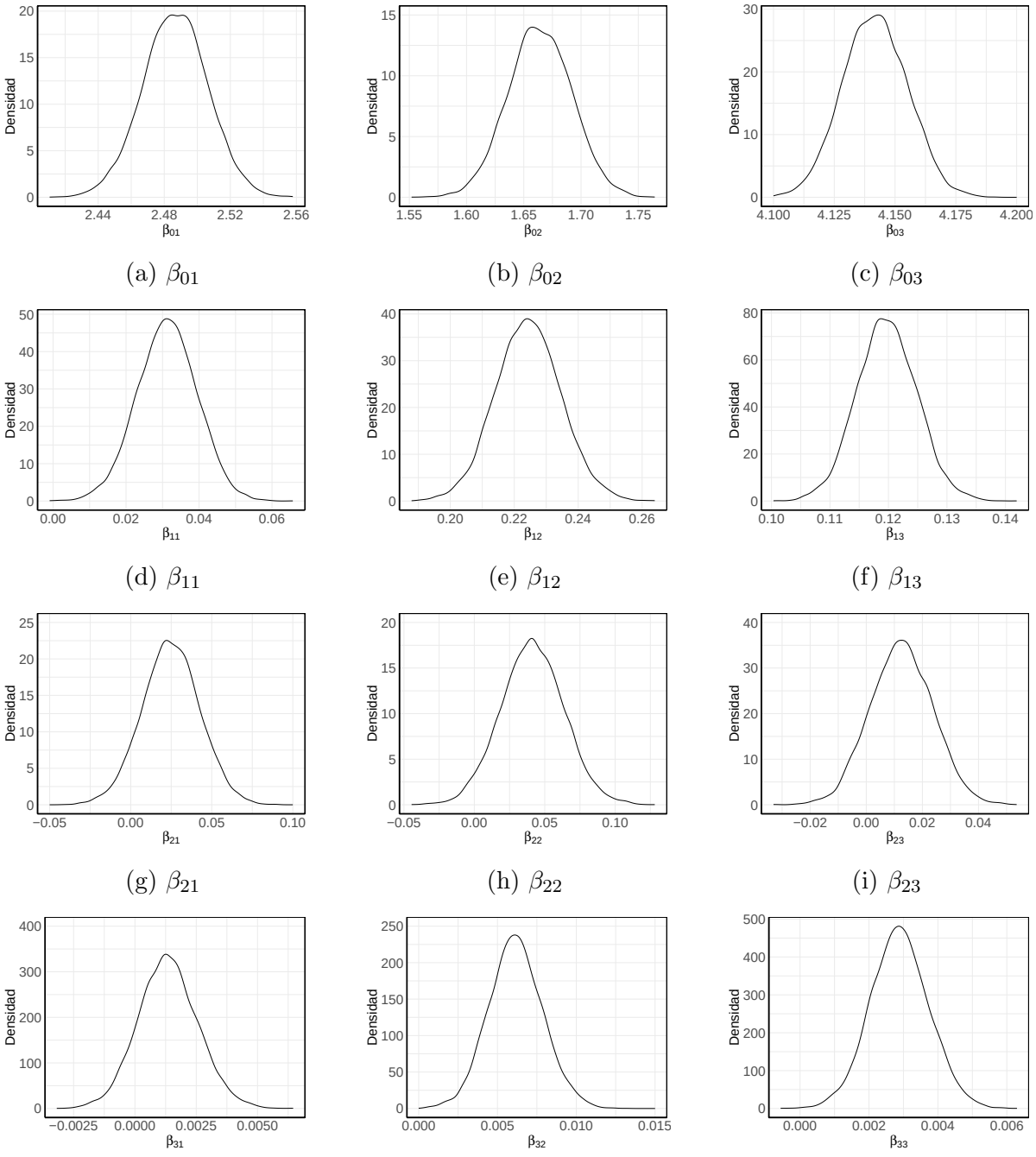
$$\text{Log}(\mu_{i2}) = \beta_{02} + \beta_{12}\text{Edad}_i + \beta_{22}\text{Sexo}_i + \beta_{32}\text{Tiempo leche materna}_i,$$

$$\text{Log}(\mu_{i3}) = \beta_{03} + \beta_{13}\text{Edad}_i + \beta_{23}\text{Sexo}_i + \beta_{33}\text{Tiempo leche materna}_i,$$

para $i = 1, \ldots, 173$. En este contexto, empleamos las distribuciones t trivariada y slash trivariada como distribuciones asociadas al vector de respuesta, con la respectivas distribuciones a priori (Liu, 1996) para los parámetros del modelo t trivariado dada por $P(\boldsymbol{\beta}, \boldsymbol{\Psi}, \nu) \propto |\boldsymbol{\Psi}|^{-(p+1)/2} / \nu^2$ $(\nu > 1)$, y para el modelo slash trivariado dada por $P(\boldsymbol{\beta}, \boldsymbol{\Psi}, \nu) \propto |\boldsymbol{\Psi}|^{-(p+1)/2} \Gamma(\mathrm{a} = 6, \mathrm{b} = 2)$.

Las Figuras (**6-6**) y (**6-7**) muestran las distribuciones posteriores de los coeficientes tanto del modelo slash como del modelo t. Estas distribuciones se derivan de una muestra de 10000 tomadas de sus respectivas distribuciones posteriores, con un período de "Burn-In" de 1000 para garantizar la estabilidad de las estimaciones y eliminar (Schafer, 1997) el efecto de los valores iniciales. Estos gráficos proporcionan una representación visual detallada de la incertidumbre asociada con los coeficientes del modelo, ofreciendo información valiosa sobre la variabilidad esperada. El análisis de estas distribuciones posteriores contribuye significativamente a la comprensión de la robustez y la confiabilidad de los resultados del modelo, permitiendo una evaluación más informada de la incertidumbre inherente en las estimaciones de los coeficientes.

Además, la Figura (**6-8**) exhibe las distribuciones posteriores del parámetro de cola del modelo slash trivariado y los grados de libertad del modelo t trivariado. La interpretación de estos resultados es fundamental para comprender la forma y la flexibilidad inherentes a los modelos empleados. Esta información permite una evaluación más profunda de la calidad de ajuste y la robustez de las inferencias realizadas.



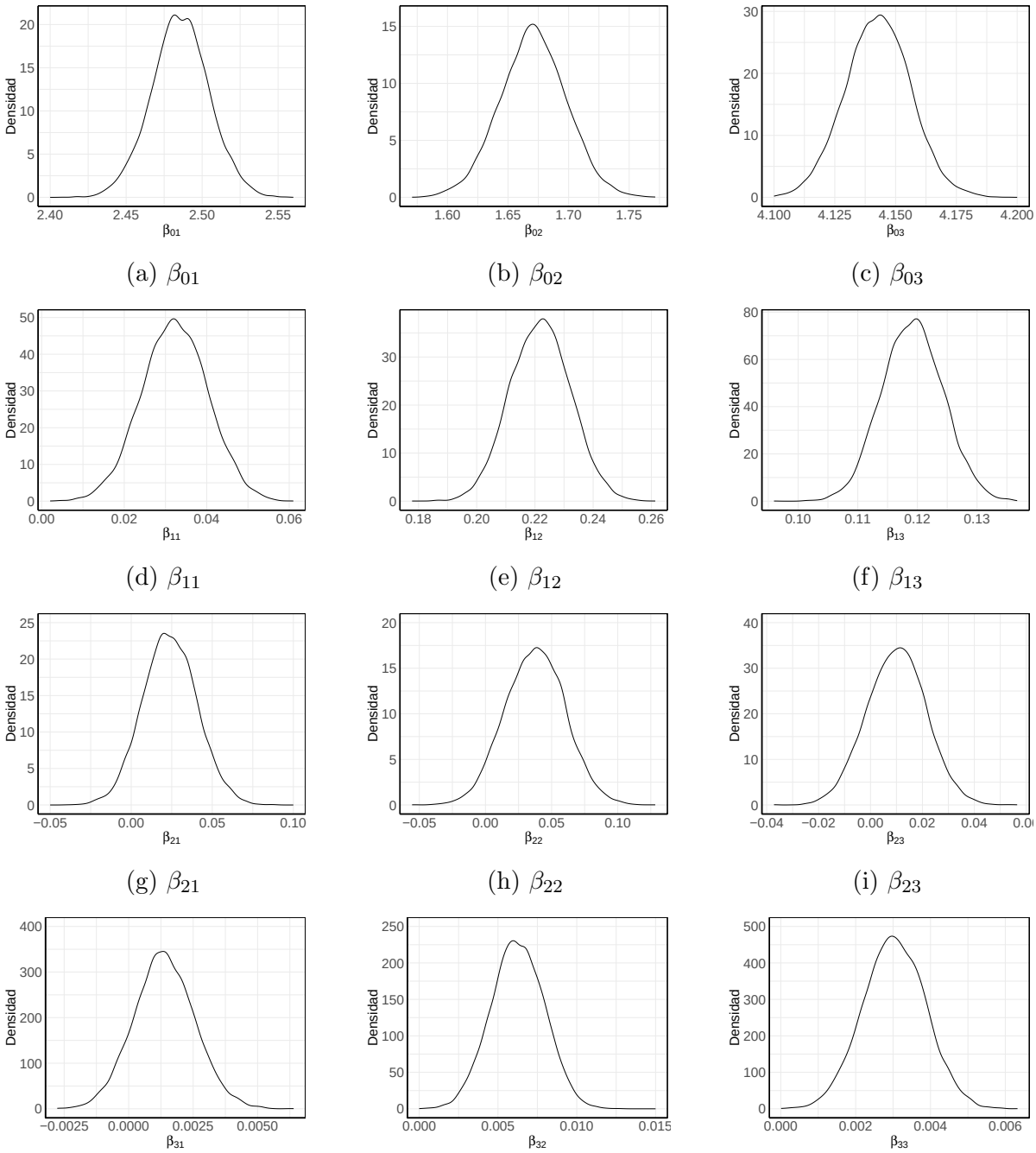
(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 6-6: Distribuciones posteriores de los coeficientes del modelo slash multivariado



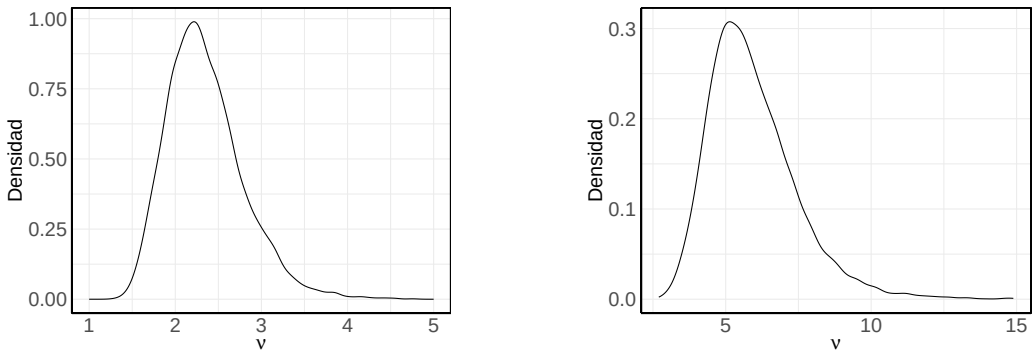
(a) β_{01} (b) β_{02} (c) β_{03}

(d) β_{11} (e) β_{12} (f) β_{13}

(g) β_{21} (h) β_{22} (i) β_{23}

(j) β_{31} (k) β_{32} (l) β_{33}

Figura 6-7: Distribuciones posteriores de los coeficientes del modelo t multivariado



(a) Distribución posterior del parámetro ν del modelo slash multivarido

(b) Distribución posterior del parámetro ν del modelo t multivariado

Figura 6-8: Distribuciones posteriores de los hiperparámetros del modelo slash multivariado y t multivariado basadas en una muestra de 10000 tomadas desde su distribución posterior con periodo de "Burn-in" de 1000.

La Tabla (**6-1**) presenta las estimaciones puntuales (mediana posterior) de los coeficientes del modelo slash, acompañadas de sus correspondientes errores estándar (respecto a la mediana posterior) y los intervalos de probabilidad del 95 %. Las respectivas estimativas de los parámetros de dispersión, $\psi_{1,1}$, $\psi_{2,2}$ y $\psi_{3,3}$, fueron 0.005, 0.013, y 0.003. Así mismo, las estimativas de los parámetros de asociación entre el par de variables, $\psi_{1,2}$, $\psi_{1,3}$ y $\psi_{2,3}$, fueron 0.002, 0.001, y 0.006, revelando una asociación positiva entre el vector de variables respuesta. La mediana posterior del parámetro de cola ν fue igual a 2.291. Estos resultados proporcionan información detallada de las estimaciones de los coeficientes y la incertidumbre asociada, siendo fundamentales para la interpretación y comprensión del modelo.

Así mismo, la Tabla (**6-2**) presenta las estimaciones puntuales (mediana posterior) de los coeficientes del modelo t, acompañadas de sus correspondientes errores estándar y los intervalos de probabilidad del 95 %. Las respectivas estimativas de los parámetros de dispersión, $\psi_{1,1}$, $\psi_{2,2}$ y $\psi_{3,3}$, fueron 0.007, 0.017, y 0.004. Asimismo, las estimativas de los parámetros de asociación entre el par de variables, $\psi_{1,2}$, $\psi_{1,3}$ y $\psi_{2,3}$, fueron 0.003, 0.001, y 0.008, revelando de igual manera la asociación positiva entre el vector de variables respuesta. La mediana posterior de los grados de libertad ν fue igual a 5.688.

Evaluamos la bondad del ajuste de los modelos de regresión lineal normal/independiente multivariados (3-1), bajo la presencia de datos faltantes, mediante gráficos de cuantiles-cuantiles, comparando las distancias empíricas de Mahalanobis $\hat{\delta}^2_{i,\text{obs}} = \delta^2(\boldsymbol{y}_{i,\text{obs}}, \text{Log}(\hat{\boldsymbol{\mu}}_{i,\text{obs}}), \hat{\boldsymbol{\Psi}}_{i,\text{obs}})$, $i = 1, \ldots, n$, con los cuantiles teóricos $\delta^2_{\alpha_i}$, donde $\alpha_i = 1/(n+1)$, $i = 1, \ldots, n$, obtenidos desde (2-6) con $\boldsymbol{\nu} = \hat{\boldsymbol{\nu}}$. Además, construimos envolventes simuladas (Atkinson, 1981) para los gráficos de cuantiles-cuantiles con el fin de ayudar a la comparación entre cuantiles y juzgar la adecuación de los modelos.

	Per. braq.				Peso			
Var. Expl.	**Est.**	**SE**	**Inf.**	**Sup.**	**Est.**	**SE**	**Inf.**	**Sup.**
Intercepto	2.4867	0.0004	2.4473	2.5253	1.66389	0.00077	1.6097	1.7183
Edad	0.0313	0.0001	0.0150	0.0470	0.2243	0.00011	0.2041	0.2451
Sexo	0.0247	0.0003	-0.0103	0.0594	0.0411	0.00050	-0.0028	0.0855
tlm	0.0013	0.0000	-0.0010	0.0038	0.0061	0.00000	0.0028	0.0097

	Talla			
Var. Expl.	**Est.**	**SE**	**Inf.**	**Sup.**
Intercepto	4.1417	0.0002	4.1159	4.1677
Edad	0.1197	0.0000	0.1097	0.1300
Sexo	0.0125	0.0001	-0.0087	0.0342
tlm	0.0029	0.0000	0.0012	0.0046

Tabla 6-1: Estimativas del Modelo Slash multivariado

	Per. braq.				Peso			
Var. Expl.	**Est.**	**SE**	**Inf.**	**Sup.**	**Est.**	**SE**	**Inf.**	**Sup.**
Intercepto	2.4855	0.0004	2.4476	2.5230	1.6709	0.0008	1.6164	1.7247
Edad	0.0323	0.0001	0.0164	0.0477	0.2220	0.0001	0.2017	0.2423
Sexo	0.0236	0.0003	-0.0084	0.0573	0.0377	0.0005	-0.0061	0.0822
tlm	0.0014	0.0000	-0.0009	0.0037	0.0062	0.0000	0.0029	0.0095

	Talla			
Var. Expl.	**Est.**	**SE**	**Inf.**	**Sup.**
Intercepto	4.1429	0.0002	4.1164	4.1695
Edad	0.1192	0.0000	0.1094	0.1292
Sexo	0.0101	0.0001	-0.0123	0.0322
tlm	0.0030	0.0000	0.0014	0.0047

Tabla 6-2: Estimativas del Modelo t Multivariado

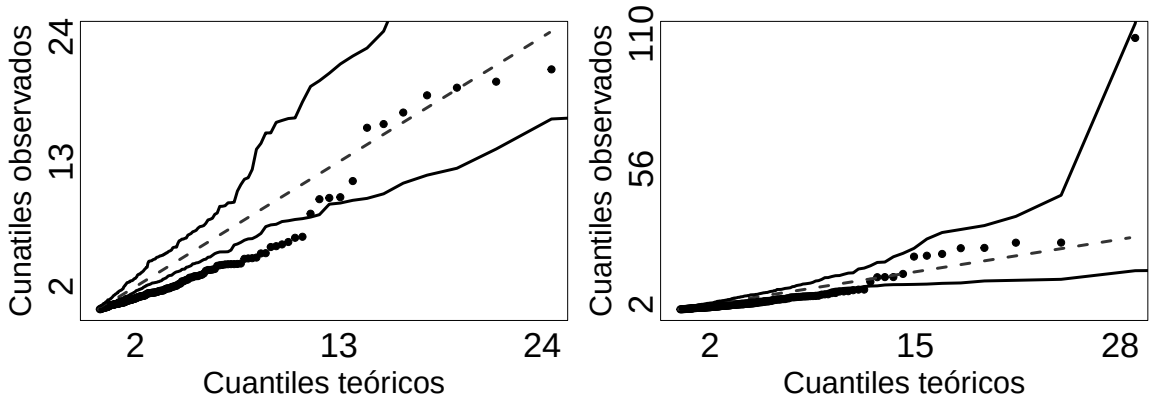
La Figura **6-9** presenta el gráfico de cuantiles con la envolvente simulada para las distancias de Mahalanobis de cada modelo. Al examinar el gráfico, se observa que el modelo que mejor se ajusta a los datos es el modelo slash. Esta representación gráfica proporciona una valiosa visualización de la capacidad de ajuste de cada modelo, destacando la superioridad del modelo slash.

Adicionalmente, para llevar a cabo la selección de un modelo, además de la información proporcionada por el gráfico de cuantiles, nos apoyamos en el factor de Bayes (BF), una

herramienta valiosa en el contexto bayesiano para la comparación de modelos. En nuestra hipótesis nula, supusimos que el modelo adecuado era el t, mientras que bajo la hipótesis alternativa consideramos que el modelo adecuado era el slash. Dado que los parámetros del modelo son conocidos, no es necesario asumir ninguna distribución a priori para ellos, y el BF se reduce a una razón de las probabilidades posteriores.

El valor del Log10(BF), conocido como el peso de la evidencia proporcionada por los datos (De Santis y Spezzaferri, 1999), fue igual a -49.95644. Dado que el Log10 (BF) es menor a -2, esto ofrece evidencia decisiva en contra de la hipótesis nula, favoreciendo al modelo slash como el que mejor se ajusta. Este análisis basado en el factor de Bayes refuerza la robustez de la elección del modelo y respalda la conclusión de que el modelo slash proporciona una descripción más adecuada de los datos.

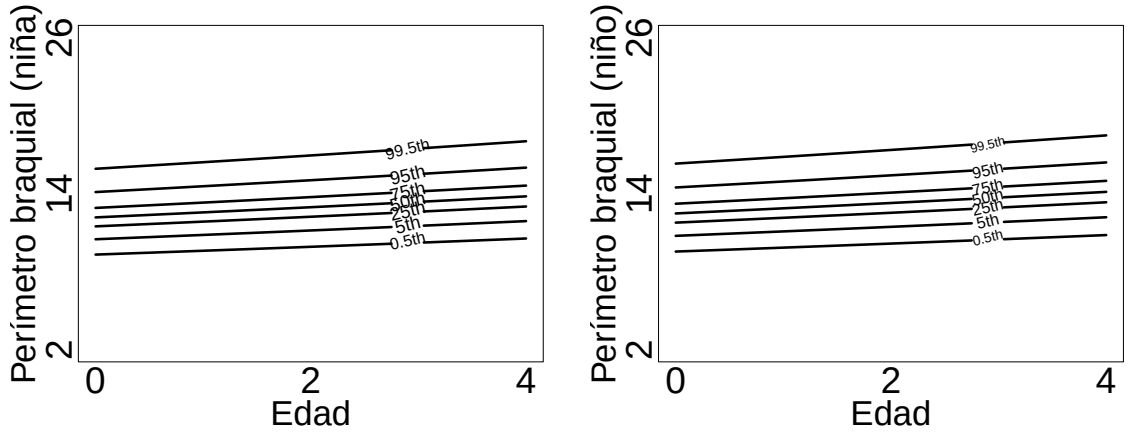
Basados en el modelo final seleccionado, se ajustaron las curvas de cuantiles. Estas son ilustradas en los gráficos (**6-10**), (**6-11**) y (**6-12**), para las variables de perímetro braquial, peso y talla de los niños en función de la edad y tiempo de leche materna, manteniendo constante la variable del tiempo de leche materna en su media.



(a) Envolvente simulada modelo t multivariado

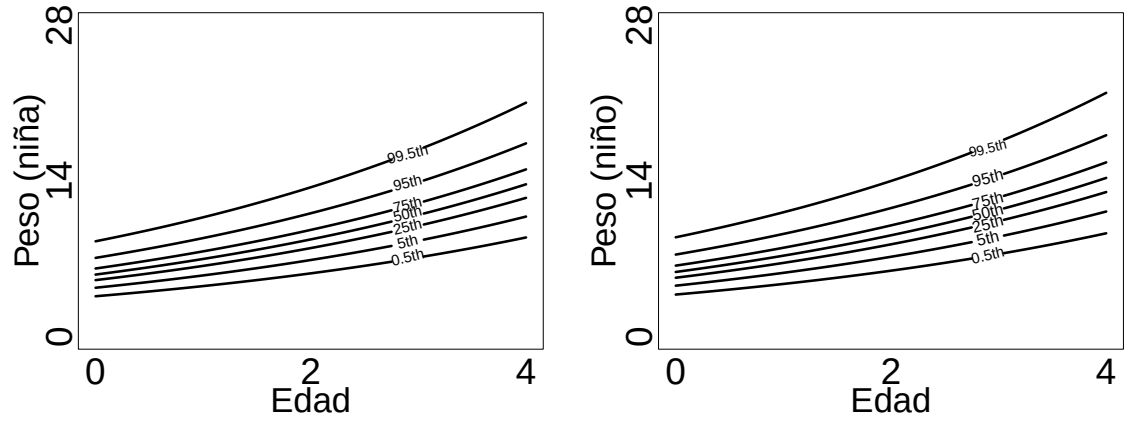
(b) Envolvente simulada modelo slash multivariado

Figura 6-9: Envolventes simuladas para los modelos t multivariado y slash multivariado



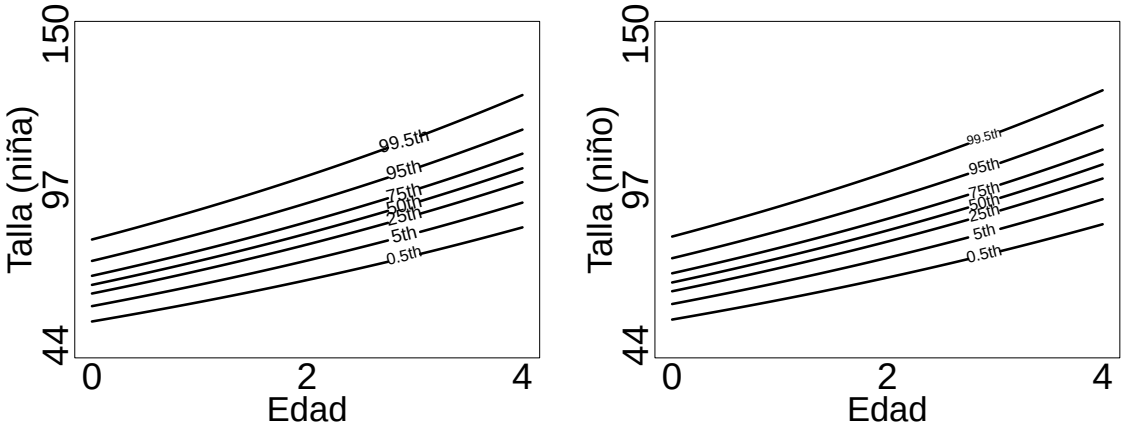
(a) (b)

Figura 6-10: Curvas de cuantiles ajustadas (para los percentiles 0.5, 5, 25, 50, 75, 95, 99.5 de abajo a arriba) para el perímetro braquial frente a la edad del niño/a, para el tiempo de leche materna fijada en su media



(a) (b)

Figura 6-11: Curvas de cuantiles ajustadas (para los percentiles 0.5, 5, 25, 50, 75, 95, 99.5 de abajo a arriba) para el peso frente a la edad del niño/a, para el tiempo de leche materna fijada en su media



(a) (b)

Figura 6-12: Curvas de cuantiles ajustadas (para los percentiles 0.5, 5, 25, 50, 75, 95, 99.5 de abajo a arriba) para la talla frente a la edad del niño/a, para el tiempo de leche materna fijada en su media

En el próximo capítulo se expondrán algunas de las conclusiones obtenidas después de realizar el trabajo expuesto en este documento.

7 Conclusiones y recomendaciones

7.1. Conclusiones

En este trabajo de investigación, se propuso un enfoque para el modelado de cuantiles marginales en presencia de datos faltantes a través de la clase de modelos de regresión lineal asociado a la clase de distribuciones normal/independiente multivariadas, adoptando un enfoque Bayesiano. Este enfoque considera la asociación entre las variables del vector de respuesta.

Se abordaron varias propiedades esenciales de la clase de distribuciones normal/independiente multivariadas, así como su relación con la clase de distribuciones log-normal/independiente multivariadas. Se describió la conexión entre los modelos de regresión lineal multivariados asociados a cada clase, demostrando cómo esta relación facilita la estimación robusta de los parámetros, especialmente para datos con colas posiblemente más pesadas que la distribución normal, además, la modelación de cuantiles marginales, considerando la asociación entre las variables respuesta.

El trabajo también se centró en la problemática de los datos faltantes, proporcionando una descripción de diferentes enfoques para abordarlos y explicando la tipología de este tipo de datos. Se detalló el procedimiento de simulación posterior de los parámetros del modelo mediante el algoritmo MDA, que aprovecha una estructura de datos faltantes importante, la estructura monótona, facilitando así el tratamiento de este tipo de datos y la simulación de los parámetros desde su distribución posterior.

Se llevó a cabo un estudio de simulación, orientado a la estimación de los parámetros de los modelos log-normal/independiente multivariados, utilizando el modelo t multivariado y el modelo slash multivariado, respaldando de manera sólida la metodología propuesta y demostrando un desempeño satisfactorio en la estimación de cuantiles.

Finalmente, se presentó y discutió una aplicación práctica utilizando un conjunto de datos reales de niños de 0 a 4 años pertenecientes a la comuna de Robledo, Medellín. Este análisis proporcionó una alternativa valiosa para la construcción de curvas de cuantiles en presencia de datos faltantes, mostrando la aplicabilidad y robustez de la metodología propuesta en un contexto empírico.

7.2. Recomendaciones

- Sería de interés explorar la optimización del código en el software estadístico R, con el objetivo de mejorar su eficiencia en términos algorítmicos. Además, se podría considerar la creación de una biblioteca específica en R para organizar y modularizar el código, facilitando su reutilización y mantenimiento a lo largo del tiempo.
- Realizar un procedimiento inferencial más exhaustivo, es decir, una evaluación más profunda en cuanto a la significancia de los coeficientes del modelo, además de incorporar pruebas de hipótesis Bayesianas para dichos coeficientes.
- Proponer la implementación de un procedimiento para la regresión robusta de modelos lineales mixtos, asociados a la clase de distribuciones log-normal/independiente multivariadas, bajo la presencia de datos faltantes. Este proceso se basaría en la relación entre esta clase y la clase de distribuciones normal/independiente multivariadas.

Referencias

Abanto-Valle, C. A., Lachos, V. H., y Ghosh, P. (2012). A bayesian approach to term structure modeling using heavy-tailed distributions. *Applied stochastic models in business and industry*, *28*(5), 430–447.

Andrews, D. F., y Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(1), 99–102.

Arellano-Valle, R., Galea-Rojas, M., y Zuazola, P. I. (2000). Bayesian sensitivity analysis in elliptical linear regression models. *Journal of Statistical Planning and Inference*, *86*(1), 175–199.

Arslan, O. (2008). An alternative multivariate skew-slash distribution. *Statistics & Probability Letters*, *78*(16), 2756–2761.

Arslan, O., y Genç, A. İ. (2009). A generalization of the multivariate slash distribution. *Journal of Statistical Planning and Inference*, *139*(3), 1164–1170.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, *68*(1), 13–20.

Bartlett, M. S. (1934). Xx.—on the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, *53*, 260–283.

Beale, E. M., y Little, R. J. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *37*(1), 129–145.

Boris Choy, S., y Chan, J. S. (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zealand Journal of Statistics*, *50*(2), 135–146.

Box, G. E., y Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *26*(2), 211–243.

Box, G. E., y Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. John Wiley & Sons.

Buchinsky, M. (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, 88–126.

Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Quartagno, M., y Kenward, M. G. (2023). *Multiple imputation and its application*. John Wiley & Sons.

Chakraborty, B. (2003). On multivariate quantile regression. *Journal of statistical planning and inference*, *110*(1-2), 109–132.

Chen, G., y Luo, S. (2016). Robust bayesian hierarchical model using normal/independent distributions. *Biometrical Journal*, *58*(4), 831–851.

De la Cruz, R. (2014). Bayesian analysis for nonlinear mixed-effects models under heavy-

tailed distributions. *Pharmaceutical statistics*, *13*(1), 81–93.

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

De Santis, F., y Spezzaferri, F. (1999). Methods for default and robust bayesian model comparison: the fractional bayes factor approach. *International Statistical Review*, *67*(3), 267–286.

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

Ferrari, S. L., y Fumes, G. (2017). Box–cox symmetric distributions and applications to nutritional data. *AStA Advances in Statistical Analysis*, *101*, 321–344.

Galarza Morales, C., Lachos Davila, V., Barbosa Cabral, C., y Castro Cepero, L. (2017). Robust quantile regression using a generalized class of skewed distributions. *Stat*, *6*(1), 113–130.

Garay, A. M., Bolfarine, H., Lachos, V. H., y Cabral, C. R. (2015). Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. *Journal of Applied Statistics*, *42*(12), 2694–2714.

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

Genç, A. İ. (2007). A generalization of the univariate slash by a scale-mixtured exponential power distribution. *Communications in Statistics—Simulation and Computation®*, *36*(5), 937–947.

Gómez, H. W., Quintana, F. A., y Torres, F. J. (2007). A new family of slash-distributions with elliptical contours. *Statistics & probability letters*, *77*(7), 717–725.

Han, P., Kong, L., Zhao, J., y Zhou, X. (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *81*(2), 305–333.

Howarth, T., Ben Saad, H., y Heraganahally, S. S. (2023). The impact of lung function parameters on sleep among aboriginal australians–a polysomnography and spirometry relationship study. *Nature and Science of Sleep*, 449–464.

Hunter, D. R., y Lange, K. (2000). Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, *9*(1), 60–77.

Kafadar, K. (2004). Slash distribution. *Encyclopedia of statistical sciences*, *12*.

Kibria, B. G., y Joarder, A. H. (2006). A short review of multivariate t-distribution. *Journal of Statistical research*, *40*(1), 59–72.

Kleinke, K., Fritsch, M., Stemmler, M., Reinecke, J., y Lösel, F. (2021). Quantile regression-based multiple imputation of missing values—an evaluation and application to corporal punishment data. *Methodology*, *17*(3), 205–230.

Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge university press.

Koenker, R., y Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the*

Econometric Society, 33–50.

Kotz, S., y Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Lachos, V. H., Bandyopadhyay, D., y Dey, D. K. (2011). Linear and nonlinear mixed-effects models for censored hiv viral loads using normal/independent distributions. *Biometrics*, *67*(4), 1594–1604.

Lange, K., y Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, *2*(2), 175–198.

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

Lee, S. X., y McLachlan, G. J. (2014). Scale mixture distribution. *Wiley StatsRef: Statistics Reference Online*, 1–16.

Li, K.-H. (1988). Imputation using markov chains. *Journal of Statistical Computation and Simulation*, *30*(1), 57–79.

Lin, P.-E. (1972). Some characterizations of the multivariate t distribution. *Journal of Multivariate Analysis*, *2*(3), 339–344.

Little, R. J. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *37*(1), 23–38.

Little, R. J. A., y Rubin, D. B. (1987). *Statistical analysis with missing data* (First ed.). Hoboken, NJ : Wiley: Wiley Series in Probability and Statistics.

Liu, C. (1993). Bartlett s decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *Journal of Multivariate Analysis*, *46*(2), 198–206.

Liu, C. (1994). *Statistical analysis using the multivariate t distribution*. Harvard University.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of multivariate analysis*, *53*(1), 139–158.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association*, *91*(435), 1219–1227.

Liu, C. (1997). Ml estimation of the multivariatetdistribution and the em algorithm. *Journal of Multivariate Analysis*, *63*(2), 296–312.

Liu, C., y Rubin, D. B. (1995). Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, 19–39.

Meng, X.-L., y Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, *80*(2), 267–278.

Montgomery, D. C., Peck, E. A., y Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Morán-Vásquez, R. A., y Ferrari, S. L. (2019). Box–cox elliptical distributions with application. *Metrika*, *82*(5), 547–571.

Morán-Vásquez, R. A., Giraldo-Melo, A. D., y Mazo-Lopera, M. A. (2023). Quantile estimation using the log-skew-normal linear regression model with application to children's weight data. *Mathematics*, *11*(17), 3736.

Morán-Vásquez, R. A., Mazo-Lopera, M. A., y Ferrari, S. L. (2021). Quantile modeling through multivariate log-normal/independent linear regression models with application to newborn data. *Biometrical Journal*, *63*(6), 1290–1308.

Morán-Vásquez, R. A., Roldán-Correa, A., y Nagar, D. K. (2023). Quantile-based multivariate log-normal distribution. *Symmetry*, *15*(8), 1513.

Petrella, L., y Raponi, V. (2019). Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. *Journal of Multivariate Analysis*, *173*, 70–84.

Quiroz, A. J., Nakamura, M., y Pérez, F. J. (1996). Estimation of a multivariate box-cox transformation to elliptical symmetry via the empirical characteristic function. *Annals of the Institute of Statistical Mathematics*, *48*, 687–709.

Raghunathan, T. (2015). *Missing data analysis in practice*. CRC press.

Reyes, J., Gallardo, D. I., Bolfarine, H., y Gómez, H. W. (2019). A new class of slash-elliptical distributions. *Communications in Statistics-Theory and Methods*, *48*(12), 3105–3121.

Reyes, J., y Iriarte, Y. A. (2023). A new family of modified slash distributions with applications. *Mathematics*, *11*(13), 3018.

Ritter, C., y Tanner, M. A. (1992). Facilitating the gibbs sampler: the gibbs stopper and the griddy-gibbs sampler. *Journal of the American Statistical Association*, *87*(419), 861–868.

Rogers, W. H., y Tukey, J. W. (1972). Understanding some long-tailed symmetrical distributions. *Statistica Neerlandica*, *26*(3), 211–226.

Rosa, G., Padovani, C. R., y Gianola, D. (2003). Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *45*(5), 573–590.

Roth, S., M'Pembele, R., Stroda, A., Voit, J., Lurati Buse, G., Sixt, S. U., ... others (2022). Days alive and out of hospital after left ventricular assist device implantation. *ESC Heart Failure*, *9*(4), 2455–2463.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1983). Iteratively reweighted least squares. *Encyclopedia of statistical sciences*, *6*.

Rubin, D. B., y Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. En *Proceedings of the statistical computing section of the american statistical association* (Vol. 83, p. 88).

Sánchez, B., Lachos, H., y Labra, V. (2013). Likelihood based inference for quantile regression using the asymmetric laplace distribution. *Journal of Statistical Computation and Simulation*, *81*, 1565–1578.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schmeiser, B. W., y Lal, R. (1980). Squeeze methods for generating gamma variates. *Journal of the American Statistical Association*, *75*(371), 679–682.

Stasinopoulos, M., Rigby, B., Voudouris, V., Akantziliotou, C., Enea, M., y Kiose, D. (2023). Package 'gamlss'. *Dist'2020Available online: http://www. gamlss. org (accessed on 16 July 2021)*.

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25.

Tang, Y. (2015). An efficient monotone data augmentation algorithm for bayesian analysis of incomplete longitudinal data. *Statistics & Probability Letters*, *104*, 146–152.

Tanner, M. A., y Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Tian, Y., Tian, M., y Zhu, Q. (2014). Linear quantile regression based on em algorithm. *Communications in Statistics-Theory and Methods*, *43*(16), 3464–3484.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van Dyk, D. A., y Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1–50.

Van Dyk, D. A., y Meng, X.-L. (2010). Cross-fertilizing strategies for better em mountain climbing and da field exploration: A graphical guide book.

Vanegas, L. H., y Paula, G. A. (2016). Log-symmetric distributions: statistical properties and parameter estimation.

Verhasselt, A., Flórez, A. J., Van Keilegom, I., y Molenberghs, G. (2019). The impact of incomplete data on quantile regression for longitudinal data. *FEB Research Report KBI_1906*.

Wang, C., Tian, M., y Tang, M.-L. (2022). Nonparametric quantile regression with missing data using local estimating equations. *Journal of Nonparametric Statistics*, *34*(1), 164–186.

Wang, J., y Genton, M. G. (2006). The multivariate skew-slash distribution. *Journal of Statistical Planning and Inference*, *136*(1), 209–220.

Wei, Y. (2008). An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *Journal of the American Statistical Association*, *103*(481), 397–409.

WHO. (2006). *Who child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development*. World Health Organization.

WHO. (2007). *World health organization child growth standards: head circumference-for-age, arm circumference-for-age, triceps skinfold-for-age and subscapular skinfold-for-age: methods and development*. World Health Organization.

Wichitaksorn, N., Choy, S. B., y Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *Canadian Journal of Statistics*, *42*(4), 579–596.

Yu, K., y Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability*

Letters, *54*(4), 437–447.