PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Modelo de Regresión Lineal con Censura Basado en una Mixtura Finita de una Distribución Normal Asimétrica

Tesis para obtener el grado académico de Magíster en Estadística que presenta:

Ingrid Alicia Yábar Geldres

Asesor:

Dr. Luis Enrique Benites Sánchez

Lima, 2023

Informe de Similitud

Yo, Luis Enrique Benites Sánchez, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulada Modelo de Regresión Lineal con Censura Basado en una Mixtura Finita de una Distribución Normal Asimétrica, de la autora Ingrid Alicia Yábar Geldres.
Dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 9 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 03/05/2022.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 10 de mayo 2023

Apellidos y nombres del asesor / de la asesora: Benites Sánchez, Luis Enrique	
DNI : 42987865	Firma : 
ORCID : 0000000159987098	

Dedicatoria

Con gran admiración, este trabajo está dedicado a mis amados padres, por haber sido el apoyo para no rendirme y por acompañarme con su amor incondicional durante todo el recorrido de preparación en mi vida profesional.

Agradecimientos

A Dios, por ser mi guía y la luz en cada paso y decisión que he tomado.

A mis queridos profesores de la Universidad Nacional Mayor de San Marcos (Rosario Bullón, Estela Ponce y Erwin Kraenau) quienes lograron fortalecer mi vocación y amor hacia la carrera Estadística.

A los profesores Luis Valdivieso, José Flores, Enver Tarazona y Giancarlo Sal y Rosas, profesores de la Maestría en Estadística, quienes fortalecieron mis conocimientos hacia la carrera, en todos los cursos llevados.

A mi estimado asesor Luis Benites, por haberme respaldado con sus conocimientos durante el desarrollo de mi trabajo.

Resumen

El presente trabajo de tesis propone estudiar el modelo de regresión lineal con censura basado en una mixtura finita de una distribución normal asimétrica (NA), con adaptación a diferente número de componentes. Este enfoque permite modelar datos continuos con gran flexibilidad, acomodando simultáneamente multimodalidad, colas pesadas y asimetría, dependiendo de la estructura de los componentes de la mixtura. Se implementa un algoritmo de tipo EM analíticamente manejable y eficiente para calcular iterativamente las estimaciones de máxima verosimilitud de los parámetros, mediante aproximaciones estocásticas (SAEM). El algoritmo propuesto tiene algunas expresiones cerradas en el paso-E, por lo que la obtención de los errores estándar se da por el método Bootstrap.

Asimismo, se realiza un estudio de simulación con el fin de evaluar si el método propuesto permite recuperar los parámetros del modelo mediante el uso del algoritmo SAEM.

Por otro lado, se realiza la aplicación del modelo propuesto para el estudio de la participación en la fuerza laboral de las mujeres casadas usando la base de datos de la Universidad de Michigan (Mroz, 1987). Como segunda aplicación se utiliza un conjunto de datos de clientes que entraron en campaña en una entidad financiera local con el fin de estimar sus ingresos.

Palabras-clave: Mixtura finita, Modelo de regresión no-lineal, Algoritmo EM, Algoritmo SAEM, Distribución de Mixtura de Escala Normal, Bootstrap.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Los modelos de regresión lineal asociados a una variable de interés Y en función de un conjunto de covariables $\boldsymbol{x_i}$, se expresan de la siguiente forma:

$$y_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \varepsilon_i \ , \tag{1.1}$$

representando y_i a la variable dependiente, $\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ al vector de covariables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ a los parámetros de la regresión lineal y ε_i a los errores por cada observación i. Con fines prácticos es común asumir normalidad en los errores, supuesto que no siempre se satisface debido a que pueden presentarse casos con multimodalidad y asimetría, ver Wei y Tanner (1990). Un tratamiento para dar soluciones en caso se presente multimodalidad y asimetría es mediante transformaciones de datos, que pueden generar una normalidad aproximada con resultados empíricos razonables. A pesar de ello, esta medida puede generar pérdida de información y problemas de interpretabilidad de los resultados, ver Garay et al. (2016).

En ese sentido, trabajos previos han mostrado la importancia de considerar distribuciones más robustas que la distribución normal simétrica, por ejemplo, Fernandez y Steel (1999), discuten las dificultades en los modelos de regresión cuando el error tiene una distribución t de Student multivariada. Benites et al. (2019), utiliza la mixtura como método semiparamétrico conveniente, que se encuentra entre los modelos paramétricos y los estimadores de densidad del kernel para modelar la forma de distribución desconocida de los errores.

Por otro lado, Garay et al. (2017) establece un nuevo vínculo entre el modelo de regresión censurado y una clase de distribuciones simétricas. Estas estructuras proporcionan modelos robustos y adaptables, por ejemplo, al modelo con distribuciones de escala normal, presentado por Andrews y Mallows (1974) y al modelo de regresión lineal con censura basado en mixtura finita en el error presentado por Benites et al. (2019) y Benites et al. (2018). En este último se muestra el buen funcionamiento del modelo en presencia de valores atípicos y multimodalidad ya que se considera una mixtura finita de la distribución t de student y una mixtura de escala normal (MEN), en el término del error, para el caso lineal. La clase de distribuciones de mixtura de escala normal asimétrica (MENA) contiene varios miembros con colas pesadas más que en una distribución normal, tales como la normal asimétrica, t asimétrica, Slash asimétrica y la normal contaminada asimétrica. Asimismo, Branco y Dey

(2001), propusieron unas clases de mixturas normales asimétricas tales como la mixtura finita de normales asimétricas, logística asimétrica, asimétria equilibrada, t asimétrica, etc. Por otro lado, cuando la información sobre la variable de interés Y es incompleta, una alternativa a considerar es el modelo de regresión censurado basado en el desarrollo del modelo Tobit, en términos del supuesto de normalidad. Para esto, la variable Y presenta una concentración de las observaciones dependiendo del tipo de censura que tenga (izquierda, derecha o de forma intervalar), lo cual viola los supuestos del modelo de regresión múltiple, por lo que es necesario un modelo que permita representar mejor este tipo de datos (Tobin, 1958). El modelo Tobit asume que los errores cumple con el supuesto de normalidad, pero en situaciones reales la variable de interés no cumple con dicho supuesto, por lo cual, el estimador de máxima verosimilitud resulta inconsistente, ver Zeller et al. (2018). En este sentido, Caudill (2012) introduce un estimador parcialmente adaptativo para el modelo de regresión censurado basado en una estructura de errores con mixtura de distribuciones normales, y Garay et al. (2016), establece un vínculo entre el modelo de regresión no lineal censurado y una clase de distribuciones simétricas recientemente estudiadas, que se extiende a la normal mediante la inclusión de la curtosis, llamada distribucion de mixtura de escala de normales.

Según lo anteriormente mencionado, en la presente tesis se propone un modelo lineal censurado basado en una mixtura finita de una distribución normal asimétrica (NA), la cual proporciona flexibilidad para admitir el efecto de la asimetría y cola pesada para la variable respuesta censurada a la izquierda, con adaptación para diferentes números de componentes, con base en el estudio realizado por Thalita et al. (2017).

1.1. Objetivos

El objetivo general de la tesis es generalizar el modelo de regresión lineal con censura basado en una mixtura finita de una distribución normal asimétrica (MRL-CR-MF-NA) propuesto por Thalita et al. (2017), mediante la inclusión de varias componentes.

De manera específica:

- Revisar la literatura acerca de las diferentes propuestas de los modelos de regresión lineal censurado.
- Proponer y estudiar las propiedades y beneficios del modelo propuesto.
- Implementar el método de estimación en el lenguaje R (R Core Team, 2021).
- Desarrollar estudios de simulación mediante el algoritmo SAEM, que permitan ilustrar el desempeño de la metodología utilizada.
- Aplicar el modelo a un conjunto de datos reales.

1.2. Organización del Trabajo

El trabajo es organizado como sigue: En el Capítulo 2 se presenta una serie de conceptos preliminares asociados al modelo a desarrollar en la presente tesis. Conceptos como

censura, truncamiento, así como del algoritmo EM y el algoritmo SAEM, siendo este último un caso particular del primer algoritmo, el cual permite estimar de manera más eficiente los parámetros con observaciones incompletas. También se presenta la distribución normal asimétrica (NA). En el Capítulo 3, se aborda el modelo propuesto, es decir, el MRL-CI-MF-NA adicionando algunas propiedades, gráficos de la función de densidad de probabilidad y el algoritmo SAEM para la estimación de máxima verosimilitud. Asimismo, se detalla el método para la estimación de los parámetros mediante un método de aproximación estocástica (SAEM) y el método Bootstrap para la obtención de los errores estándar con un procedimiento para la regla de parada. En el Capítulo 4, se presenta un estudio de simulación con diferentes escenarios para la generación de datos, con el fin de evaluar si el método propuesto permite recuperar los parámetros del modelo de la regresión lineal con censura a la izquierda, cuando los errores siguen una mixtura finita de distribuciones normales asimétricas. Con el fin de elegir el mejor modelo se utiliza el criterio de información de Akaike (AIC, por sus siglas en inglés) y el criterio de información bayesiano (BIC, por sus siglas en inglés), así como el sesgo y el error absoluto medio (MAE, por sus siglas en inglés). En el Capítulo 5, se muestra dos aplicaciones del modelo con datos reales. Finalmente, en el Capítulo 6 se presentan las principales conclusiones y sugerencias para futuras investigaciones.

En el anexo se presentará algunos cuadros adicionales del estudio de simulación (Apéndice A).

Capítulo 2

Conceptos preliminares

A continuación, resaltamos algunos conceptos relevantes que emplearemos en la presente tesis. Dado que utilizamos los modelos de regresión lineal cuando la variable respuesta presenta observaciones incompletas por causa de la censura, será importante diferenciar en primer lugar los conceptos de truncamiento y censura, que son dos de las principales causas de datos incompletos.

2.1. Preliminares

En esta sección, proporcionamos algunos conceptos útiles sobre censura, truncamiento y describimos el algoritmo EM y algoritmo SAEM para la estimación de un modelo de regresión lineal.

2.1.1. Censura

La censura se define como la ocurrencia de un evento de interés Y en un tiempo T en el que la información que proporciona el sujeto de estudio es incompleta o no está completamente disponible para algunas unidades de la muestra sobre su tiempo de vida; sin embargo, para estas unidades, los datos sobre las variables regresoras son totalmente conocidos. Así, una observación censurada contiene solo información parcial sobre el evento de interés. De esta forma las observaciones censuradas no son cuantificadas y solo son conocidas por exceder o ser inferiores a un valor umbral, al cual denotaremos por k_i para la observación i. De esta forma se pueden considerar varios tipos de censura, como censura por la derecha, censura por la izquierda y censura dentro de un intervalo. Sin embargo, en la presente tesis el enfoque estará dado solo para datos censurados por la izquierda.

- Una variable censurada a la izquierda Y se da cuando tenemos n observaciones de Y, pero solo conocemos el verdadero valor de Y para una cantidad restringida de observaciones, debido al hecho que el evento de interés sucedió antes que el sujeto haya sido incluido en el estudio, es decir, se desconoce la información sobre el tiempo exacto en que ocurrió el evento de interés, y solo se sabe que ocurrió antes (Bogaertz et al., 2017).

  De forma general, toda variable censurada por la izquierda, se define como:

$$Y_i = \begin{cases} k_i, & \text{si } y_i \leq k_i \\ y_i, & \text{si } y_i > k_i \end{cases} \tag{2.1}$$

Note que $P(Y_i = k_i) = P(Y_i \leq k_i) \geq 0$, ya que $Y \in \mathbb{R}$. Luego,

$$Y_i = \text{máx}\{y_i, k_i\}, \ \forall i.$$

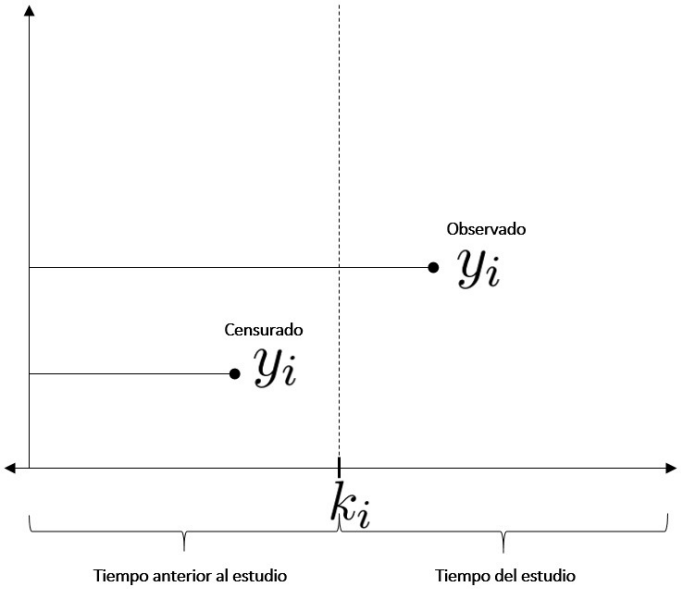


Figura 2.1: Censura por la izquierda

En la Figura 2.1 se observa un caso particular donde el evento de interés Y_i se ha presentado antes del tiempo de inicio del estudio ($y_i \leq k_i$), lo cual implica de forma general el desconocimiento del tiempo exacto en que se inició el evento de interés, siendo para este caso censurado por la izquierda y tomando como valor k_i y para el caso en que la observación se presenta dentro del tiempo de estudio tomará el valor de y_i.

- Según Klein y Moeschberger (2003), una variable es censurada por la derecha cuando el tiempo de ocurrencia del evento es mayor que un tiempo definido para el final de dicho evento, es decir, la información es desconocida.

  De forma general, toda variable censurada por la derecha, se define como:

$$Y_i = \begin{cases} k_i, & \text{si } y_i > k_i \\ y_i, & \text{si } y_i \leq k_i \end{cases} \tag{2.2}$$

Note que $P(Y_i = k_i) = P(Y_i > k_i) \geq 0$, ya que $Y \in \mathbb{R}$. Luego,

$$Y_i = \text{mín}\{y_i, k_i\}, \ \forall i.$$

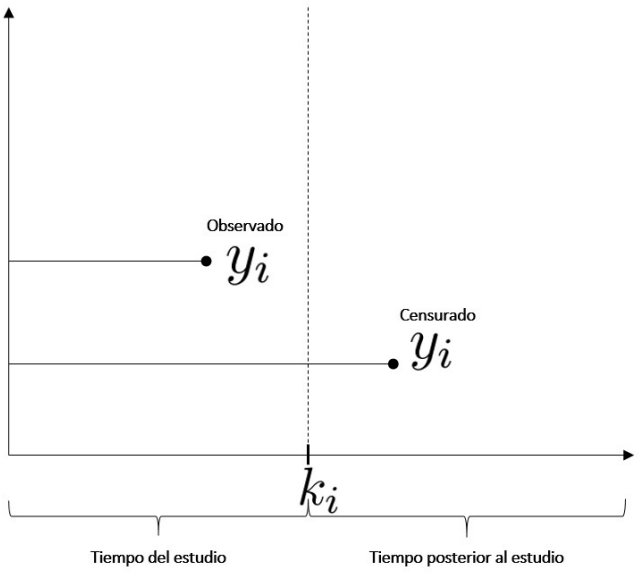


Figura 2.2: Censura por la derecha

En la Figura 2.2 se observa un caso particular donde el evento de interés Y_i se ha presentado después del tiempo del fin del estudio ($y_i > k_i$), lo cual implica de forma general el desconocimiento del tiempo exacto en que se iniciará el evento de interés, siendo para este caso censurado por la derecha y tomando como valor k_i y para el caso en que la observación se presenta dentro del tiempo de estudio tomará el valor de y_i.

2.1.2. Truncamiento

El truncamiento se define como la ocurrencia de un evento de interés Y en un tiempo T, donde las observaciones de las variables explicativas X solo se observan cuando se cumple una determinada condición. Es decir, el truncamiento proporciona una muestra aleatoria de una parte de los individuos a estudiar; para aquellos que verifiquen la ocurrencia del evento de interés sobre la variable Y, dentro de una ventana observacional, para lo cual la inferencia para datos truncados, esta restringida a la estimación condicionada. Esto es contrastado con la censura, donde hay por lo menos información parcial sobre la variable Y, en cambio para el truncamiento, estos individuos nunca fueron incluidos en el estudio (Klein y Moeschberger, 2003). Se considera truncamiento por la izquierda si la observación y_i para el cual su tiempo de falla T_i es superior de un tiempo o cota inferior U_i. U_i representa el tiempo en el que ocurre el evento de truncamiento. Las observaciones cuyo tiempo de falla es inferior al tiempo de truncamiento no son considerados en el estudio. Este caso particular será estudiado en la presente tesis.

De forma general, toda variable truncada por la izquierda, se define como:

$$Y = Y_i, \;\; si \;\; T_i > U_i.$$

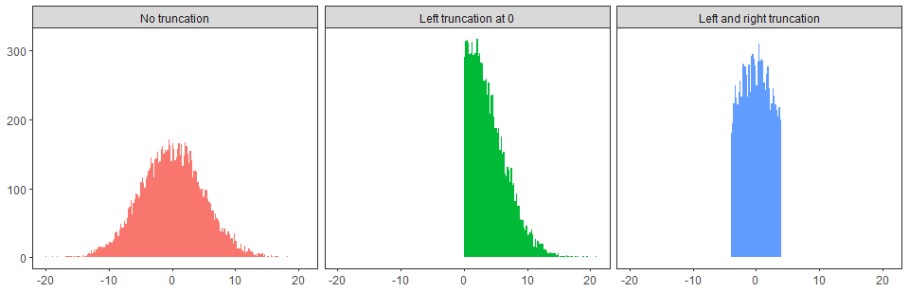


Figura 2.3: Distribución normal con $\mu = 0$ y $\sigma = 5$ | Primer gráfico: Sin truncar | Segundo gráfico: Truncada a la izquierda en 0 | Tercer gráfico: Truncada a la izquierda y a la derecha en $(-4, 4)$

2.1.3. Algoritmo Esperanza - Maximización (Algoritmo EM)

El algoritmo de Esperanza - Maximización (EM), es un proceso iterativo que permite la maximización de la función de verosimilitud de manera más eficiente, cuando los procedimientos tradicionales resultan ser analíticamente intratables, no viables o las observaciones pueden ser incompletas. Según Dempster et al. (1977), el término de "datos incompletos" en su forma general implica la existencia de dos espacios muestrales Ω_X y Ω_Y y una función de muchos a uno de Ω_X a Ω_Y. Los datos observados y son una realización de Ω_Y. El correspondiente $x \in \Omega_X$ no es directamente observable, sino indirectamente observable a través de Ω_Y. Específicamente, se asume que existe una función de $x \mapsto y(x)$ de Ω_X a Ω_Y, y se sabe que x se encuentra en un subconjunto de Ω_X tal que $y = y(x)$, donde y es el dato observado.

Sea $\boldsymbol{z} = (y(x), x) \in \Omega_Y \times \Omega_X$ el vector de datos completos, donde y son los datos observados y x los datos perdidos, con su correspondiente función densidad de probabilidad $f_Z(\boldsymbol{z}|\boldsymbol{\theta})$ dependiente del parámetro $\boldsymbol{\theta}$ y se calcula la función densidad de probabilidad de los datos observados $g_Y(y|\boldsymbol{\theta})$. La especificación de los datos completos $f_Z(\boldsymbol{z}|\boldsymbol{\theta})$ está directamente relacionado con la datos observados (datos incompletos) $g_Y(y|\boldsymbol{\theta})$ por:

$$g_Y(y|\boldsymbol{\theta}) = \int_{\Omega_X} f_Z(\boldsymbol{z}|\boldsymbol{\theta})\, dx = \int_{\Omega_X} f_{X\times Y}(y, x|\boldsymbol{\theta})\, dx$$

Este algoritmo fue introducido con la finalidad de calcular un valor de $\boldsymbol{\theta}$, el cual maximiza $g_Y(y|\boldsymbol{\theta})$ dado una observación y, pero lo hace a través de la función de densidad de probabilidad de los datos completos $f_Z(\boldsymbol{z}|\boldsymbol{\theta})$. Del vector de datos completos se encuentra la función log-verosimilitud completa $\ell_c(\boldsymbol{\theta}|\boldsymbol{z})$.

El algoritmo EM en primer lugar encuentra una estimación de $\ell_c(\boldsymbol{\theta}|\boldsymbol{z})$ que representa la función log-verosimilitud completa a través de la función de densidad de probabilidad condicional de $Y|X$ y la función de densidad de probabilidad de X. Lo siguiente es calcular la esperanza condicional de $\ell_c(\boldsymbol{\theta}|\boldsymbol{z})$ en términos de la función de densidad de probabilidad de $X|Y$ para la k-ésima estimación de los parámetros (Paso E) y luego se maximiza esta esperanza, obteniendo la estimación de $\boldsymbol{\theta}$ para la iteración $k+1$ (Paso M), generando estabilidad, es decir, por cada iteración aumenta la verosimilitud incompleta (Delyon et al., 1999). Formalmente, si:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \;=\; E_{\boldsymbol{\theta}^{(k)}}\left[\ell_c(\boldsymbol{\theta}|\boldsymbol{z})|\widehat{\boldsymbol{\theta}}^{(k)}, x\right]. \tag{2.3}$$

El algoritmo EM consta de los siguientes pasos:

- **Paso E:** En este paso se calculará $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$.
- **Paso M:** En este paso se procede a maximizar $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ con respecto a cada $\boldsymbol{\theta}$, lo cual genera la estimación $\widehat{\boldsymbol{\theta}}^{(k+1)}$, es decir:

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \underset{\boldsymbol{\theta}}{argmax} Q(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(k)}).$$

Estos dos pasos se repiten iterativamente hasta la convergencia de la sucesión $\widehat{\boldsymbol{\theta}}^{(k)}$.

Como lo mencionan Meza y De la Cruz (2012), cada iteración del algoritmo EM incrementa la función verosimilitud $\ell(\boldsymbol{\theta}|y)$ y la secuencia del algoritmo EM, $\boldsymbol{\theta}^{(k)}$ converge a un punto estacionario de la verosimilitud observada bajo condiciones regulares leves, para más detalles, ver Wu (1983) y Vaida (2005).

2.1.4. Algoritmo de Aproximación Estocástica EM (SAEM)

En ciertas situaciones calcular la función $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ en (2.3) puede ser númericamente complicado o incluso intratable. Por esto, para lidiar con esta limitación Delyon et al. (1999), propuso una aproximación estocástica a partir del algoritmo EM, llamado el algoritmo SAEM, el cual consiste en reemplazar el paso E, por un paso S y un paso AE, donde el paso S consiste en generar observaciones del vector de datos perdidos x con la distribución a posteriori de los datos perdidos dado los datos observados, y el paso AE consiste en una aproximación estocástica obtenida a partir de los datos simulados. Por otra, parte el paso M permenace sin cambios.

Las fases del algoritmo SAEM en la k-ésima iteración, se detallan a continuación:

- El paso S consiste en generar m simulaciones $x_k(l)$ con $l = 1, ..., m$ de los datos perdidos de la distribución aposteriori $f(x|y, \widehat{\boldsymbol{\theta}}^{(k)})$.
- En el paso de aproximación estocástica (AE) se actualiza $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k+1)})$ acorde a:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k+1)}) \approx Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) + \gamma_k \left[\frac{1}{m}\sum_{l=1}^{m} l_c(\boldsymbol{\theta}|y, x_k(l)) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})\right], \tag{2.4}$$

según Kuhn y Lavielle (2004), donde γ_k es una sucesión decreciente de números positivos tal que:

$$\sum_{k=1}^{\infty}\gamma_k = \infty \;\text{ y }\; \sum_{k=1}^{\infty}\gamma_k^2 < \infty.$$

- El paso M maximiza $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ en $\boldsymbol{\theta}$ obteniéndose $\widehat{\boldsymbol{\theta}}^{(k+1)}$ tal que:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k+1)}) \geq Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}).$$

Es importante mencionar que el algoritmo SAEM no solo usa la simulación actual de los datos faltantes x en la iteración k-ésima, sino también utiliza algunas o todas las simulaciones anteriores, representando esta "memoria" la propiedad para suavizar el parámetro γ_k. De esta forma en (2.4) la contribución del parámetro γ_k es fuerte en cuanto a la velocidad de la convergencia del algoritmo. Esto quiere decir que, si el parámetro suavizado γ_k equivale a 1 para todo k, entonces el algoritmo SAEM presentará una convergencia rápida en la distribución (sin memoria). En caso contrario, ante una convergencia lenta (con memoria) la convergencia es casi segura.

Según la sugerencia de elección del parámetro suavizado por Galarza et al. (2015), se tiene:

$$\gamma_k = \begin{cases} 1, & \text{si} \quad 1 \leq \ k \leq \kappa_i S, \\ \dfrac{1}{k - \kappa_i S}, & \text{si} \quad \kappa_i S + 1 \leq k \leq S, \end{cases} \tag{2.5}$$

donde S representa el máximo número de iteraciones, y κ_i es el punto de corte ($0 \leq \kappa_i \leq 1$), el cual determina el porcentaje de iteraciones iniciales, donde un número κ_i entre 0 y 1 asegurará una convergencia inicial en la distribución a una vecindad de solución para las primeras iteraciones $\kappa_i S$ y una convergencia casi segura para el resto de las iteraciones. Resultando a partir de la combinación un algoritmo rápido con buenas estimaciones.

Según Thalita (2016), para implementar el SAEM, el usuario debe corregir varias constantes que coincidan con el número total de iteraciones S y el punto de corte κ_i que define el inicio del paso de suavizado del algoritmo SAEM. Sin embargo, esos parámetros variarán según el modelo y los datos. Para determinar esas constantes, se recomienda un enfoque gráfico que monitoree la convergencia de las estimaciones para todos los parámetros, y si es posible, que monitoree la siguiente diferencia (diferencia relativa) entre dos evaluaciones sucesivas de la log-verosimilitud, expresada por:

$$\parallel \ell(\boldsymbol{\theta}^{(k+1)}|y) - \ell(\boldsymbol{\theta}^{(k)}|y) \parallel o \parallel \ell(\boldsymbol{\theta}^{(k+1)}|y)/\ell(\boldsymbol{\theta}^{(k)}|y) - 1 \parallel .$$

2.2. Distribución Normal asimétrica

2.2.1. Definiciones

A continuación, se define la distribución normal asimétrica (NA), propuesta por Azzalini (1985).

Una variable aleatoria W tiene esta distribución con parámetro de localización μ, parámetro de escala σ^2 y parámetro de forma $\lambda \in \mathbb{R}$, la cual regula la asimetría de la función de densidad, si su función de densidad de probabilidad (fdp) está dada por:

$$f_{\text{NA}}(w|\mu, \sigma^2, \lambda) = 2f(w|\mu, \sigma^2)F\left(\frac{\lambda(w-\mu)}{\sigma}\right), \quad w \in \mathbb{R}, \tag{2.6}$$

donde $f(\cdot|\mu, \sigma^2)$ denota la densidad de la distribución normal con media μ y varianza σ^2, $F(\cdot)$ representa la función de distribución acumulada (fda) de la distribución normal estandar. Usaremos en adelante la notación $W \sim NA(\mu, \sigma^2, \lambda)$

La media y la varianza de W está dado por:

$$E(W) = \mu + \sqrt{\frac{2}{\pi}}\sigma\delta \quad y \quad V(W) = \sigma^2 - \frac{2}{\pi}\sigma^2\delta^2,$$

donde:

$$\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}.$$

De acuerdo con Azzalini (1985), una variable aleatoria $W \sim \text{NA}(\mu, \sigma^2, \lambda)$ tiene la siguiente representación estocástica:

$$W = \mu + \Delta \mid T_0 \mid + \tau^{1/2} T_1, \tag{2.7}$$

donde $\Delta = \sigma\delta, \tau = (1-\delta^2)\sigma^2, \delta = \dfrac{\lambda}{\sqrt{1+\lambda^2}}$, y T_0 y T_1 son variables aleatorias normales estándar e independientes.

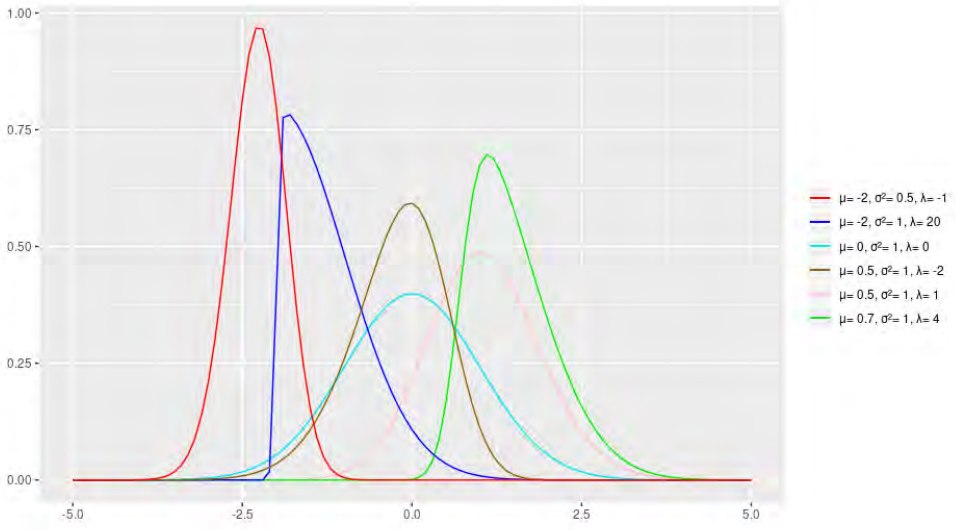


Figura 2.4: Función de densidad de probabilidad de la distribución normal asimétrica para diferentes valores de los parámetros (μ, σ^2, λ)

Capítulo 3

Modelo

3.1. Modelo de regresión lineal con censura a la izquierda basado en una mixtura finita de una distribución normal asimétrica (MRL-CI-MF-NA)

Considerando las ideas presentadas por Benites et al. (2018), definimos el siguiente modelo de regresión lineal con covariables $\boldsymbol{x_i}$ y errores que siguen una mixtura finita de distribuciones normales asimétricas (NA):

$$Y_i = \beta_0 + \boldsymbol{x_i}^\top \boldsymbol{\beta} + \varepsilon_i, \tag{3.1}$$

$$\varepsilon_i \sim \sum_{j=1}^{G} p_j \text{NA}(\mu_j + b\Delta_j, \sigma_j^2, \lambda_j), \;\; i = 1, \ldots, n, \tag{3.2}$$

donde el número de componentes G es conocido y fijo, y $\mathbf{p} = (p_1, \ldots, p_G)^\top$ es un vector de parámetros de las proporciones de la mixtura finita que satisfacen $\sum_{j=1}^{G} p_j = 1$. Adicionalmente, se asume que $\sum_{j=1}^{G} p_j \mu_j = 0$ para que $E(\varepsilon_i) = 0$ y λ_j representa al parámetro de asimetría de la j-ésima componente de la mixtura finita de los errores. De las ecuaciones (3.1) y (3.2) tenemos que la función de densidad de Y_i es:

$$f(y_i|\boldsymbol{x_i}, \boldsymbol{\theta}) = \sum_{j=1}^{G} p_j \, f_{NA}(y_i|\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j), \tag{3.3}$$

donde, $\mu_{ij} = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \vartheta_j$, $\vartheta_j = \beta_0 + \mu_j$, $b = -\sqrt{\frac{2}{\pi}}$, y $\Delta_j = \sigma_j \delta_j$, siendo $\delta_j = \dfrac{\lambda_j}{\sqrt{1+\lambda_j^2}}$. Aquí, estamos representando por $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$, al vector con todos los parámetros, donde $\boldsymbol{\beta} \subset \mathbb{R}^p$ y $\boldsymbol{\theta}_j = (p_j, \; \sigma_j^2, \; \mu_j, \; \lambda_j)^\top$ es el vector específico de parámetros para la componente j. Para cada observación $i = 1, \ldots, n$ en la ecuación (3.3), se define el vector aleatorio de componentes $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})^\top$, donde:

$$Z_{ij} = \begin{cases} 1 & \text{,si la } i\text{-ésima observación pertenece al grupo } j, \\ 0 & \text{,en otro caso} \end{cases}$$

Consecuentemente, $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iG})^\top \sim \text{Multinomial}(1; p_1 \ldots, p_G)$, con probabilidades $p_1,\ p_2, \ldots,\ p_G$, tal que $\sum_{j=1}^{G} Z_{ij} = 1$. La función de probabilidad de $\mathbf{Z}_i$ es:

$$f(\mathbf{Z}_i = z_i) = p_1^{z_{i1}} p_2^{z_{i2}} \ldots p_G^{z_{iG}},$$

donde $\sum_{j=1}^{G} p_j = 1$. Además, si $Z_{ij} = 1$, la media de la variable respuesta Y_i depende del predictor $\boldsymbol{x_i}$ en una forma lineal, tal que:

$$Y_i | Z_{ij} = 1 \sim NA(\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j), \quad j = 1, \ldots, G. \tag{3.4}$$

Note que, para $Z_{ij} = 1$, tenemos la siguiente representación jerárquica para (3.4):

$$\begin{aligned} Y_i | T_i = t_i, Z_{ij} = 1 &\sim \text{N}(\mu_{ij} + \Delta_j t_i, \tau_j), \\ T_i | Z_{ij} = 1 &\sim \text{TN}(b, 1, (b, \infty)), \\ \mathbf{Z}_i &\sim \text{Multinomial}(1, p_1, \ldots, p_G), \end{aligned} \tag{3.5}$$

donde, $\mu_{ij} = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \vartheta_j$, $\vartheta_j = \beta_0 + \mu_j$, $\Delta_j = \sigma_j \delta_j$, $\tau_j = (1 - \delta_j^2)\sigma_j^2$, y $\delta_j = \frac{\lambda_j}{\sqrt{1+\lambda_j^2}}$ para $i = 1,\ \ldots,\ n$, donde $\text{TN}(b, 1, (b, \infty))$ denota una distribución normal con media b y varianza 1, truncada en el intervalo (b, ∞).

Estaremos interesados en que las observaciones de la variable respuesta se encuentren incompletas. Así, para la i-ésima observación y para el supuesto de censura por la izquierda, Y_i representa una variable latente, cuya data observada toma la siguiente forma $(V_i, \mathbb{I}_i)$, donde:

$$V_i = \begin{cases} \kappa_i & \text{si} \quad \mathbb{I}_i = 1\ (Y_i \leq \kappa_i) \\ Y_i & \text{si} \quad \mathbb{I}_i = 0\ (Y_i > \kappa_i), \end{cases}$$

$i = 1, \ldots, n$, para algún punto de corte conocido κ_i e $\mathbb{I}_i$ es la función indicadora de censura para la i-ésima observación. Se define la función indicadora de censura:

$$\mathbb{I}_i = \begin{cases} 0, & \text{si la observación no es censurada} \\ 1, & \text{si la observación es censurada} \end{cases} \tag{3.6}$$

De las ecuaciones (3.4) y (3.6) tenemos la siguiente función de densidad de la normal asimétrica con censura a la izquierda de la i-ésima observación que pertenece a la componente j, el cual se expresa de la siguente manera:

$$g_{ij}(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta}_j) = \begin{cases} f_{NA}(y_i | \mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j) & , \text{si } \mathbb{I}_i = 0 \\ F_{NA}\left(\frac{\kappa_i - \mu_{ij} - b\Delta_j}{\sigma_j}\right) & , \text{si } \mathbb{I}_i = 1, \end{cases}$$

donde $f_{NA}(\cdot)$ denota la densidad de distribución normal asimétrica y $F_{NA}(\cdot)$ representa la función de distribución acumulada de la distribución normal asimétrica con media 0, varianza 1 y parámetro de forma λ_j. La expresión anterior se puede escribir de la siguiente forma:

$$g_{ij}(y_i|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_j) \;=\; \left[F_{NA}\left(\frac{\kappa_i-\mu_{ij}-b\Delta_j}{\sigma_j}\right)\right]^{\mathbb{I}_i}\left[f_{NA}(y_i|\mu_{ij}+b\Delta_j,\sigma_j^2,\lambda_j)\right]^{1-\mathbb{I}_i},$$

donde $g_{ij}(\cdot)$ es la función de densidad de Y_i que pertenece a la componente j con censura a la izquierda. Por lo tanto, la función de densidad de probabilidad de Y_i dado en la ecuación (3.3) cuando existen datos censurados, es la siguiente:

$$f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j g_{ij}(y_i|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_j). \tag{3.7}$$

donde la ecuación (3.7) es el modelo de regresión lineal con censura a la izquierda basado en una mixtura finita de una distribución normal asimétrica, que será llamado el modelo MRL-CI-MF-NA.

3.2. Especificación del modelo

En esta sub-sección, según Dempster et al. (1977), consideramos el algoritmo de aproximación estocástica de Esperanza-Maximización (SAEM) para la estimación de máxima verosimilitud del modelo MRL-CI-MF-NA. Para explorar el algoritmo, presentaremos el modelo MRL-CI-MF-NA en un contexto de datos incompletos. Para mayor comodidad en la notación, consideremos el vector de datos completos $Y = (\boldsymbol{y},\ \boldsymbol{y}^0)$, donde $\mathbf{y}$ representa los datos observados e $\mathbf{y}^0$ representa las observaciones incompletas o censuradas. La función de log-verosimilitud del modelo MRL-CI-MF-NA (3.7) se expresa de la siguiente forma:

$$\ell_c(\boldsymbol{\theta}) = \prod_{i=1}^{n}\log\{\sum_{j=1}^{G} p_j\left[F_{NA}\left(\frac{\kappa_i-\mu_{ij}-b\Delta_j}{\sigma_j}\right)\right]^{\mathbb{I}_i}\left[f_{NA}(y_i|\mu_{ij}+b\Delta_j,\sigma_j^2,\lambda_j)\right]^{1-\mathbb{I}_i}\}.$$

Dado que la función de log-verosimilitud observada implica expresiones complejas e intratables, esta función $\ell_c(\boldsymbol{\theta})$ es muy difícil de maximizar de manera directa. Para superar este problema usaremos el algoritmo SAEM. Para ello, observe que, dada una muestra de tamaño n del modelo, el vector de respuestas censuradas $Y = (Y_1,...,Y_n)^\top$ se asumirá como un vector aleatorio latente (parcialmente no observable). De la representación jerárquica (3.5), obtenemos la función de verosimilitud completa $L_c(\boldsymbol{\theta})$,

$$\begin{aligned}
L_c(\boldsymbol{\theta}) \;&\propto\; \prod_{i=1}^{n}\prod_{j=1}^{G} f(y_i|T_i=t_i,Z_{ij}=1)^{z_{ij}} f(T_i|Z_{ij}=1)^{z_{ij}} f(\mathbf{Z}_i=z_i)^{z_{ij}} \\
&\propto\; \prod_{i=1}^{n}\prod_{j=1}^{G} f(y_i|T_i=t_i,Z_{ij}=1)^{z_{ij}}\ p_1^{z_{i1}}p_2^{z_{i2}}\ldots(1-p_1-\ldots p_{G-1})^{z_{ij}} \\
&\propto\; \prod_{i=1}^{n}\prod_{j=1}^{G}\left[f(y_i|T_i=t_i,Z_{ij}=1)^{z_{ij}}\ p_j^{z_{ij}}\right],
\end{aligned}$$

denotaremos a $L_{ic}(\boldsymbol{\theta})$ de la siguiente forma:

$$L_{ic}(\boldsymbol{\theta}) \propto \prod_{j=1}^{G} \left[f(y_i|T_i = t_i, Z_{ij} = 1)^{z_{ij}} \; p_j^{z_{ij}} \right].$$

La log-verosimilitud completa es dado por :

$$\begin{aligned} \ell_c(\boldsymbol{\theta}) &\propto \sum_{i=1}^{n} log\left[L_{ic}(\boldsymbol{\theta})\right] \\ &\propto \sum_{i=1}^{n} \left[\sum_{j=1}^{G} log(f(y_i|T_i = t_i, Z_{ij} = 1)^{z_{ij}} \; p_j^{z_{ij}}) \right]. \end{aligned} \tag{3.8}$$

de la ecuación (3.8) se obtendrá la verosimilitud para el i-ésimo elemento denotado como $\ell_{ic}(\boldsymbol{\theta})$:

$$\begin{aligned} \ell_{ic}(\boldsymbol{\theta}) &\propto \sum_{j=1}^{G} log(f(y_i|T_i = t_i, Z_{ij} = 1)^{z_{ij}} \; p_j^{z_{ij}}) \\ &\propto \sum_{j=1}^{G} \left[z_{ij} \log[f(y_i|T_i = t_i, Z_{ij} = 1)] + z_{ij} log(p_j) \right] \\ &\propto \sum_{j=1}^{G} z_{ij} \left[\log\left(\frac{1}{(2\pi\tau_j)^{1/2}} \exp\left(-\frac{(y_i - \mu_{ij} - \Delta_j t_i)^2}{2\tau_j} \right) \right) + \log(p_j) \right] \\ &\propto \sum_{j=1}^{G} z_{ij} \left[-\frac{1}{2} log(2\pi\tau_j) - \frac{(y_i - \mu_{ij} - \Delta_j t_i)^2}{2\tau_j} + log(p_j) \right]. \end{aligned}$$

Resultando:

$$\ell_{ic}(\boldsymbol{\theta}) \propto \sum_{j=1}^{G} z_{ij} log(p_j) - \frac{1}{2} \sum_{j=1}^{G} z_{ij} log(|\tau_j|) - \frac{1}{2} \sum_{j=1}^{G} \frac{z_{ij}}{\tau_j} (y_i - \mu_{ij} - \Delta_j t_i)^2. \tag{3.9}$$

Obteniéndose a partir de (3.8) y de (3.9), la función log-verosimilitud completa:

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{G} z_{ij} log(p_j) - \frac{1}{2} \sum_{j=1}^{G} z_{ij} log(|\tau_j|) - \frac{1}{2} \sum_{j=1}^{G} \frac{z_{ij}}{\tau_j} (y_i - \mu_{ij} - \Delta_j t_i)^2 \right].$$

Lo que es equivalente a:

$$\ell_c(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\left(log(p_j) - \frac{1}{2}log(|\tau_j|)\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}\frac{z_{ij}}{\tau_j}\left(y_i - \mu_{ij} - \Delta_j t_i\right)^2. \tag{3.10}$$

3.2.1. Algoritmo SAEM

A partir de la expresión dada en (3.10) y el paso de aproximación estocástica dada en la ecuación (2.4), se obtiene la actualización de la función $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ de la siguiente manera:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) \approx Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}) + \gamma_k\left[\frac{1}{m(k)}\sum_{l=1}^{m(k)} l_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{y}^{0(l,k)}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)})\right]. \tag{3.11}$$

La función $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)})$ se obtiene de la siguiente manera:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}) \quad = \quad E_{\boldsymbol{\theta}^{(k-1)}}\left[\ell_c(\boldsymbol{\theta})|\boldsymbol{\theta}^{(k-1)}\right].$$

El superíndice (k) indica la estimación del parámetro relacionado a la etapa k del algoritmo, donde:

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= E_{\boldsymbol{\theta}^{(k)}}\left[\sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\left(log(p_j)-\frac{1}{2}log(|\tau_j|)\right)\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&\quad -\frac{1}{2}E_{\boldsymbol{\theta}^{(k)}}\left[\sum_{i=1}^{n}\sum_{j=1}^{G}\frac{Z_{ij}}{\tau_j}\left(Y_i-\mu_{ij}-\Delta_j T_i\right)^2\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&= E_{\boldsymbol{\theta}^{(k)}}\left[\sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\left(log(p_j)-\frac{1}{2}log(|\tau_j|)\right)\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}E_{\boldsymbol{\theta}^{(k)}}\left[\frac{Z_{ij}}{\tau_j}\left((Y_i-\Delta_j T_i)^2-2(Y_i-\Delta_j T_i)\mu_{ij}+\mu_{ij}^2\right)\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&= E_{\boldsymbol{\theta}^{(k)}}\left[\sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\left(log(p_j)-\frac{1}{2}log(|\tau_j|)\right)\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}E_{\boldsymbol{\theta}^{(k)}}\left[\frac{Z_{ij}}{\tau_j}\left(Y_i^2-2\Delta_j T_i Y_i+\Delta_j^2 T_i^2-2\mu_{ij}Y_i+2\mu_{ij}\Delta_j T_i+\mu_i^2\right)\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{G}\left(log(p_j)-\frac{1}{2}log(|\tau_j|)\right)E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}\,\middle|\,\boldsymbol{\theta}^{(k)}\right]-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}E_{\boldsymbol{\theta}^{(k)}}\left[\frac{Z_{ij}Y_i^2}{\tau_j}\right. \\
&\quad \left.-\frac{2\Delta_j}{\tau_j}Z_{ij}T_iY_i+\frac{\Delta_j^2}{\tau_j}Z_{ij}T_i^2-\frac{2\mu_{ij}}{\tau_j}Z_{ij}Y_i+\frac{2\mu_{ij}\Delta_j}{\tau_j}Z_{ij}T_i+\frac{\mu_{ij}^2}{\tau_j}Z_{ij}\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{G}\left(log(p_j)-\frac{1}{2}log(|\tau_j|)-\frac{\mu_{ij}^2}{2\tau_j}\right)E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}\,\middle|\,\boldsymbol{\theta}^{(k)}\right] \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}\frac{1}{\tau_j}\left(E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}Y_i^2\,\middle|\,\boldsymbol{\theta}^{(k)}\right]-2\Delta_j E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}T_iY_i\,\middle|\,\boldsymbol{\theta}^{(k)}\right]\right. \\
&\quad \left.+\Delta_j^2 E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}T_i^2\,\middle|\,\boldsymbol{\theta}^{(k)}\right]-2\mu_{ij}E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}Y_i\,\middle|\,\boldsymbol{\theta}^{(k)}\right]+2\mu_{ij}\Delta_j E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}T_i\,\middle|\,\boldsymbol{\theta}^{(k)}\right]\right),
\end{aligned}
$$

Definamos de forma similar a Benites et al. (2019) los términos:

$$
\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}\,\middle|\,\boldsymbol{\theta}^{(k-1)}\right] \tag{3.12}
$$

$$
S_{1ij}(\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{k}}\left[Z_{ij}T_i\,\middle|\,\boldsymbol{\theta}^{(k-1)}\right] \tag{3.13}
$$

$$S_{2ij}(\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^k}\left[Z_{ij}T_i^2 \,\middle|\, \boldsymbol{\theta}^{(k-1)}\right], \tag{3.14}$$

y

$$\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = E\left[Z_{ij}T_i^r Y_i^s \,\middle|\, \boldsymbol{\theta}^{(k-1)}\right] \tag{3.15}$$

para $r = 0, 1$ y $s = 1, 2$.
Usando propiedades de esperanza condicional, se obtiene:

$$\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) = \frac{p_j^k g_{ij}\left(y_i \,\middle|\, x_i, \gamma^k, \boldsymbol{\theta}_j^{(k)}\right)}{\sum p_j^k g_{ij}\left(y_i \,\middle|\, x_i, \gamma^k, \boldsymbol{\theta}_j^{(k)}\right)}$$

$$S_{1ij}(\boldsymbol{\theta}^{(k)}) = \mathcal{Z}_{ij}\left(\widehat{\mu}_{Tij} + \widehat{M}_{T_j}\widehat{\tau}_{1ij}\right) \tag{3.16}$$

$$S_{2ij}(\boldsymbol{\theta}^{(k)}) = \mathcal{Z}_{ij}\left(\widehat{\mu}_{Tij}^2 + \widehat{M}_{T_j}^2 + \widehat{M}_{T_j}\widehat{\mu}_{T_{ij}}\widehat{\tau}_{1ij}\right), \tag{3.17}$$

donde:

$$\widehat{\tau}_{1ij} = E\left[W_{\phi_1}\left(\frac{\widehat{\mu}_{T_{ij}}}{\widehat{M}_{T_j}}\right) \,\middle|\, \boldsymbol{\theta}^{(k-1)}, y_i, Z_{ij} = 1\right] \tag{3.18}$$

$$W_\Phi(\alpha) = \frac{\phi(\alpha)}{\Phi(\alpha)},$$

$$\widehat{M}_{T_j}^2 = \frac{\tau_j}{\tau_j + \Delta_j^2},$$

$$\widehat{\mu}_{T_{ij}} = b + \frac{\Delta_j}{\tau_j + \Delta_j^2}\left(y_i - \mu_{ij} - \Delta_j b\right),$$

para $i = 1,\ 2, ...,\ n$ y $j = 1,\ 2, ...,\ g$.
Obteniéndose:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = {} & \sum_{i=1}^{n}\sum_{j=1}^{G}\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})\left(\log(p_j) - \frac{1}{2}log(|\tau_j|) - \frac{\mu_{ij}^2}{2\tau_j}\right) \\ & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G}\frac{1}{\tau_j}\left(\mathcal{E}_{02i}(\boldsymbol{\theta}^{(k)}) - 2\Delta_j\mathcal{E}_{11i}(\boldsymbol{\theta}^{(k)}) + \Delta_j^2 S_{2ij}(\boldsymbol{\theta}^{(k)})\right. \\ & \left. -2\mu_{ij}\mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)}) + 2\mu_{ij}\Delta_j S_{1ij}(\boldsymbol{\theta}^{(k)})\right). \end{aligned} \tag{3.19}$$

Así, para cada paso se calcula $\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})$, $S_{1ij}(\boldsymbol{\theta}^{(k)})$, $S_{2ij}(\boldsymbol{\theta}^{(k)}), \mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)})$ y $\widehat{\tau}_{1ij}$ para la distribución NA y considerando dos situaciones:

1. Para una observación no censurada i:

En este caso, se tiene que $\mathbb{I}_i = 0$ así :

$$Y_i \sim \sum_{j=1}^{G} p_j \phi_{NA}\left(\beta_0 + \boldsymbol{x_i}^\top \boldsymbol{\beta} + \mu_j - \sqrt{\frac{2}{\pi}}\Delta_j, \sigma_j^2, \lambda_j\right)$$

y por lo tanto:

$$\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = y_i^s E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}T^r \,\middle|\, \boldsymbol{\theta}^{(k-1)}\right] = y_i^s S_{rij}(\boldsymbol{\theta}^{(k)}),\ r = 1, 2$$

donde $S_{rij}(\boldsymbol{\theta}^{(k)})$ puede ser obtenido usando la ecuación (3.16) y (3.17) y a partir de los resultados dado por Basso et al. (2010).

2. Para una observación censurada i:
   En este caso, se tiene que $\mathbb{I}_i = 1$, es decir $Y_i \leq k_i$, por lo tanto:

$$\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij}T_i^r Y_i^s \,\middle|\, \mathbb{I}_i, Y_i \leq k_i, \boldsymbol{\theta}^{(k-1)}\right], \tag{3.20}$$

   para $r = 0, 1$ y $s = 1, 2$.

Como esta esperanza condicional no tiene forma conocida, se necesita introducir dos pasos intermedios en orden para reemplazar el paso E por una aproximación estocástica, usando data simulada. Así, la iteración k consiste en los siguientes pasos.

Paso S (Sampling)

Sea $Y^c = (Y_1^c, Y_2^c, ..., Y_n^c)$ el vector de n^c casos censurados, donde Y_i^c es generado a partir de:

$$\sum_{j=1}^{G} p_j TSN(\boldsymbol{x_i}^\top \boldsymbol{\beta} - \sqrt{\frac{2}{\pi}}k_i\Delta_j, \sigma_j^2, \lambda_j, < -\infty, k_i]),$$

para $i = 1, 2, ..., n^c$. Así, el nuevo vector de observaciones $Y^{(l,k)} = (Y_{i1}^{(l,k)}, ..., Y_{in^c}^{(l,k)}, Y_{n_i^c+1}, ..., Y_n)$ es una muestra generada por los n^c casos censurados y los valores observados (casos no censurados) para $l = 1, 2, ..., m$.

Paso AE (Aproximación estocástica)

Se tiene la secuencia $y^{(l,k)}$ en la iteración k_i, considerando las ecuaciones (3.16), (3.17) y los resultados presentados por Basso et al. (2010), se reemplaza la esperanza condicional en (3.20) para la siguiente aproximación estocástica:

$$\mathcal{E}_{rsi}(\boldsymbol{\theta}^{*(k)}) = \mathcal{E}_{rsi}(\boldsymbol{\theta}^{*(k-1)}) + \gamma_j\left(\frac{1}{m}\sum_{l=1}^{m} E\left[Z_{ij}T_i^r Y_i^{s(l,k)} \,\middle|\, \mathbb{I}_i, \boldsymbol{\theta}^{*(k)}\right] - \mathcal{E}_{rsi}(\boldsymbol{\theta}^{*(j-1)})\right),$$

para $r = 0, 1$ y $s = 1, 2$.

Según Thalita (2016), se requiere un tamaño de muestra Montecarlo $m = 20$.

Paso M

Maximizar $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ con respecto a $\boldsymbol{\theta}$ obteniendo $\boldsymbol{\theta}^{(k+1)}$, donde γ_k es una sucesión decreciente de números positivos tales que:

$$\sum_{k=1}^{\infty}\gamma_k = \infty \text{ y } \sum_{k=1}^{\infty}\gamma_k^2 < \infty,$$

como lo presentado por Kuhn y Lavielle (2004). Según lo sugerido por Galarza et al. (2017), se usa la siguiente opción del parámetro suavizado:

$$\gamma_k = \begin{cases} 1, & \text{si } 1 \leq k \leq \kappa_i S \\ \dfrac{1}{k-\kappa_i S}, & \text{si } \kappa_i S + 1 \leq k \leq S, \end{cases}$$

donde S, es el número máximo de iteraciones, y κ_i es el punto de corte $(0 \leq \kappa_i \leq 1)$ que determina el porcentaje de iteraciones iniciales sin memoria.
De esta manera, se obtiene las siguientes expresiones:

$$\begin{aligned}
p_j^{(k+1)} &= \frac{\sum_{i=1}^{n}\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})}{n}, \\
\vartheta_j^{(k+1)} &= \frac{\sum_{i=1}^{n}\left[\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})(y_i - \mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}}) - \widehat{\Delta}_j^{(k)}S_{1ij}(\boldsymbol{\theta}^{(k)})\right]}{\sum_{i=1}^{n}\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})}, \\
\widehat{\boldsymbol{\beta}}_j^{(k+1)} &= \left(\sum_{i=1}^{n}\sum_{j=1}^{G}\frac{\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})\mathbf{x}_i\mathbf{x}_i^{\top}}{\widehat{\tau}_j^{(k)}}\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{G}\frac{1}{\widehat{\tau}_j^{(k)}}\left[\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})(y_i - \vartheta_j^{(k+1)}) - \widehat{\Delta}_j^{(k)}S_{1ij}(\boldsymbol{\theta}^{(k)})\right]\mathbf{x}_i, \\
\widehat{\Delta}_j^{(k+1)} &= \frac{\sum_{i=n}^{n}(y_i - \mu_{ij}^{(k)})S_{1ij}(\boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n}S_{2ij}(\boldsymbol{\theta}^{(k)})}, \\
\widehat{\tau}_j^{(k+1)} &= \frac{\sum_{i=1}^{n}\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})(y_i - \mu_{ij}^{(k+1)})^2 - 2(y_i - \mu_{ij}^{(k+1)})\widehat{\Delta}_j^{(k+1)}S_{1ij}(\boldsymbol{\theta}^{(k)}) + \widehat{\Delta}_j^{2(k+1)}S_{2ij}(\boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n}\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})}.
\end{aligned}$$

Continuando las ideas de Bartolucci y Scaccia (2005), se tiene que:

$$\widehat{\mu}_j^{(k+1)} = \widehat{\vartheta}_j^{(k+1)} - \widehat{\boldsymbol{\beta}}_0^{(k+1)},\ \widehat{\boldsymbol{\beta}}_0^{(k+1)} = \sum_{j=1}^{G}\widehat{p}_j^{(k+1)}\widehat{\vartheta}_j^{(k+1)}.$$

respectivamente, para $j = 1, ..., g$. Este proceso es iterativo hasta que se satisface un criterio de parada adecuado.

3.2.2. Especificación de los valores iniciales

Es bien conocido que, los modelos de mixtura pueden proporcionar una función de verosimilitud multimodal, en ese sentido, el método de estimación de máxima verosimilitud a través del algoritmo EM puede no dar una solución máxima global si los valores iniciales están lejos de los valores reales de los parámetros. El procedimiento de obtención de los valores iniciales se resume a continuación:

- El valor inicial $\boldsymbol{\beta}^{(0)}$ resulta de aplicar un modelo de regresión lineal al conjunto de datos completos (no censurados) en el software R.
- Ejecutar el algoritmo de agrupamiento k-medias (Benites et al., 2019) o k-medoids (Kaufman y Rousseeuw, 1990) para inicializar con respecto a un centro de clúster elegido al azar.
- Inicializar el indicador $\widehat{Z}_j^{(0)} = \{\widehat{z}_{ij}^{(0)}\}_{i=1}^{G}$ de acuerdo con el algoritmo de agrupamiento escogido en el item anterior.
- Inicializar las proporciones de mixtura como sigue: $\widehat{p}_j^{(0)} = (1/n)\sum_{i=1}^{n} \widehat{z}_{ij}^{(0)}$. Esto, es el valor inicial para p_j.
- Para cada grupo j, calcular los valores iniciales $\mu_j^{(0)}$, $(\sigma_j^2)^{(0)}$ usando el método de momento.

3.2.3. Regla de parada

La evaluación de la convergencia del algoritmo EM para el modelo MRL-CI-MF-NA, se realiza a partir del criterio de parada basado en la aceleración de Aitken;

$$|\ell^{(k+1)} - \ell_{\infty}^{(k+1)}| < \varepsilon,$$

donde, $\ell^{(k+1)}$ es la log-verosimilitud observada en $\boldsymbol{\theta}^{(k)}$ y ε es la tolerancia deseada. Para los análisis numéricos que serán presentados en el estudio de simulación y las aplicaciones de esta tesis, se considera un $\varepsilon = 10^{-6}$. Asumiendo convergencia de los estimadores de máxima verosimilitud (EMV) $\widehat{\boldsymbol{\theta}}$, también $\ell_{\infty}^{(k+1)}$ es estimado asintóticamente de la log-verosimilitud en la iteración $k+1$ (McLachlan y Krishnan, 2008, Chap. 4.9), donde:

$$\ell_{\infty}^{(k+1)} = \ell^{(k)} + \frac{\ell^{(k+1)} - \ell^{(k)}}{1 - c^{(k)}},$$

con $c^{(k)}$ denotando el acelerador de Aitke en la iteración k, dado por:

$$c^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}}.$$

El procedimiento mostrado arriba es también aplicable para un caso más simple, es decir considerando $Z_{ij} = 1$, $(G = 1)$.

3.2.4. Aproximación del error estándar

Una manera de obtener los errores estándar de las estimaciones de MV de los parámetros de un modelo de mixtura es mediante la aproximación de la matriz de covarianza asintótica de $\widehat{\boldsymbol{\theta}}$ por la inversa de la matriz de información observada. Sea $\mathbf{I}_o(\boldsymbol{\theta}) = -\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ la matriz de información observada, donde $\ell(\boldsymbol{\theta})$ es la función de log-verosimilitud completa, considerando $\boldsymbol{\theta}_j = (p_j,\ \sigma_j^2,\ \mu_j,\ \lambda_j)^\top$. En esta tesis usaremos el método alternativo sugerido por Basford et al. (1997), que consiste en aproximar la inversa de la matriz de covarianza por:

$$\mathbf{I}_o(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \widehat{\mathbf{s}}_i \widehat{\mathbf{s}}_i^\top, \quad \text{donde} \quad \widehat{\mathbf{s}}_i = \mathrm{E}\left[\frac{\partial \ell_{ic}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \;\middle|\; \mathbf{y}, \boldsymbol{\theta} \right], \tag{3.21}$$

donde, $\ell_{ic}(\boldsymbol{\theta})$ es dado por (3.10) y

$$\widehat{\mathbf{s}}_i \;=\; (\widehat{s}_{i,\boldsymbol{\beta}}, \widehat{s}_{i,p_1}, \ldots, \widehat{s}_{i,p_{G-1}}, \widehat{s}_{i,\sigma_1^2}, \ldots, \widehat{s}_{i,\sigma_G^2}, \widehat{s}_{i,\mu_1}, \ldots, \widehat{s}_{i,\mu_G}, \widehat{s}_{i,\lambda_1}, \ldots, \widehat{s}_{i,\lambda_G})^\top.$$

Entonces:

- $\widehat{s}_{i,\boldsymbol{\beta}} = E\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \;\middle|\; \mathbf{y}, \boldsymbol{\theta} \right]$

Resultando a partir de (3.12) y (3.13):

$$\widehat{s}_{i,\boldsymbol{\beta}} = -\sum_{j=1}^{G} \frac{1}{\tau_j} \left(\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i^\top \boldsymbol{\beta} + \mu_{ij})\mathbf{x}_i - \mathbf{x}_i^\top \mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)}) + \mathbf{x}_i^\top \Delta_j S_{1ij}(\boldsymbol{\theta}^{(k)}) \right)$$

- $\widehat{s}_{i,p_j} = E\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial p_j} \;\middle|\; \mathbf{y}, \boldsymbol{\theta} \right]$

Resultando a partir de (3.12):

$$\widehat{s}_{i,p_j} = \frac{1}{p_j}\mathcal{Z}_{ij} - \frac{1}{p_G}\mathcal{Z}_{iG}$$

- $\widehat{s}_{i,\sigma_j^2} = E\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2 j} \;\middle|\; \mathbf{y}, \boldsymbol{\theta} \right]$

Resultando a partir de (3.12), (3.13) y (3.14) :

$$\widehat{s}_{i,\sigma^2 j} = \frac{-1}{2\tau_j}\left[\mathcal{Z}_{ij}(1+\lambda_j^2)^{-1} - \mathcal{Z}_{ij}\sigma_j^{-2}\mu_{ij}^2 \right] + \left(\frac{1}{2\tau_j\sigma_j^2} \right)\left[\mathcal{E}_{02i}(\boldsymbol{\theta}^{(k)}) + 2\mu_{ij}\mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)}) \right] -$$

$$\left(\frac{\delta_j}{2\tau_j\sigma_j} \right)\left[\mathcal{E}_{11i}(\boldsymbol{\theta}^{(k)}) - \mu_{ij}S_{1ij}(\boldsymbol{\theta}^{(k)}) \right]$$

- $\widehat{s}_{i,\mu_j} = E\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu_j} \;\middle|\; \mathbf{y}, \boldsymbol{\theta} \right]$

Resultando a partir de (3.15) y (3.14):

$$\widehat{s}_{i,\mu_j} = \frac{-1}{\tau_j}\left[\mathcal{Z}_{ij}(\boldsymbol{x_i}^\top \boldsymbol{\beta} + \boldsymbol{\beta}_0 + \mu_j) - \mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)}) + \sigma_j\delta_j S_{ij}(\boldsymbol{\theta}^{(k)}) \right]$$

- $\widehat{s}_{i,\lambda_j} = E\left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \lambda_j} \,\middle|\, \mathbf{y}, \boldsymbol{\theta}\right]$

Resultando a partir de (3.15), (3.13), (3.14), (3.16) y (3.17):

$$\widehat{s}_{i,\lambda_j} = \frac{\lambda_j}{1+\lambda_j^2}\mathcal{Z}_{ij}\left(1-\frac{\mu_{ij}^2}{\tau_j}\right) - \frac{\lambda_j}{\sigma_j^2}\left[\mathcal{E}_{02i}(\boldsymbol{\theta}^{(k)}) + \sigma_j^2 S_{2ij}(\boldsymbol{\theta}^{(k)}) - 2\mu_{ij}\mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)})\right] +$$

$$\frac{\delta_j(1+2\lambda_j^2)}{\sigma_j\lambda_j}\left[2\mathcal{E}_{11i}(\boldsymbol{\theta}^{(k)}) - \mu_{ij}S_{1ij}(\boldsymbol{\theta}^{(k)})\right]$$

3.2.5. Selección de modelos

El criterio utilizado en esta tesis para seleccionar el mejor modelo MRL-CI-MF-NA, con diferente número de componentes se realiza con el criterio de Akaike (1974), lo cual es denotado por AIC. De esta manera se define como:

$$\text{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2\rho,$$

y el criterio de información Bayesiana (Schwarz, 1978) (BIC).

$$\text{y} \quad \text{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + \rho\log(n),$$

donde, $\ell(\boldsymbol{\theta})$ es la actual log-verosimilitud, ρ es el número de parámetros libres que serán estimados en el modelo, y n es el tamaño de muestra.

Capítulo 4

Estudio de simulación

En el presente capítulo, se desarrolló un estudio de simulación con la finalidad de examinar el desempeño del modelo propuesto en el Capítulo 3, mediante el algoritmo SAEM, el cual recupera información de los datos incompletos mediante aproximaciones estocásticas con mejor precisión de las estimaciones y consistencia de los errores estándar. Para este propósito el modelo fue implementado en el programa R y se realizó un estudio de simulación de Monte Carlo considerándose dos componentes ($G = 2$) con datos censurados por la izquierda.

4.1. Consideraciones para la simulación

La Figura 4.1 muestra la gráfica de la curva de la distribución de los errores que se emplearán en el estudio de simulación.

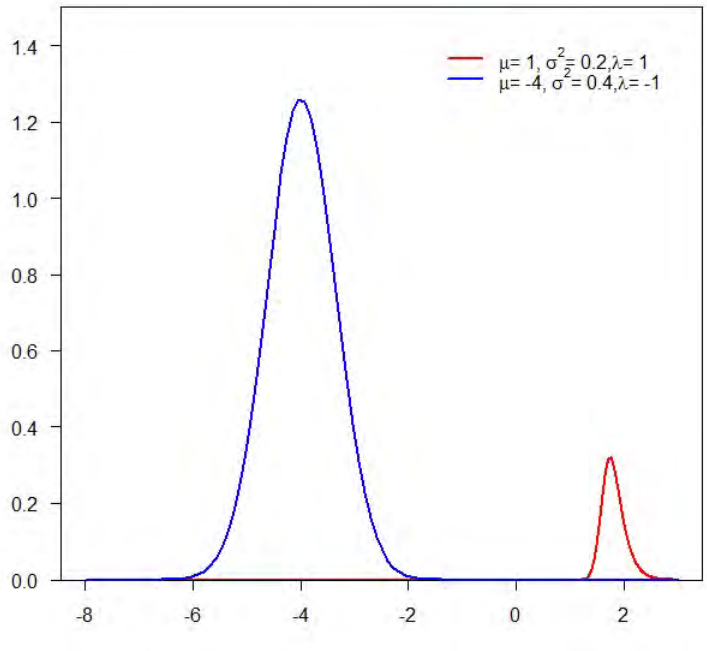


Figura 4.1: Funciones de densidad de normales asimétricas que se emplearán en el estudio de simulación (La curva azul representa la función de densidad de la NA con $\mu = -4, \sigma^2 = 0.4, \lambda = -1$ y la curva roja representa la función de densidad de la NA con $\mu = 1, \sigma^2 = 0.2, \lambda = 1$)

Se generará 1000 simulaciones para cada tamaño de muestra $n = 150,\ 250,\ 500,\ 1000$ y 2000 para las covariables x_i y para los errores ε_i. Los valores de x_i, $i = 1, \ldots, n$, fueron generados en forma independiente de una distribución uniforme $U(2, 20)$ y los errores ε_i con

dos componentes ($G = 2$), fueron simulados a partir de la Figura 4.1. Estos valores serán mantenidos constantes en el desarrollo de la simulación. Posteriormente, a partir de los datos antes mencionados se generará la respuesta y_i, según 4.1. Como se desea obtener valores censurados de y_i, se define un punto de corte c_i con la finalidad que los valores de y_i que son menores a c_i, sean los valores censurados a la izquierda (y_i incompletos). Cabe resaltar que, el punto de corte c_i corresponde al percentil, que está asociado al porcentaje de censura, de todos los y_i generados.

La función de densidad de Y_i se presenta de la siguiente forma:

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = p_1 g_{i1}(y_i|\mathbf{x}_i, \boldsymbol{\beta} + \mu_1 + b\Delta_1, \sigma_1^2, \lambda_1) + p_2 g_{i2}(y_i|\mathbf{x}_i, \boldsymbol{\beta} + \mu_2 + b\Delta_2, \sigma_2^2, \lambda_2), \quad (4.1)$$

donde $\boldsymbol{\beta}^\top = (\beta_0,\ \beta_1,\ \beta_2) = (1,\ 2,\ 3)$, $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$, $\mu_1 = 1$, $\mu_2 = -4$, $\lambda_1 = 1$, $\lambda_2 = -1$, $p_1 = 0.8$ y $p_2 = 0.2$ y $\mathbf{x}_i^\top = (1,\ x_1,\ x_2)$. Además, se ha considerado tres escenarios con diferentes niveles de censura a la izquierda de $p = 8\,\%$, $p = 20\,\%$ y $p = 35\,\%$. Para obtener el nivel de censura de $p = 8\,\%$, los valores de y_i que son menores al punto de corte $c_i = 26.92712$, serán los valores censurados, para obtener el nivel de censura de $p = 20\,\%$, el punto de corte será $c_i = 38.35708$. Para obtener el nivel de censura de $p = 35\,\%$, el punto de corte será $c_i = 47.22542$. Adicionalmente, consideramos $m = 20$, $c = 0.3$ y $S = 400$ para la implementación del SAEM. La regla de convergencia utilizada en esta tesis, para el estudio de simulación y aplicación es basada en el acelerador de Aitken como fue mencionado en la subsección 3.2.3

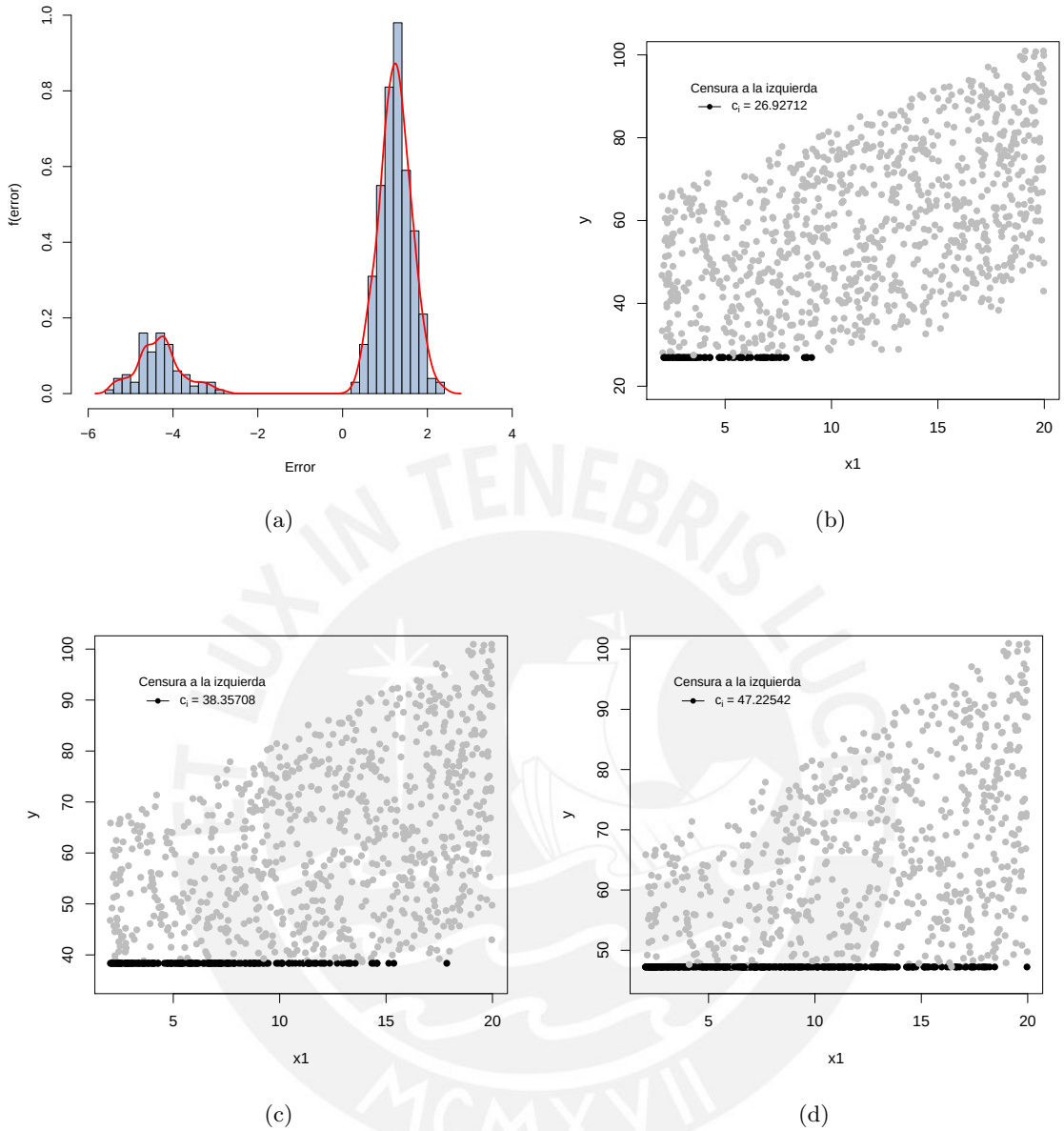


Figura 4.2: Densidad de las componentes de los errores con distribución Normal Asimétrica (NA) (a), gráficos de dispersión de la covariable x_1 y la variable respuesta Y censurada (b), (c) y (d), para diferentes niveles de censura tales como 8 %, 20 % y 35 % respectivamente.

4.2. Criterios para la evaluación de la simulación

Los criterios para validar la precisión de las estimaciones a través de las simulaciones realizadas en el modelo MRL-CI-MF-NA serán el de la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE) como lo muestra Willmott (2005). Para el cálculo de la precisión a través del RMSE se utilizará la siguiente fórmula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{\theta_i} - \theta_i)^2}{N}}. \tag{4.2}$$

El cálculo del MAE, nos da el promedio de las diferencias entre los valores estimados y observados, la fórmula es la siguiente:

$$MAE = \sum_{i=1}^{N} \frac{|\widehat{\theta_i} - \theta_i|}{N}, \tag{4.3}$$

donde:

N representa la cantidad de réplicas que se darán en los estudios de simulación

$\widehat{\theta_i}$ representa el valor estimado para la celda i-ésima

θ_i representa el valor observado para la celda i-ésima

k representa los valores simulados en la k-ésima iteración

4.3. Propiedad de consistencia para los estimadores de la simulación

A partir de los estudios de simulación, se realizó una evaluación de las propiedades asintóticas de los estimadores por máxima verosimilitud a partir de diferentes tamaños de muestra $n = 150,\ 250,\ 500,\ 1000$ y 2000, diferentes niveles de censura (8 %, 20 % y 35 %) para 1000 réplicas, y con muestras de tamaño 400.

En la presente investigación se utiliza el mismo criterio de Garay et al. (2017) para la evaluación de las estimaciones mediante el algoritmo SAEM por medio del Sesgo, así como el Error cuadrático medio (MSE).

Para el cálculo del Sesgo se considera la fórmula siguiente:

$$Sesgo = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\theta_i} - \theta_i\right). \tag{4.4}$$

Para el cálculo del MSE considera la fórmula siguiente:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\theta_i} - \theta_i\right)^2, \tag{4.5}$$

donde:

$\widehat{\theta_i^k}$ representa el valor estimado para la celda i-ésima, para la muestra de tamaño n

4.4. Resultados

- Los Cuadros 4.1, 4.2, 4.3 muestran las estimaciones de los parámetros, el sesgo y los errores estándar de la simulación considerando un 8 %, 20 % y 35 % de nivel de censura, respectivamente, para cada tamaño de muestra $n = 150,\ 250,\ 500,\ 1000$ y 2000. Estos valores fueron calculados usando el método Bootstrap para 1000 réplicas, con muestras con reemplazo de tamaño 400 y los valores iniciales propuestos en la sección 4.1. En todos los casos se observa que los errores estándar son menores a medida que el tamaño de muestra aumenta y las estimaciones se encuentran cercanas a los valores iniciales.

- Analizando las Figuras 4.3, 4.4, 4.5 se aprecia que para los niveles de censura del 8 %, 20 % y 35 %, la estimación de los parámetros evidencian menos variabilidad y menos sesgo cuando n aumenta. Es decir, se refleja mejor precisión en el comportamiento de la estimación al presentar menores valores atípicos y menor variabilidad. Además, es importante resaltar que, a mayor proporción de censura, hay mayor presencia de valores atípicos; sin embargo, el algoritmo SAEM recupera la información a mayor cantidad de muestra, corroborándose que el modelo propuesto presenta buenas propiedades asintóticas.

- En las Figuras 4.6 y 4.7 se muestran los resultados resumidos mediante el promedio del MAE y RMSE para diferentes niveles de censura tales como 8 %, 20 % y 35 %, para diferentes tamaños de muestra $n = 150,\ 250,\ 500,\ 1000$ y 2000, para β_0, β_1, β_2, μ_1, μ_2, σ_1^2, σ_2^2, λ_1, λ_2 y p_1. Estos resultados evidencian que a mayor tamaño de muestra, el MAE y RMSE disminuyen para cualquier nivel de censura. En el anexo A.1 aparecen los resultados numéricos con mayor detalle.

- El Cuadro A.1 muestra la media, desviación estándar y los percentiles 25, 50 y 97.5 de los parámetros estimados para cada tamaño de muestra, considerando un 8 %, 20 % y 35 % de nivel de censura, con lo cual se evidencia que los resultados se encuentran muy cercanos a los valores teóricos.

- El Cuadro A.2 muestra las diferentes medidas de precisión consideradas para esta tesis, como la raíz del error cuadrático medio (RMSE), error absoluto medio (MAE) y el error cuadrático medio (MSE) del modelo MRL-CI-MF-NA para un 8 %, 20 % y 35 % de nivel de censura respectivamente para $n = 150,\ 250,\ 500,\ 1000$ y 2000 muestras. Los resultados de las medidas de precisión evidencia una tendencia decreciente a medida que aumenta el tamaño de muestra.

- Los Cuadros A.3, A.4 y A.5 corresponden a los estadísticos (Media, desviación estándar, percentiles 25, 50 y 97.5) de los residuales del modelo MRL-CI-MF-NA para 1000 réplicas, las cuales evidencian que las estimaciones de los parámetros son recuperados de forma satisfactoria a medida que el tamaño de muestra incrementa. Estos resultados también son corroborados en A.6, A.7 y A.8, las cuales muestran la precisión del modelo.

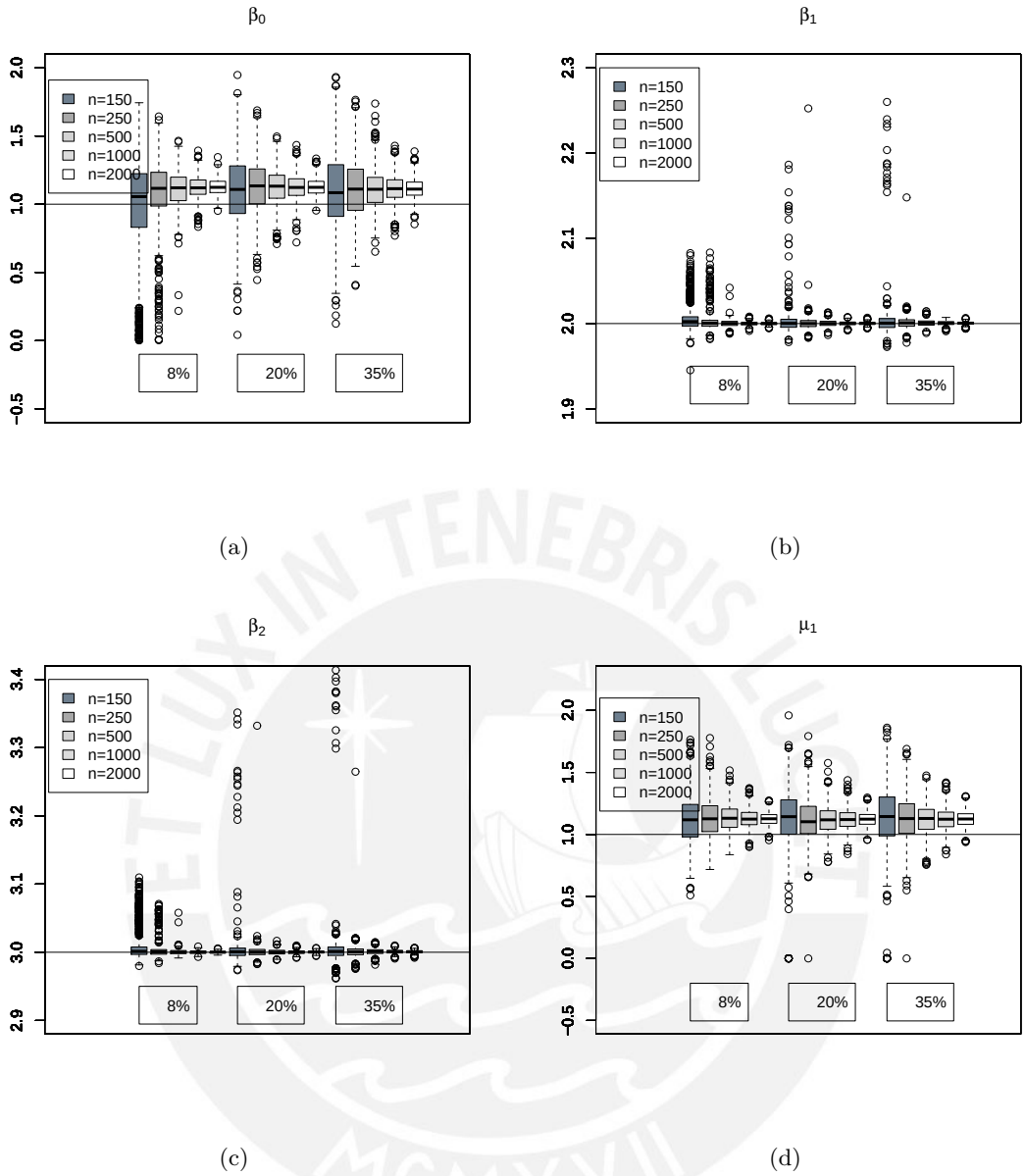


Figura 4.3: Gráfico de cajas de los parámetros estimados en las componentes con errores de distribución Normal Asimétrica (NA). (a) Gráfico de cajas para β_0, (b) Gráfico de cajas para β_1, (c) Gráfico de cajas para β_2 y (d) Gráfico de cajas de μ_1 para diferentes niveles de censura tales como 8 %, 20 % y 35 % y diferentes tamaños de muestra ($n = 150,\ 250,\ 500,\ 1000$ y 2000).

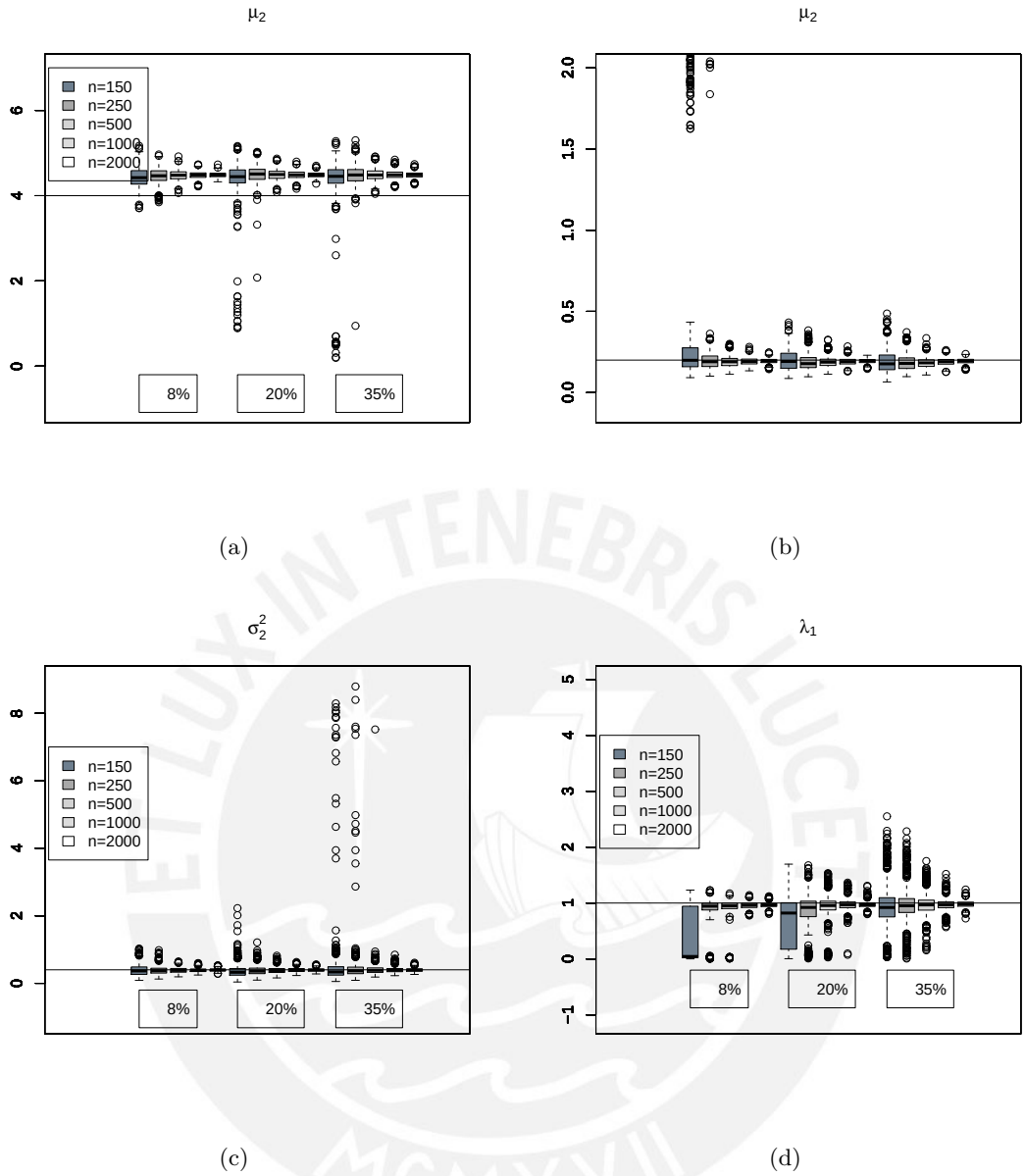
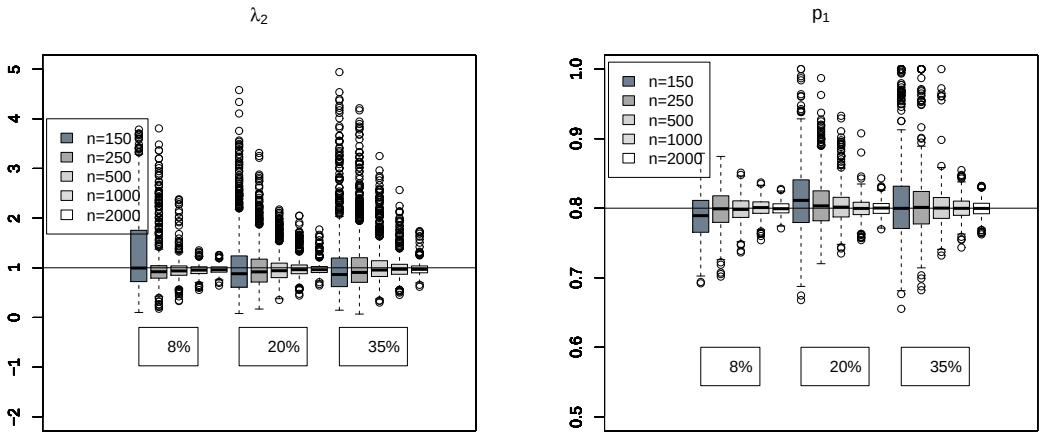


Figura 4.4: Gráfico de cajas de los parámetros estimados en las componentes con errores de distribución Normal Asimétrica (NA). (a) Gráfico de cajas para μ_2, (b) Gráfico de cajas para σ_1^2, (c) Gráfico de cajas para σ_2^2 y (d) Gráfico de cajas para λ_1 para diferentes niveles de censura tales como 8 %, 20 % y 35 % y diferentes tamaños de muestra ($n = 150,\ 250,\ 500,\ 1000$ y 2000).



(a) (b)

Figura 4.5: Gráfico de cajas de los parámetros estimados en las componentes con errores de distribución Normal Asimétrica (NA). (a) Gráfico de cajas para λ_2, (b) Gráfico de cajas para p_1, para diferentes niveles de censura tales como 8 %, 20 % y 35 % y diferentes tamaños de muestra ($n = 150,\ 250,\ 500,\ 1000$ y 2000).

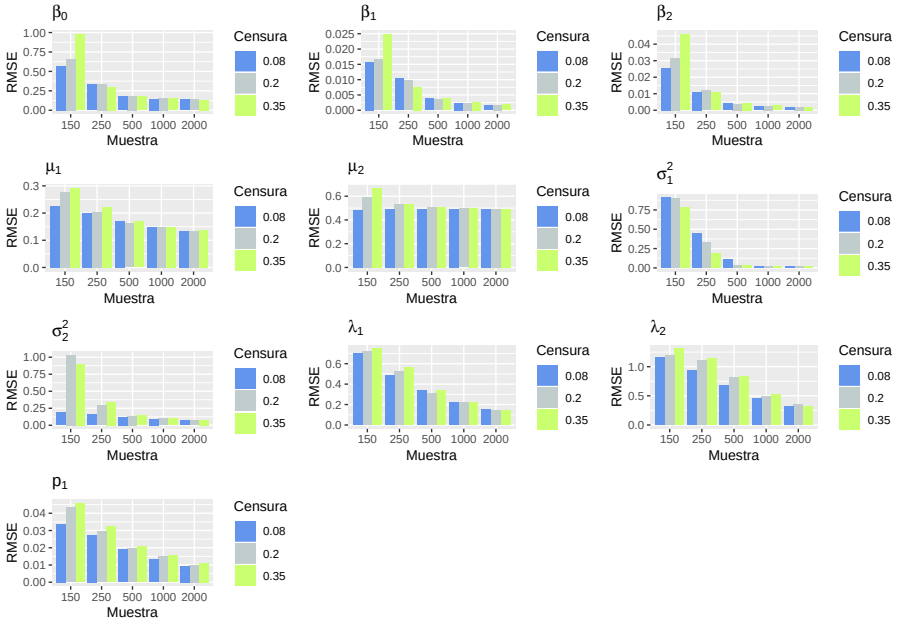


Figura 4.6: RMSE para diferentes niveles de censura tales como 8 %, 20 % y 35 % y diferentes tamaños de muestra ($n = 150,\ 250,\ 500,\ 1000$ y 2000) para β_0, β_1, β_2, μ_1, μ_2, σ_1^2, σ_2^2, λ_1, λ_2 y p_1.

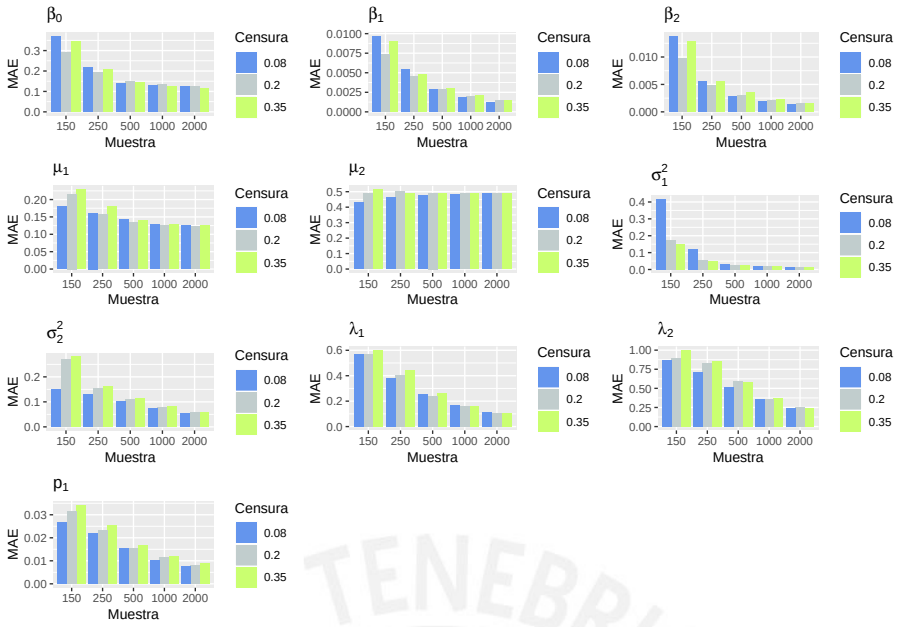


Figura 4.7: MAE para diferentes niveles de censura tales como 8 %, 20 % y 35 % y diferentes tamaños de muestra ($n = 150,\ 250,\ 500,\ 1000$ y 2000) para β_0, β_1, β_2, μ_1, μ_2, σ_1^2, σ_2^2, λ_1, λ_2 y p_1.

n=150						n=250				
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	0.8898	-0.0098	0.0173	0.0176	[0.8559, 0.9237]	1.0689	-0.0049	0.0103	0.0103	[1.0487, 1.0891]
β_1	2.0057	-0.0000	0.0005	0.0005	[2.0047, 2.0067]	2.0019	-0.0003	0.0003	0.0003	[2.0013, 2.0025]
β_2	3.0096	-0.0002	0.0008	0.0008	[3.0080, 3.0112]	3.0020	0.0001	0.0003	0.0003	[3.0014, 3.0026]
μ_1	1.1092	0.0077	0.0058	0.0061	[1.0978, 1.1206]	1.1355	-0.0078	0.0049	0.0049	[1.1259, 1.1451]
μ_2	-4.4314	0.0053	0.0072	0.0073	[-4.4455, -4.4173]	-4.4695	0.0073	0.0056	0.0056	[-4.4805, -4.4585]
σ_1^2	0.5746	-0.0005	0.0258	0.0262	[0.5240, 0.6252]	0.3008	-0.0196	0.0137	0.0138	[0.2739, 0.3277]
σ_2^2	0.3855	0.0069	0.0059	0.0060	[0.3739, 0.3971]	0.4020	0.0013	0.0051	0.0050	[0.392, 0.4120]
λ_1	0.7850	0.0166	0.021	0.0213	[0.7438, 0.8262]	0.9287	-0.0185	0.0149	0.0150	[0.8995, 0.9579]
λ_2	-1.1484	0.0218	0.036	0.0365	[-1.2190, -1.0778]	-1.1148	-0.0124	0.0299	0.0292	[-1.1734, -1.0562]
p_1	0.7968	0.0018	0.0011	0.0011	[0.7946, 0.7990]	0.7984	-0.0001	0.0009	0.0009	[0.7966, 0.8002]

n=500						n=1000				
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	1.1096	0.0042	0.0044	0.0043	[1.1010, 1.1182]	1.1180	0.0038	0.0025	0.0026	[1.1131, 1.1229]
β_1	2.0004	-0.0001	0.0001	0.0001	[2.0002, 2.0006]	2.0000	0.0001	0.0001	0.0001	[1.9998, 2.0002]
β_2	3.0002	0.0002	0.0001	0.0001	[3.0000, 3.0004]	3.0001	0.0000	0.0001	0.0001	[2.9999, 3.0003]
μ_1	1.1312	-0.0006	0.0034	0.0034	[1.1245, 1.1379]	1.1272	0.0002	0.0023	0.0023	[1.1227, 1.1317]
μ_2	-4.4743	-0.0041	0.0037	0.0037	[-4.4816, -4.467]	-4.4810	-0.0012	0.0025	0.0025	[-4.4859, -4.4761]
σ_1^2	0.1945	0.0008	0.0034	0.0033	[0.1878, 0.2012]	0.1922	-0.0007	0.0007	0.0007	[0.1908, 0.1936]
σ_2^2	0.3969	0.0042	0.0040	0.0039	[0.3891, 0.4047]	0.4022	0.0001	0.0029	0.0029	[0.3965, 0.4079]
λ_1	0.9215	-0.0074	0.0107	0.0103	[0.9005, 0.9425]	0.9185	0.0060	0.0067	0.0066	[0.9054, 0.9316]
λ_2	-1.0474	-0.0071	0.0219	0.0213	[-1.0903, -1.0045]	-1.0408	0.0046	0.0146	0.0145	[-1.0694, -1.0122]
p_1	0.7982	0.0002	0.0006	0.0006	[0.7970, 0.7994]	0.7992	-0.0001	0.0004	0.0004	[0.7984, 0.8000]

n= 2000					
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	1.1268	-0.0005	0.0019	0.0019	[1.1231, 1.1305]
β_1	2.0002	-0.0001	0.0001	0.0001	[2.0000, 2.0004]
β_2	3.0000	0.0001	0.0001	0.0001	[2.9998, 3.0002]
μ_1	1.1225	0.0021	0.0016	0.0017	[1.1194, 1.1256]
μ_2	-4.4873	0.0011	0.0018	0.0018	[-4.4908, -4.4838]
σ_1^2	0.1933	0.0007	0.0005	0.0005	[0.1923, 0.1943]
σ_2^2	0.4004	0.0004	0.0021	0.0021	[0.3963, 0.4045]
λ_1	0.9423	0.0001	0.0044	0.0045	[0.9337, 0.9509]
λ_2	-1.0090	0.0024	0.0103	0.0100	[-1.0292, -0.9888]
p_1	0.7998	-0.0002	0.0003	0.0003	[0.7992, 0.8004]

Cuadro 4.1: (Censura al 8 % para $n = 150,\ 250,\ 500,\ 1000$ y 2000) Parámetros estimados del modelo MRL-CI-MF-NA mediante el método Bootstraps, para 1000 réplicas

	n=150					n=250				
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	1.0157	-0.0011	0.0209	0.0209	[0.9747, 1.0567]	1.1299	-0.0089	0.0103	0.0100	[1.1097, 1.1501]
β_1	2.0014	0.0005	0.0005	0.0005	[2.0004, 2.0024]	2.0005	-0.0002	0.0003	0.0003	[1.9999, 2.0011]
β_2	3.0040	0.0003	0.0010	0.0010	[3.0020, 3.0060]	3.0005	0.0001	0.0004	0.0004	[2.9997, 3.0013]
μ_1	1.1074	0.0206	0.0078	0.0077	[1.0921, 1.1227]	1.1098	0.0052	0.0053	0.0053	[1.0994, 1.1202]
μ_2	-4.3992	-0.0109	0.0136	0.0136	[-4.4259, -4.3725]	-4.4912	-0.0037	0.0059	0.0062	[-4.5028, -4.4796]
σ_1^2	0.3423	-0.0128	0.0279	0.0281	[0.2876, 0.3970]	0.2073	-0.0043	0.0104	0.0104	[0.1869, 0.2277]
σ_2^2	0.4629	0.0513	0.0335	0.0325	[0.3972, 0.5286]	0.4311	-0.0086	0.0090	0.0091	[0.4135, 0.4487]
λ_1	1.0342	0.0103	0.0230	0.0228	[0.9891, 1.0793]	0.9010	0.0058	0.0166	0.0163	[0.8685, 0.9335]
λ_2	-1.0999	-0.0226	0.0365	0.0372	[-1.1714, -1.0284]	-1.2402	0.0263	0.0349	0.0342	[-1.3086, -1.1718]
p_1	0.7985	-0.0001	0.0014	0.0014	[0.7958, 0.8012]	0.7996	0.0018	0.0009	0.0009	[0.7978, 0.8014]

	n=500					n=1000				
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	1.1221	0.0069	0.0040	0.0040	[1.1143, 1.1299]	1.1253	0.0002	0.0030	0.0030	[1.1194, 1.1312]
β_1	2.0001	0.0001	0.0001	0.0001	[1.9999, 2.0003]	2.0003	-0.0001	0.0001	0.0001	[2.0001, 2.0005]
β_2	3.0003	-0.0001	0.0001	0.0001	[3.0001, 3.0005]	3.0001	0.0000	0.0001	0.0001	[2.9999, 3.0003]
μ_1	1.1160	0.0021	0.0034	0.0035	[1.1093, 1.1227]	1.1208	0.0006	0.0027	0.0026	[1.1155, 1.1261]
μ_2	-4.4895	-0.0017	0.0039	0.0038	[-4.4971, -4.4819]	-4.4892	0.0015	0.0029	0.0029	[-4.4949, -4.4835]
σ_1^2	0.1897	-0.0012	0.0010	0.0010	[0.1877, 0.1917]	0.1902	0.0001	0.0007	0.0007	[0.1888, 0.1916]
σ_2^2	0.4159	-0.0027	0.0046	0.0044	[0.4069, 0.4249]	0.4001	0.0044	0.0030	0.0031	[0.3942, 0.406]
λ_1	0.8799	0.0099	0.0095	0.0091	[0.8613, 0.8985]	0.9040	0.0098	0.0064	0.0063	[0.8915, 0.9165]
λ_2	-1.1182	-0.0250	0.0250	0.0255	[-1.1672, -1.0692]	-1.0317	-0.0036	0.0153	0.0157	[-1.0617, -1.0017]
p_1	0.7997	0.0010	0.0006	0.0006	[0.7985, 0.8009]	0.7998	0.0003	0.0005	0.0005	[0.7988, 0.8008]

	n= 2000				
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	1.1235	0.0011	0.0021	0.0020	[1.1194, 1.1276]
β_1	2.0002	-0.0000	0.0001	0.0001	[2.0000, 2.0004]
β_2	3.0002	0.0001	0.0001	0.0001	[3.0000, 3.0004]
μ_1	1.1189	0.0033	0.0018	0.0018	[1.1154, 1.1224]
μ_2	-4.4860	-0.0013	0.0019	0.0020	[-4.4897, -4.4823]
σ_1^2	0.1940	0.0001	0.0004	0.0004	[0.1932, 0.1948]
σ_2^2	0.4006	0.0001	0.0023	0.0023	[0.3961, 0.4051]
λ_1	0.9395	0.0019	0.0038	0.0040	[0.9321, 0.9469]
λ_2	-1.0321	0.0175	0.0105	0.0109	[-1.0527, -1.0115]
p_1	0.8006	-0.0006	0.0003	0.0003	[0.8000, 0.8012]

Cuadro 4.2: (Censura al 20 % para $n = 150,\ 250,\ 500,\ 1000$ y 2000) Parámetros estimados del modelo MRL-CI-MF-NA mediante el método Bootstraps, para 1000 réplicas

n=150						n=250				
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	0.9871	-0.0140	0.0306	0.0307	[0.9271, 1.0471]	1.0865	0.0157	0.0088	0.0089	[1.0693, 1.1037]
β_1	2.0036	-0.0006	0.0008	0.0008	[2.0020, 2.0052]	2.0010	-0.0001	0.0002	0.0002	[2.0006, 2.0014]
β_2	3.0052	0.0009	0.0014	0.0014	[3.0025, 3.0079]	3.0006	0.0002	0.0003	0.0003	[3.0000, 3.0012]
μ_1	1.1331	-0.0014	0.0081	0.0082	[1.1172, 1.1490]	1.1219	0.0027	0.0060	0.0058	[1.1101, 1.1337]
μ_2	-4.4005	-0.0025	0.0170	0.0168	[-4.4338, -4.3672]	-4.4917	0.0084	0.0072	0.0072	[-4.5058, -4.4776]
σ_1^2	0.2726	0.0155	0.0248	0.0246	[0.2240, 0.3212]	0.1892	0.0019	0.0056	0.0058	[0.1782, 0.2002]
σ_2^2	0.5058	0.0018	0.0272	0.0283	[0.4525, 0.5591]	0.4300	-0.0103	0.0101	0.0107	[0.4102, 0.4498]
λ_1	0.9384	-0.0163	0.0234	0.0235	[0.8925, 0.9843]	0.8743	0.0092	0.0172	0.0175	[0.8406, 0.9080]
λ_2	-1.1719	-0.0261	0.0426	0.0409	[-1.2554, -1.0884]	-1.1224	-0.0619	0.0358	0.0356	[-1.1926, -1.0522]
p_1	0.7962	0.0018	0.0014	0.0014	[0.7935, 0.7989]	0.7999	-0.0002	0.0010	0.0010	[0.7979, 0.8019]

n=500						n=1000				
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	1.1119	-0.0030	0.0045	0.0045	[1.1031, 1.1207]	1.1146	-0.0003	0.0030	0.0031	[1.1087, 1.1205]
β_1	2.0006	-0.0000	0.0001	0.0001	[2.0004, 2.0008]	2.0004	0.0001	0.0001	0.0001	[2.0002, 2.0006]
β_2	3.0011	-0.0001	0.0001	0.0001	[3.0009, 3.0013]	3.0007	0.0000	0.0001	0.0001	[3.0005, 3.0009]
μ_1	1.1289	-0.0067	0.0036	0.0037	[1.1218, 1.1360]	1.1190	0.0028	0.0027	0.0028	[1.1137, 1.1243]
μ_2	-4.4868	-0.0038	0.0042	0.0042	[-4.4950, -4.4786]	-4.4912	0.0008	0.0030	0.0030	[-4.4971, -4.4853]
σ_1^2	0.1821	0.0008	0.0010	0.0010	[0.1801, 0.1841]	0.1902	-0.0006	0.0007	0.0007	[0.1888, 0.1916]
σ_2^2	0.4179	-0.0070	0.0044	0.0045	[0.4093, 0.4265]	0.4072	-0.0010	0.0031	0.0032	[0.4011, 0.4133]
λ_1	0.8373	0.0012	0.0092	0.0095	[0.8193, 0.8553]	0.8934	0.0029	0.0059	0.0062	[0.8818, 0.9050]
λ_2	-1.1486	0.0117	0.0257	0.0259	[-1.1990, -1.0982]	-1.0653	0.0030	0.0164	0.0163	[-1.0974, -1.0332]
p_1	0.7995	0.0006	0.0007	0.0007	[0.7981, 0.8009]	0.8006	-0.0005	0.0005	0.0005	[0.7996, 0.8016]

n= 2000					
Parámetro	**Estimado**	**Sesgo**	**SE**	$\hat{SE}$	**IC 95 %**
β_0	1.1145	-0.0005	0.0022	0.0023	[1.1102, 1.1188]
β_1	2.0003	0.0001	0.0001	0.0001	[2.0001, 2.0005]
β_2	3.0006	0.0001	0.0001	0.0001	[3.0004, 3.0008]
μ_1	1.1253	-0.0020	0.0020	0.0020	[1.1214, 1.1292]
μ_2	-4.4832	-0.0034	0.0023	0.0022	[-4.4877, -4.4787]
σ_1^2	0.1935	0.0001	0.0005	0.0005	[0.1925, 0.1945]
σ_2^2	0.4023	-0.0000	0.0024	0.0023	[0.3976, 0.4070]
λ_1	0.9324	0.0038	0.0042	0.0041	[0.9242, 0.9406]
λ_2	-1.0162	-0.0023	0.0102	0.0103	[-1.0362, -0.9962]
p_1	0.7996	0.0002	0.0004	0.0004	[0.7988, 0.8004]

Cuadro 4.3: (Censura al 35 % para $n = 150,\ 250,\ 500,\ 1000$ y 2000) Parámetros estimados del modelo MRL-CI-MF-NA mediante el método Bootstraps, para 1000 réplicas

Capítulo 5

Aplicaciones

5.1. Aplicación 1: Conjunto de datos de tasas salariales

Estos datos provienen de un estudio panel realizado por la Universidad de Michigan (Mroz, 1987) en el año 1975 orientado a obtener información sobre el mercado laboral de las esposas. La muestra consiste en 753 mujeres blancas casadas entre 30 y 60 años de edad durante el año 1975, donde 428 mujeres que lograron trabajar en algún momento del año. La variable dependiente son tasas salariales, definidas como las ganancias promedio por hora, las cuales en algunos casos se establecen como cero, lo que significa que las esposas no trabajaron en 1975 y por tanto en estos casos las observaciones se consideran censuradas a cero.

Las variables a utilizar son:

- y_i : Tasas salariales.
- x_{i1} : Edad de las esposas (en años)
- x_{i2} : Años de escolaridad de las esposas
- x_{i3} : Años de escolaridad del padre
- x_{i4} : Experiencia real en el mercado laboral

En la presente tesis, se utiliza este conjunto de datos para evaluar el rendimiento del algoritmo SAEM en el modelo propuesto, considerando diferente número de componentes para su evaluación.

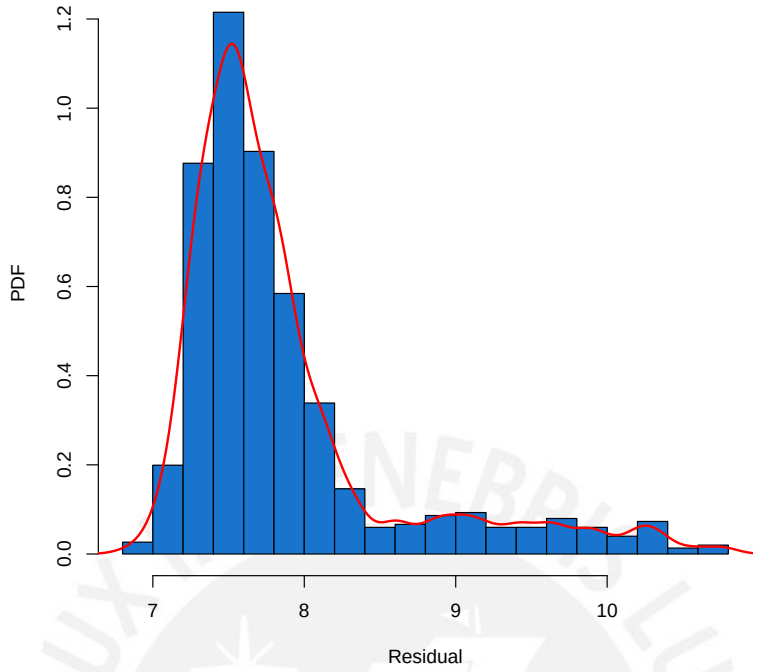


Figura 5.1: Histograma con la función de probabilidad de los residuales del las horas de trabajo anuales

5.1.1. Resultados

A fin de evaluar el desempeño del modelo MRL-CI-MF-NA definido en el capítulo 3 y ajustar los datos de las tasas salariales con tres covariables, se consideró de manera conveniente realizar una comparación del modelo propuesto con diferente número de componentes. Para tal fin, considerando el algoritmo SAEM, para estimar los parámetros $\boldsymbol{\theta}_j = (p_j, \sigma_j^2, \mu_j, \lambda_j)^\top$, se realiza una evaluación para los diferentes números de componentes. Los resultados son detallados en el Cuadro 5.1.

Criterio de selección

El Cuadro 5.1 contiene los resultados a partir de los indicadores AIC y BIC, definidos en la subsección 3.2.5 considerando diferente número de componentes (G=1, G=2 y G=3) comparando el modelo propuesto (MRL-CI-MF-NA) con los siguientes modelos:

- MRL-CI-MN: Modelo de regresión lineal con censura a la izquierda, cuando los errores siguen una mixtura de normales.
- MRL-CI-N: Modelo de regresión lineal con censura a la izquierda, cuando los errores siguen una distribución normal.

Este hallazgo confirma que, al inspeccionar los criterios mencionados anteriormente se demuestra que el modelo propuesto (MRL-CI-MF-NA) presenta estimaciones más precisas para G=2 componentes, ver (5.1).

Modelo	Criterio	G=1	G=2	G=3
MRL–CI-MF-NA	**AIC**	2914.794	2803.673	2808.221
	BIC	2965.411	2876.786	2903.83
MRL-CI-MN	**AIC**	2914.809	2981.526	2806.78
	BIC	2965.426	3054.639	2902.389
MRL-CI-N	**AIC**	2914.854	-	-
	BIC	2965.471	-	-

Cuadro 5.1: ($G = 1, 2$ y 3 componentes) Criterios de selección AIC y BIC para el modelo MRL-CI-MF-NA, MRL-CI-MN y MRL-CI-N aplicado al conjunto de datos de las tasas salariales para diferente número de componentes

Parámetros estimados

A continuación, se presentan las estimaciones por máxima verosimilitud de los parámetros para el conjunto de datos de tasas salariales, considerando el algoritmo SAEM para el modelo MRL-CI-MF-NA. Además, se observa el sesgo y los errores estándar obtenidos por el método Bootstraps, utilizando 400 muestras para 1000 réplicas. Estos resultados se muestran considerando 2 componentes, al ser seleccionado como el mejor candidato (Ver Cuadro 5.1).

Parámetro	Estimado	Sesgo	SE	IC 95 %
β_0	-4.9728	-0.0009	0.0464	[-5.0637, -4.8819]
β_1	-0.0366	-0.0012	0.0012	[-0.039 , -0.0342]
β_2	0.2499	0.0024	0.0019	[0.2462 , 0.2536]
β_3	0.0370	0.0006	0.0008	[0.0354 , 0.0386]
β_4	0.1098	0.0013	0.0018	[0.1063 , 0.1133]
μ_1	0.5726	-0.0005	0.0273	[0.5191 , 0.6261]
μ_2	0.0545	-0.0701	0.1504	[-0.2403, 0.3493]
σ_1^2	20.0354	0.0287	0.2643	[19.5174, 20.5534]
σ_2^2	122.4826	-2.0492	2.2563	[118.0603, 126.9049]
λ_1	-0.9661	-0.0378	0.0335	[-1.0318 , -0.9004]
λ_2	1.4090	-0.0287	0.0467	[1.3175 , 1.5005]
p_1	0.8256	-0.0040	0.0033	[0.8191 , 0.8321]

Cuadro 5.2: (G = 2 componentes) Estimaciones de los parámetros por el Algoritmo SAEM, sesgo, errores estándar obtenidos por bootstrap (SE), e intervalo de credibilidad al 95 % para el modelo MRL-CI-MF-NA aplicado al conjunto de datos de las tasas salariales

A partir de los resultados vistos en el Cuadro 5.2, se observan las estimaciones, sesgo y errores estándar obtenidos por el método bootstrap , considerando un número de réplicas de tamaño 1000 para 400 muestras con reemplazo, muestran el error estándar obtenido a través de los bootstrap (SE).

Perturbación de las observaciones

Definida la perturbación de la variable respuesta para una observación i: $y_i(\delta) = \delta Desv.Est(\mathbf{y})$, donde Desv.Est(.) denota la desviación estándar para diferentes valores $\delta = 5, 6, 7, 8, 9$ perturbaremos la observación n° 210, con la finalidad de ver el cambio relativo de las estimaciones

con respecto a las estimaciones sin perturbar la variable. Para ello se realiza una evaluación del cambio relativo de las estimaciones perturbadas frente a las estimaciones originales y se comparan los modelos de regresión lineal con censura a la izquierda, cuando los errores siguen una mixtura de normales (MRL-CI-MN) cuando los errores siguen una distribución normal (MRL-CI-N) y bajo el modelo propuesto (MRL-CI-MF-NA).

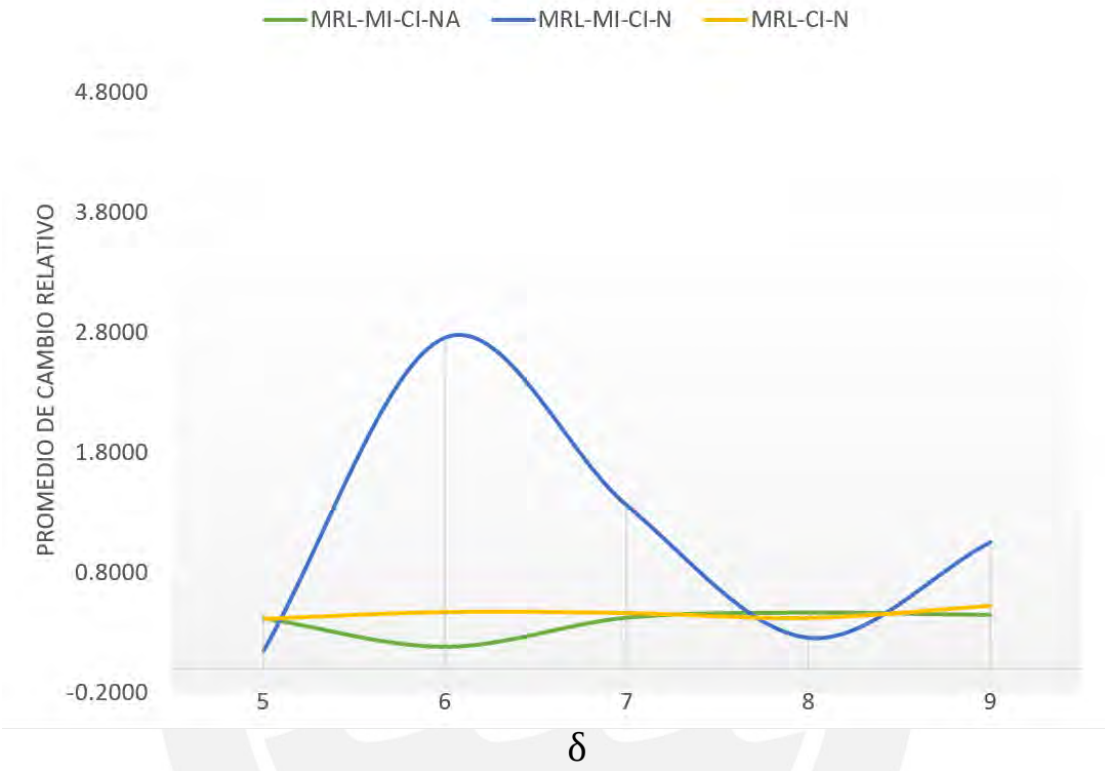


Figura 5.2: Promedio de cambio relativo de las estimaciones para cada modelo y diferentes perturbaciones

El interés se encuentra en observar el cambio relativo de las estimaciones $|(\hat{\beta}_j(\delta)\text{-}\hat{\beta}_j)/\hat{\beta}_j|$. Se concluye que el modelo popuesto (MRL-CI-MF-NA), presenta mayor robustez frente a las perturbaciones que presenta la variable respuesta.

5.2. Aplicación 2: Conjunto de ingresos en una entidad financiera

Estos datos corresponden a una población de N =2,587 clientes que entraron en campaña (Conjunto de estrategias para la captación de nuevos clientes, fidelización, etc., a partir del ofrecimiento de mejores tasas de interés para acceder a los productos de la entidad como préstamos vehiculares, hipotecarios, tarjetas de crédito, entre otros) en una entidad financiera en el mes de julio del 2019 por presentar una buena clasificación en los últimos 48 meses y presentar información de saldo retail en los últimos 24 meses (Bancarizados) en el sistema financiero. La fuente de información corresponde a la base de datos internas de clientes que se encuentran vinculados con saldos pasivos en la entidad desde enero del 2019.

Además, es importante mencionar que, según las políticas internas de la entidad, un cliente puede entrar en campaña si al menos tiene un ingreso básico al 2019 de 930 soles, siendo este

el valor de corte para clasificar una observación censurada, en caso no se cuente con información del ingreso. Por lo tanto, cuando no se cuenta con información del ingreso del cliente este será clasificado como una observación censurada y su ingreso será reasignado con el valor de 930 soles. El propósito es fidelizar a los clientes seleccionados, ofreciéndoles productos internos (tarjetas, préstamos o compras de deuda) debido a su buen historial crediticio en el sistema financiero en los últimos 48 meses.

Del total de clientes evaluados, el 6.9 % fueron censurados (178), es decir clientes con información incompleta en los ingresos. Para estudiar que factores podrían estar asociados con el ingreso, se evaluaron 2 variables que presentaron mayor peso en las versiones documentadas del modelo de estimación de ingresos en la entidad y que por la confidencialidad se reserva la información para la presente tesis.

5.2.1. Información derivada del sistema financiero

Las variables de esta subsección reflejan la información de deuda y del comportamiento de pago de los clientes en el Sistema Financiero.

- Saldo Retail: Corresponde a la suma de saldos deudores por: Pequeñas empresas, microempresas, consumo revolvente, consumo no revolvente e hipotecario que mantiene un cliente con todas las entidades del sistema financiero obtenida el último reporte disponible del RCC.
- Clasificación de riesgo: Es la clasificación que determina la SBS en función a los días de atraso registrados por el cliente en cada cuenta de las entidades donde se mantenga saldo deudor.
- Calificativo normal: Registra el valor 1, si el individuo no presenta ninguna clasificación peor a normal en más del 5 % de su saldo total en el sistema financiero.
- Edad en años de las personas

5.2.2. Descripción del portafolio

	Media	**Desv.Est**	**Perc. 25**	**Perc. 50**	**Perc. 97.5**
Ingreso	1,224.09	237.22	993.61	1,175.38	1,693.15
SaldoRetailProm24M	2,609.16	2,169.64	1,300.30	2,135.10	7,654.81
Edad	45.49	12.51	35.00	43.00	74.00

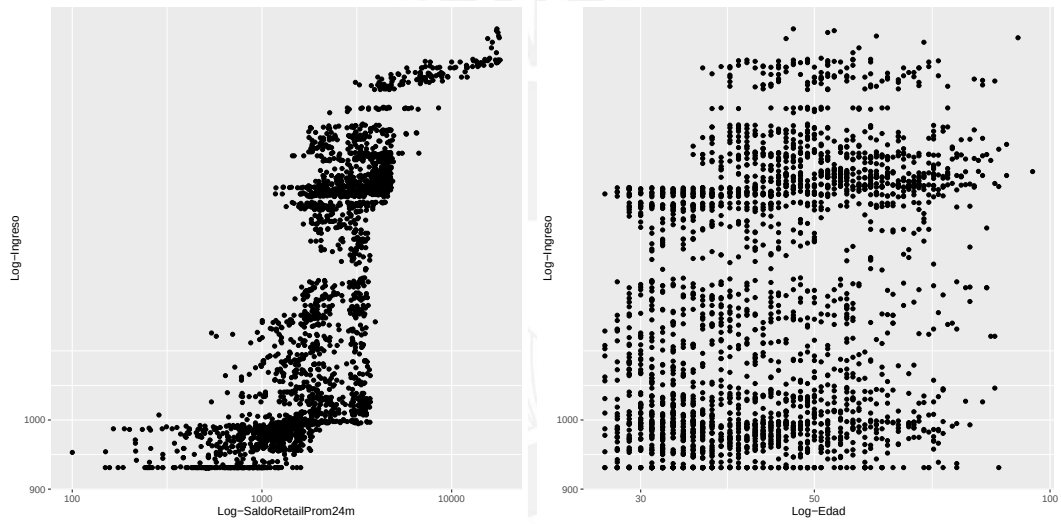
Cuadro 5.3: Resumen descriptivo de las variables ingreso, edad y saldo promedio retail en los últimos 24 meses (SaldoRetailProm24M) aplicado al conjunto de ingresos en una entidad financiera

Este conjunto de datos es evaluado en el modelo propuesto (MRL-CI-MF-NA), con la finalidad de poder estimar el ingreso del cliente (Y) a partir de las covariables x_i. Estimar el ingreso del cliente surge ante la necesidad de darle una mejor asignación de ofertas a los clientes con ingresos estimados más altos y presentar una actitud conservadora para aquellos clientes con un ingreso estimado más bajo. El interés de evaluar la eficiencia del modelo

propuesto en este conjunto de datos, surge a partir que en la muestra, la distribución de los datos presentan asimetría y se evidencia grupos diferenciados en cuanto al ingreso real (multimodalidad), por lo cual, una mixtura de normales asimétricas podría resultar ser más conveniente.

El conjunto de datos está conformado por las siguientes variables:

- Ingreso: Ingreso del cliente (Y)
- SaldoRetailProm24M: Promedio de saldo retail en los últimos 24 meses (X_1).
- Edad: Edad en años del cliente (X_2)



(a) (b)

Figura 5.3: Gráfico de dispersión del . (a) Gráfico de dispersión del Logarítmo del ingreso vs Logaritmo del Promedio de saldo retail en los últimos 24 meses, (b)Gráfico de dispersión del Logarítmo del ingreso vs Logarítmo de la Edad.

Por ende, la representación líneal de este modelo es la siguiente:

$$Ingreso_i = \beta_0 + \beta_1 SaldoRetailProm24M_i + \beta_2 Edad_i + \varepsilon_i,$$

donde los errores siguen una mixtura finita de una distribución normal asimétrica (MRL-MF-CI-NA)

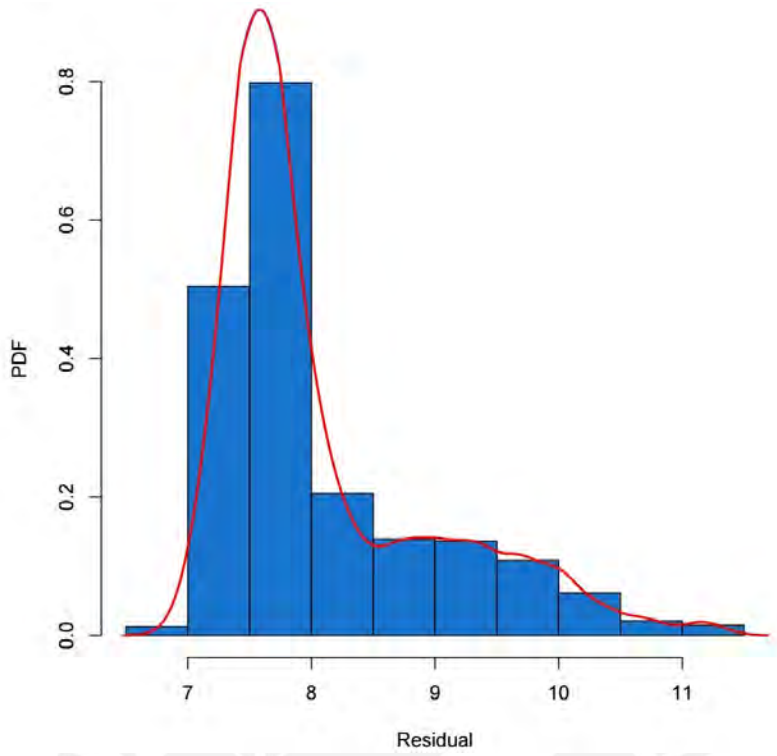


Figura 5.4: Histograma con la función de densidad de probabilidad (PDF) de los residuales del ingreso

5.2.3. Fuentes de información

Para asegurar la calidad de la información, se escogieron las siguientes fuentes de información crediticia y sociodemográfica:

1. Reporte crediticio consolidado (RCC): Es un reporte mensual que la Superintendencia de Banca, Seguros y AFP (SBS) emite a todas las entidades financieras bancarias. El RCC comprende a todas las personas (jurídicas y naturales) que participan en el mercado crediticio y muestra la información de los activos financieros que las entidades utilizan, para calcular el gasto de provisión. La información es enviada mensualmente con un desfase de hasta dos meses.

2. Bases internas: Base de datos de uso interno que contienen información de los clientes de la entidad.

5.2.4. Resultados

A fin de evaluar el desempeño del modelo MRL-CI-MF-NA definido en el capítulo 3 y ajustar los datos de los ingresos con la covariable edad y promedio de saldo retail en los últimos 24 meses, se consideró de manera conveniente realizar una comparación del modelo propuesto con una y dos componentes. Para tal fin, se consideró el algoritmo SAEM, para estimar los parámetros $\boldsymbol{\theta}_j = (p, \sigma^2, \mu, \lambda)^\top$. Los resultados son detallados en el Cuadro 5.4

Criterio de selección

El cuadro 5.4 contiene los resultados basados en los indicadores AIC y BIC, definido en la subsección 3.2.5 considerando diferente número de componentes (G=1,G=2 y G=3) las

cuales son evaluadas para el modelo de regresión lineal con censura a la izquierda, cuando los errores siguen una mixtura de normales (MRL-CI-MN), para el modelo de regresión lineal con censura a la izquierda, cuando los errores siguen una distribución normal (MRL-CI-N) y para el modelo propuesto (MRL-CI-MF-NA). Este hallazgo confirma que al inspeccionar los criterios mencionados anteriormente se demuestra que el modelo propuesto (MRL-CI-MF-NA) presenta estimaciones más precisas para G=2 componentes.

Modelo	**Criterio**	**G=1**	**G=2**	**G=3**
MRL–CI-MF-NA	**AIC**	5719.694	5046.435	5530.505
	BIC	5767.702	5121.876	5633.379
MRL-CI-MN	**AIC**	5720.022	5646.749	5534.747
	BIC	5768.03	5722.19	5637.621
MRL-CI-N	**AIC**	5719.838	-	-
	BIC	5767.846	-	-

Cuadro 5.4: ($G = 1, 2$ y 3 componentes) Criterios de selección AIC y BIC para el modelo MRL-CI-MF-NA, MRL-CI-MN y MRL-CI-N aplicado al conjunto de ingresos de una entidad financiera para diferente número de componentes

Parámetros estimados

A continuación, se presentan las estimaciones por máxima verosimilitud de los parámetros para el conjunto de datos de ingresos en una entidad financiera, considerando el algoritmo SAEM para el modelo MRL-CI-MF-NA. Además, se observa el sesgo y los errores estándar obtenidos por el método bootstrap, utilizando 400 muestras con reemplazamiento para 1000 réplicas. Estos resultados se muestran considerando 2 componentes, al ser seleccionado como el mejor candidato (Ver Cuadro 5.4).

Parámetro	**Estimado**	**Sesgo**	**SE**	**IC 95 %**
β_0	-0.1225	-0.0134	0.0120	[-0.1460, -0.0990]
β_1	0.0001	0.0000	0.0000	[0.0001 , 0.0001]
β_2	-0.0019	0.0000	0.0001	[-0.0021, -0.0017]
μ_1	0.2153	-0.0196	0.0132	[0.1894 , 0.2412]
μ_2	-0.6479	0.0035	0.0159	[-0.6791, -0.6167]
σ_1^2	0.4102	-0.0130	0.0090	[0.3926 , 0.4278]
σ_2^2	2.1908	0.0504	0.0685	[2.0565 , 2.3251]
λ_1	-0.2013	0.0072	0.0168	[-0.2342, -0.1684]
λ_2	-2.9712	0.0863	0.0547	[-3.0784, -2.8640]
p_1	0.7322	0.0041	0.0059	[0.7206 , 0.7438]

Cuadro 5.5: (G = 2 componentes) Estimaciones de los parámetros por el algoritmo SAEM, sesgo, errores estándar obtenidos por bootstrap (SE) e intervalo de credibilidad al 95 % para el modelo MRL-CI-MF-NA aplicado al conjunto de ingresos en una entidad financiera

Capítulo 6

Conclusiones

6.1. Conclusiones y discusión

- En este trabajo se ha propuesto un modelo de regresión lineal con censura basado en una mixtura finita de una distribución normal asimétrica, caso particular del modelo propuesto por Thalita et al. (2017), pero ampliada para el caso en que se admitan varias componentes. La inferencia del modelo es realizada desde la perspectiva clásica usándose el algoritmo de aproximaciones estocásticas de esperanza Maximización (SAEM), propuesto por Delyon et al. (1999) y desarrollado por Galarza et al. (2017)

- Se obtuvo la función log-verosimilitud completa y se estimaron los parámetros a partir del algoritmo SAEM, lo cual facilitó la convergencia de las estimaciones.

- Basado en los resultados del estudio de simulación para todos los niveles de censura (Ver Cuadro A.1), se concluye que el modelo propuesto tiene un mejor ajuste cuando el tamaño de muestra aumenta, independientemente del nivel de censura que tome.

- Mediante el estudio de simulación, se verificó que el método descrito permite recuperar los parámetros del modelo de regresión lineal con censura basada en una mixtura finita de escala normal asimétrica al presentar menos sesgo a medida que aumenta el número de observaciones.

- El método bootstrap utilizado en la presente tesis para la obtención de los errores estándar (SE) preserva las ideas teóricas considerando un número de réplicas de tamaño 1000 para 400 muestras con reemplazo. Es así que se corrobora, a través de los resultados vistos en los Cuadros 4.1, 4.2 y 4.3, que el error es menor a medida que aumenta el tamaño de muestra independientemente del nivel de censura utilizado. Además se evidencia que el error estándar obtenido a través de bootstrap (SE) se aproxima al estimador del error estándar obtenido a partir de las 1000 simulaciones.

- En el Cuadro A.2 se muestra las medidas de la raíz cuadrática media del error (RMSE), del error absoluto medio (MAE) y del error cuadrático medio (MSE), las cuales son medidas de precisión de las estimaciones de los parámetros frente a los datos teóricos. Los resultados de estas medidas evidencian que los parámetros para una data censurada al 8 %, 20 % y 35 %, recupera de forma satisfactoria los parámetros a mayor cantidad de réplicas, dado que a mayor tamaño de réplica, los errores son menores.

- Se considera como aplicación del modelo propuesto en primer lugar al conjunto de datos reales de tasas salariales, mostrando un desempeño adecuado en el ajuste de las estimaciones por lo cual se concluye:

  - En general, la gráfica de cajas y densidad (Ver Figura 5.1), evidencia una asimetría a la derecha con al menos una bimodalidad en la distribución de la variable respuesta la cual sugiere hacer comparaciones para diferentes tamaños de muestra y ver el rendimiento de las estimaciones.
  - Por otro lado, para los diferentes componentes y modelos evaluados, la distribución con mejor ajuste, según el criterio AIC y BIC es para $g = 2$, siendo el modelo propuesto el más óptimo. (Ver Cuadro 5.1).
  - Se observa en la figura 5.2 mayor robustez en las estimaciones del modelo propuesto (MRL-CI-MFNA) frente a los demás modelos comparados, dado que el promedio del cambio relativo de las estimaciones de los coeficientes conservan a pesar de las perturbaciones menor variabilidad en promedio.

- Se considera como aplicación del modelo propuesto en segundo lugar al conjunto de ingresos de una entidad financiera, mostrando un desempeño adecuado en el ajuste de las estimaciones por lo cual se concluye:

  - Con respecto a la evaluación del modelo de regresión lineal con censura a la izquierda cuando los errores siguen una mixtura de normales (MRL-CI-MN) y para el modelo de regresión lineal con censura a la izquierda cuando los errores siguen una distribución normal (MRL-CI-N) con respecto al modelo propuesto (MRL-CI-MF-NA) se corrobora que los datos se adecuan mejor utilizando el modelo propuesto (MRL-CI-MFNA) (Ver 5.4) .
  - A partir del cuadro 5.5, se aprecia las estimaciones obtenidas con el modelo propuesto (MRL-CI-MFNA), por lo cual se evidencia que las estimaciones, sesgo y errores estándar obtenidos por el método bootstrap , considerando un número de réplicas de tamaño 1000 para 400 muestras con reemplazo y el error estándar obtenido a través de los bootstrap (SE).

6.2. Sugerencias para investigaciones futuras

- Una extensión a considerar que se puede realizar para los resultados mostrados en esta tesis es considerar otras distribuciones de la clase normal asimétrica, tales como la t-asimétrica, slash-asimétrica y normal-contaminada-asimétrica, de esta forma conseguiremos extender diferentes trabajos tales como el de Thalita et al. (2017), Benites et al. (2018) y Benites et al. (2019).

- Inspirados en el artículo reciente de Zeller et al. (2018), la consideración de mixturas finitas de regresiones con censura y, donde el error sigue una distribución MF-MENA es posible de realizar como un trabajo futuro.

Apéndice A

Anexos de Cuadros

		Censura al 8 %					Censura al 20 %					Censura al 35				
	n	**Media**	**Desv. Est**	**Perc.25**	**Perc.50**	**Perc.97.5**	**Media**	**Desv. Est**	**Perc.25**	**Perc.50**	**Perc.97.5**	**Media**	**Desv. Est**	**Perc.25**	**Perc.50**	**Perc.97.5**
	150	0.880	0.556	0.825	1.054	1.498	1.015	0.660	0.913	1.097	1.556	0.972	0.973	0.901	1.075	1.643
	250	1.064	0.325	0.987	1.116	1.450	1.121	0.316	1.003	1.134	1.507	1.103	0.281	0.954	1.110	1.540
$\beta_0 = 1$	500	1.114	0.136	1.026	1.120	1.345	1.129	0.127	1.044	1.132	1.379	1.109	0.142	1.013	1.109	1.399
	1000	1.122	0.083	1.072	1.120	1.283	1.125	0.094	1.065	1.123	1.312	1.115	0.097	1.050	1.114	1.301
	2000	1.126	0.060	1.085	1.124	1.245	1.125	0.064	1.082	1.124	1.243	1.114	0.072	1.066	1.111	1.258
	150	2.006	0.015	1.997	2.002	2.049	2.002	0.016	1.996	2.000	2.019	2.003	0.025	1.995	2.000	2.020
	250	2.002	0.010	1.997	2.000	2.037	2.000	0.010	1.996	2.000	2.011	2.001	0.007	1.997	2.001	2.012
$\beta_1 = 2$	500	2.000	0.004	1.998	2.000	2.007	2.000	0.004	1.998	2.000	2.007	2.001	0.004	1.998	2.001	2.008
	1000	2.000	0.002	1.999	2.000	2.005	2.000	0.002	1.999	2.000	2.005	2.000	0.003	1.999	2.001	2.005
	2000	2.000	0.002	1.999	2.000	2.003	2.000	0.002	1.999	2.000	2.003	2.000	0.002	1.999	2.000	2.004
	150	3.009	0.024	2.997	3.001	3.080	3.004	0.031	2.995	3.001	3.020	3.006	0.045	2.995	3.001	3.024
	250	3.002	0.011	2.997	3.001	3.043	3.001	0.012	2.997	3.000	3.011	3.001	0.011	2.996	3.001	3.013
$\beta_2 = 3$	500	3.000	0.004	2.998	3.000	3.007	3.000	0.004	2.998	3.000	3.008	3.001	0.004	2.998	3.001	3.010
	1000	3.000	0.002	2.998	3.000	3.005	3.000	0.003	2.998	3.000	3.005	3.001	0.003	2.999	3.001	3.006
	2000	3.000	0.002	2.999	3.000	3.003	3.000	0.002	2.999	3.000	3.004	3.001	0.002	3.000	3.001	3.004
	150	1.117	0.193	0.981	1.118	1.513	1.128	0.245	1.002	1.143	1.563	1.132	0.260	0.988	1.144	1.586
	250	1.128	0.155	1.023	1.126	1.431	1.115	0.168	1.007	1.102	1.447	1.125	0.183	1.009	1.128	1.462
$\mu_1 = 1$	500	1.131	0.108	1.057	1.130	1.333	1.118	0.110	1.042	1.118	1.343	1.122	0.117	1.041	1.129	1.344
	1000	1.127	0.074	1.079	1.124	1.277	1.121	0.083	1.065	1.120	1.290	1.122	0.088	1.062	1.123	1.298
	2000	1.125	0.052	1.090	1.126	1.229	1.122	0.057	1.082	1.123	1.230	1.123	0.064	1.082	1.124	1.247
	150	-4.426	0.230	-4.588	-4.422	-3.952	-4.410	0.432	-4.604	-4.445	-3.910	-4.402	0.532	-4.610	-4.453	-3.841
	250	-4.463	0.176	-4.584	-4.466	-4.094	-4.495	0.196	-4.618	-4.508	-4.136	-4.483	0.229	-4.613	-4.486	-4.109
$\mu_2 = -4$	500	-4.478	0.117	-4.559	-4.480	-4.243	-4.491	0.120	-4.571	-4.497	-4.255	-4.491	0.132	-4.580	-4.485	-4.247
	1000	-4.482	0.079	-4.534	-4.485	-4.315	-4.488	0.092	-4.551	-4.488	-4.306	-4.490	0.096	-4.553	-4.488	-4.302
	2000	-4.486	0.058	-4.527	-4.485	-4.373	-4.487	0.063	-4.529	-4.484	-4.373	-4.487	0.069	-4.530	-4.487	-4.347
	150	0.575	0.830	0.159	0.199	2.674	0.327	0.889	0.150	0.192	0.381	0.288	0.778	0.141	0.177	0.375
	250	0.281	0.436	0.161	0.190	2.355	0.203	0.329	0.154	0.178	0.301	0.191	0.182	0.148	0.180	0.305
$\sigma_1^2 = 0.2$	500	0.195	0.105	0.166	0.189	0.265	0.188	0.030	0.167	0.187	0.254	0.183	0.030	0.161	0.183	0.246
	1000	0.192	0.022	0.177	0.191	0.237	0.190	0.022	0.176	0.191	0.233	0.190	0.022	0.175	0.190	0.233
	2000	0.194	0.015	0.185	0.195	0.222	0.194	0.014	0.185	0.195	0.219	0.194	0.015	0.184	0.194	0.221
	150	0.392	0.190	0.256	0.350	0.849	0.513	1.028	0.255	0.340	1.040	0.508	0.897	0.237	0.342	1.252
	250	0.403	0.158	0.280	0.369	0.757	0.422	0.289	0.272	0.368	0.865	0.420	0.337	0.262	0.370	0.873
$\sigma_2^2 = 0.4$	500	0.401	0.124	0.304	0.377	0.670	0.413	0.138	0.309	0.383	0.734	0.411	0.143	0.305	0.382	0.733
	1000	0.402	0.092	0.333	0.392	0.607	0.404	0.097	0.331	0.390	0.626	0.406	0.103	0.333	0.394	0.643
	2000	0.401	0.068	0.352	0.396	0.543	0.401	0.072	0.349	0.395	0.559	0.402	0.074	0.350	0.397	0.559
	150	0.802	0.674	0.251	0.734	2.426	1.044	0.720	0.469	0.940	2.794	0.923	0.743	0.312	0.793	2.778
	250	0.911	0.475	0.572	0.905	1.981	0.907	0.514	0.518	0.869	2.188	0.883	0.555	0.432	0.842	2.265
$\lambda_1 = 1$	500	0.914	0.325	0.702	0.918	1.650	0.890	0.289	0.716	0.903	1.527	0.838	0.301	0.639	0.868	1.410
	1000	0.924	0.207	0.796	0.940	1.323	0.914	0.199	0.799	0.939	1.275	0.896	0.197	0.781	0.927	1.239
	2000	0.943	0.141	0.868	0.958	1.170	0.941	0.127	0.876	0.955	1.139	0.936	0.130	0.871	0.957	1.130
	150	-1.126	1.155	-1.701	-0.698	-0.057	-1.123	1.178	-1.692	-0.655	-0.045	-1.199	1.293	-1.923	-0.631	-0.037
	250	-1.127	0.923	-1.640	-0.880	-0.114	-1.215	1.080	-1.818	-0.877	-0.079	-1.185	1.127	-1.790	-0.843	-0.064
$\lambda_2 = -1$	500	-1.055	0.673	-1.399	-0.919	-0.220	-1.144	0.807	-1.497	-0.971	-0.170	-1.137	0.819	-1.470	-0.953	-0.152
	1000	-1.036	0.457	-1.304	-0.985	-0.307	-1.035	0.495	-1.285	-0.957	-0.296	-1.062	0.516	-1.285	-0.978	-0.282
	2000	-1.007	0.318	-1.199	-0.975	-0.407	-1.015	0.343	-1.187	-0.972	-0.420	-1.018	0.325	-1.192	-0.988	-0.450
	150	0.799	0.034	0.775	0.799	0.862	0.798	0.043	0.773	0.796	0.875	0.798	0.046	0.770	0.795	0.881
	250	0.798	0.027	0.780	0.800	0.849	0.801	0.030	0.782	0.803	0.856	0.800	0.032	0.779	0.800	0.860
$p_1 = 0.8$	500	0.798	0.019	0.785	0.799	0.835	0.801	0.019	0.788	0.801	0.837	0.800	0.021	0.786	0.799	0.839
	1000	0.799	0.013	0.790	0.799	0.823	0.800	0.015	0.791	0.800	0.828	0.800	0.015	0.790	0.800	0.830
	2000	0.800	0.009	0.793	0.799	0.818	0.800	0.010	0.793	0.800	0.820	0.800	0.011	0.792	0.800	0.823

Cuadro A.1: (Censura al 8 %, 20 % Y 35 %) Media, desviación estándar, percentil 25, percentil 50, percentil 97.5 para los parámetros β_0, β_1, β_2, μ_1, μ_2, σ_1^2, σ_2^2, λ_1, λ_2 y p_1 del modelo MRL-CI-MF-NA

A.1. Error absoluto medio (MAE), raíz del error cuadrático medio (RMSE) y error cuadrático medio (MSE) de los parámetros para el estudio de simulación

Medida de precisión	Parámetro	Censura al 8 %					Censura al 20 %					Censura al 35 %				
		150	**250**	**500**	**1000**	**2000**	**150**	**250**	**500**	**1000**	**2000**	**150**	**250**	**500**	**1000**	**2000**
RMSE	$\beta_0 = 1$	0.568	0.332	0.178	0.147	0.140	0.660	0.338	0.181	0.157	0.140	0.973	0.299	0.179	0.150	0.135
	$\beta_1 = 2$	0.016	0.010	0.004	0.002	0.002	0.017	0.010	0.004	0.002	0.002	0.025	0.008	0.004	0.003	0.002
	$\beta_2 = 3$	0.026	0.011	0.004	0.002	0.002	0.032	0.012	0.004	0.003	0.002	0.046	0.011	0.005	0.003	0.002
	$\mu_1 = 1$	0.225	0.200	0.169	0.147	0.135	0.276	0.203	0.162	0.147	0.135	0.291	0.222	0.169	0.150	0.139
	$\mu_2 = -4$	0.484	0.495	0.492	0.489	0.490	0.595	0.532	0.506	0.496	0.491	0.667	0.535	0.508	0.499	0.491
	$\sigma_1^2 = 0.2$	0.910	0.443	0.105	0.024	0.016	0.897	0.329	0.032	0.024	0.015	0.783	0.183	0.035	0.024	0.017
	$\sigma_2^2 = 0.4$	0.190	0.158	0.123	0.092	0.068	1.034	0.290	0.139	0.097	0.072	0.903	0.338	0.144	0.103	0.074
	$\lambda_1 = 1$	0.702	0.483	0.336	0.221	0.152	0.721	0.522	0.310	0.217	0.140	0.746	0.567	0.341	0.222	0.145
	$\lambda_2 = -1$	1.161	0.931	0.675	0.458	0.318	1.184	1.101	0.819	0.496	0.344	1.308	1.141	0.830	0.520	0.325
	$p_1 = 0.8$	0.034	0.027	0.019	0.013	0.009	0.043	0.030	0.019	0.015	0.010	0.046	0.032	0.021	0.015	0.011
MAE	$\beta_0 = 1$	0.370	0.217	0.141	0.128	0.127	0.289	0.195	0.150	0.133	0.126	0.345	0.206	0.143	0.125	0.117
	$\beta_1 = 2$	0.010	0.005	0.003	0.002	0.001	0.007	0.005	0.003	0.002	0.001	0.009	0.005	0.003	0.002	0.001
	$\beta_2 = 3$	0.014	0.006	0.003	0.002	0.001	0.010	0.005	0.003	0.002	0.001	0.013	0.005	0.004	0.002	0.002
	$\mu_1 = 1$	0.180	0.161	0.142	0.129	0.125	0.216	0.158	0.133	0.126	0.123	0.228	0.179	0.140	0.128	0.125
	$\mu_2 = -4$	0.433	0.464	0.478	0.482	0.486	0.490	0.500	0.491	0.488	0.487	0.512	0.490	0.491	0.490	0.487
	$\sigma_1^2 = 0.2$	0.417	0.122	0.031	0.019	0.013	0.177	0.055	0.026	0.019	0.012	0.152	0.048	0.028	0.019	0.013
	$\sigma_2^2 = 0.4$	0.149	0.128	0.100	0.074	0.055	0.271	0.155	0.110	0.077	0.057	0.282	0.161	0.114	0.081	0.059
	$\lambda_1 = 1$	0.572	0.379	0.257	0.167	0.114	0.569	0.404	0.238	0.159	0.104	0.599	0.448	0.260	0.164	0.105
	$\lambda_2 = -1$	0.872	0.714	0.520	0.355	0.246	0.897	0.827	0.591	0.366	0.258	0.997	0.858	0.584	0.375	0.244
	$p_1 = 0.8$	0.027	0.022	0.015	0.010	0.007	0.031	0.023	0.015	0.012	0.008	0.034	0.025	0.017	0.012	0.009
MSE	$\beta_0 = 1$	0.323	0.110	0.032	0.022	0.020	0.435	0.114	0.033	0.025	0.020	0.946	0.089	0.032	0.023	0.018
	$\beta_1 = 2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
	$\beta_2 = 3$	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000
	$\mu_1 = 1$	0.051	0.040	0.029	0.022	0.018	0.076	0.041	0.026	0.022	0.018	0.085	0.049	0.029	0.022	0.019
	$\mu_2 = -4$	0.234	0.245	0.243	0.239	0.240	0.354	0.283	0.256	0.246	0.241	0.445	0.286	0.258	0.249	0.242
	$\sigma_1^2 = 0.2$	0.829	0.196	0.011	0.001	0.000	0.805	0.108	0.001	0.001	0.000	0.613	0.033	0.001	0.001	0.000
	$\sigma_2^2 = 0.4$	0.036	0.025	0.015	0.009	0.005	1.068	0.084	0.019	0.009	0.005	0.815	0.114	0.021	0.011	0.005
	$\lambda_1 = 1$	0.492	0.234	0.113	0.049	0.023	0.520	0.273	0.096	0.047	0.020	0.557	0.322	0.116	0.049	0.021
	$\lambda_2 = -1$	1.349	0.867	0.456	0.210	0.101	1.402	1.213	0.671	0.246	0.118	1.710	1.303	0.689	0.270	0.106
	$p_1 = 0.8$	0.001	0.001	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.000

Cuadro A.2: (Censura al 8 %, 20 % y 35 % para n = 150, 250, 500, 1000 y 2000), RMSE: Raíz del error cuadrático medio, MAE: Error absoluto medio y MSE: Error cuadrático medio del modelo MRL-CI-MF-NA

A.2. Estadísticos de los Bootstraps para los residuales para el estudio de simulación

	n=150					n=250				
Réplicas	**Media**	**Desv.Est**	**Perc. 25**	**Perc. 50**	**Perc. 97.5**	**Media**	**Desv.Est**	**Perc. 25**	**Perc. 50**	**Perc. 97.5**
1	0.3632	2.8354	-0.0499	1.2518	4.8588	-0.151	2.560	-3.576	1.274	1.815
2	-0.5616	3.3381	-4.2014	0.8831	4.7676	1.421	2.671	0.914	1.261	6.206
3	-0.0742	3.0548	-1.3412	0.9470	3.9375	-0.207	3.274	-3.679	1.246	4.375
4	-0.5601	2.7582	-3.5478	1.2255	2.1187	1.612	4.051	0.934	1.362	11.752
5	0.6879	2.1814	1.0417	1.4795	1.9070	0.106	2.510	0.857	1.298	1.796
6	-2.0210	3.0821	-5.0591	-3.7132	1.6285	0.619	3.417	0.792	1.329	6.312
7	-0.1977	3.2468	-1.8616	0.9435	4.4209	0.855	3.554	0.713	1.389	8.527
8	0.0506	2.2386	0.8121	1.0140	1.5115	0.344	3.490	0.449	1.239	7.705
9	1.3597	5.0964	-1.4518	1.3036	12.9131	0.560	2.892	0.639	1.223	6.160
10	-0.6913	2.6708	-3.6591	0.7973	2.0050	1.959	4.562	0.865	1.190	11.897
	n=500					**n=1000**				
1	0.674	3.600	0.767	1.102	8.414	0.761	2.622	0.779	1.172	6.236
2	0.867	2.901	0.831	1.181	8.743	0.819	2.905	0.700	1.170	8.471
3	0.865	3.629	0.583	1.364	9.650	0.361	2.660	0.608	1.160	4.741
4	0.881	2.985	0.822	1.160	8.634	0.533	2.985	0.690	1.182	6.718
5	0.671	2.867	0.719	1.108	6.070	0.628	2.940	0.803	1.169	8.324
6	0.339	2.698	0.687	1.075	4.115	0.683	2.601	0.753	1.280	3.638
7	0.727	2.935	0.664	1.141	4.649	0.570	2.035	0.857	1.199	2.269
8	0.211	3.323	-2.801	1.019	9.007	0.968	2.392	1.000	1.260	5.759
9	0.107	3.309	-3.798	1.114	8.074	0.683	2.834	0.752	1.161	6.453
10	0.092	2.384	0.510	1.123	2.093	0.734	2.733	0.798	1.147	7.037
	n=2000									
1	0.549	2.553	0.679	1.099	4.610					
2	0.704	2.393	0.859	1.184	5.755					
3	0.483	2.744	0.617	1.083	6.878					
4	0.735	2.809	0.810	1.205	7.294					
5	0.729	3.168	0.697	1.104	8.130					
6	0.824	2.592	0.807	1.163	6.374					
7	0.961	3.259	0.767	1.129	11.793					
8	0.504	2.903	0.626	1.059	7.286					
9	0.607	2.395	0.787	1.054	2.439					
10	0.498	2.927	0.648	1.059	5.801					

Cuadro A.3: (Censura al 8 % para n = 150, 250, 500, 1000 y 2000) Media, desviación estándar, percentil 25, percentil 50, percentil 97.5 para los Bootstraps de los residuales del modelo MRL-CI-MF-NA, para 10 réplicas

	n=150					n=250				
Réplicas	**Media**	**Desv.Est**	**Perc. 25**	**Perc. 50**	**Perc. 97.5**	**Media**	**Desv.Est**	**Perc. 25**	**Perc. 50**	**Perc. 97.5**
1	0.8564	2.8326	0.4490	1.2803	5.7915	2.596	5.971	0.818	1.137	15.784
2	3.1539	7.3833	0.6834	1.3160	19.1935	1.621	3.591	0.837	1.131	9.748
3	0.6848	1.9801	0.8023	1.1699	3.1791	1.198	3.794	0.807	1.039	9.814
4	0.8249	4.5604	-1.1533	0.8760	9.9470	1.756	5.126	0.818	1.224	15.446
5	2.8929	7.3828	-1.6551	1.2395	17.6071	1.072	4.203	0.504	0.815	9.122
6	0.0616	3.2576	-1.6778	0.7650	5.6888	1.124	3.536	0.609	1.131	9.592
7	3.7859	8.2162	0.9011	1.5567	23.8860	2.473	4.958	0.881	1.182	14.047
8	2.4299	5.7996	0.7392	1.2203	15.6584	-0.206	2.795	-3.948	1.039	3.393
9	2.1041	5.3961	0.6965	1.4816	14.2894	1.045	4.171	0.214	1.064	10.044
10	3.2132	7.6341	0.6641	1.2824	19.9590	2.754	6.407	0.619	0.975	22.342
	n=500					**n=1000**				
1	2.286	4.419	0.826	1.190	14.699	1.768	5.786	0.141	1.195	18.779
2	1.498	4.049	0.789	1.112	11.758	1.498	4.636	0.633	1.123	13.052
3	1.672	5.316	0.615	0.984	14.659	2.122	5.848	0.793	1.250	19.762
4	1.861	4.669	0.765	1.173	15.465	1.697	5.562	0.496	1.106	17.866
5	1.522	4.095	0.713	1.135	12.787	1.884	5.607	0.570	1.206	18.457
6	2.995	5.474	1.018	1.303	17.512	1.869	5.179	0.842	1.263	17.676
7	2.751	5.982	0.930	1.254	20.567	2.110	5.663	0.819	1.147	17.679
8	1.505	4.144	0.894	1.194	14.397	1.827	5.576	0.704	1.199	19.482
9	2.655	5.226	0.900	1.319	15.721	1.758	4.831	0.714	1.115	15.651
10	1.491	4.712	0.712	1.070	16.496	2.362	5.137	0.943	1.315	18.571
	n=2000									
1	1.775	5.006	0.828	1.237	21.282					
2	1.367	4.643	0.656	1.150	15.417					
3	2.195	5.213	0.903	1.224	19.778					
4	1.947	5.266	0.706	1.193	17.684					
5	1.303	3.922	0.813	1.192	12.298					
6	1.224	4.688	0.644	1.126	15.854					
7	2.191	5.269	0.890	1.217	21.215					
8	2.511	5.484	0.898	1.321	21.489					
9	1.368	4.178	0.727	1.120	12.802					
10	2.544	6.189	0.706	1.168	19.749					

Cuadro A.4: (Censura al 20 % para n = 150, 250, 500, 1000 y 2000) Media, desviación estándar, percentil 25, percentil 50, percentil 97.5 para los Bootstraps de los residuales del modelo MRL-CI-MF-NA, para 10 réplicas.

	n=150					n=250				
Réplicas	**Media**	**Desv.Est**	**Perc. 25**	**Perc. 50**	**Perc. 97.5**	**Media**	**Desv.Est**	**Perc. 25**	**Perc. 50**	**Perc. 97.5**
1	4.6121	8.1275	1.1259	1.5354	22.0794	5.098	8.936	1.108	1.220	27.555
2	10.2710	11.1241	0.9707	9.9296	27.9518	3.562	8.853	0.829	1.302	29.395
3	5.9297	9.5330	1.1069	1.4080	21.6155	4.435	8.692	0.483	1.265	23.883
4	2.6366	7.0016	0.8627	1.2497	19.7532	2.460	5.085	1.174	1.283	14.292
5	2.9048	5.9119	0.7495	1.0026	16.5153	4.650	6.684	0.909	1.326	21.250
6	4.1453	10.4274	0.7818	1.1782	29.5975	4.675	7.854	0.908	1.252	23.889
7	2.1290	4.5680	0.7725	1.1579	13.3030	5.760	9.744	0.952	1.317	24.951
8	5.4422	8.8490	1.1755	1.5354	22.8872	6.937	9.013	1.265	1.708	27.642
9	4.3035	3.7995	1.1942	2.5535	11.6462	3.031	6.850	0.443	1.169	19.042
10	4.9660	8.0466	1.0171	1.3518	24.3382	4.978	8.646	1.053	1.346	25.952
	n=500					**n=1000**				
1	3.073	7.109	0.818	1.233	24.739	3.976	6.587	0.970	1.447	20.823
2	2.965	7.996	0.779	1.280	27.633	5.786	9.181	0.928	1.466	29.172
3	4.643	8.391	0.787	1.351	25.639	3.807	7.152	0.660	1.259	24.583
4	3.094	6.512	0.726	1.455	21.678	3.459	7.332	0.883	1.403	27.148
5	4.400	7.008	1.032	1.319	21.909	3.698	8.859	0.748	1.271	31.453
6	3.938	6.309	1.097	1.284	16.680	4.750	8.729	0.843	1.378	29.705
7	3.716	7.658	1.101	1.411	24.267	5.432	9.075	0.954	1.714	30.293
8	5.784	9.061	1.109	1.501	30.688	5.503	9.149	0.954	1.352	30.469
9	2.721	7.275	0.454	1.091	23.091	4.162	8.223	0.645	1.238	27.603
10	5.197	8.798	0.976	1.647	25.259	4.010	7.757	0.949	1.348	25.873
	n=2000									
1	4.460	7.670	0.950	1.393	26.224					
2	4.472	7.699	0.959	1.378	23.759					
3	4.462	7.773	0.943	1.485	25.602					
4	4.138	8.317	0.798	1.274	25.105					
5	4.161	8.095	0.851	1.263	27.218					
6	4.618	7.665	0.865	1.436	26.214					
7	3.979	7.463	0.845	1.253	23.302					
8	4.376	7.719	0.964	1.335	25.853					
9	3.994	7.767	0.885	1.354	26.266					
10	4.756	7.891	0.970	1.462	25.116					

Cuadro A.5: (Censura al 35 % para n = 150, 250, 500, 1000 y 2000) Media, desviación estándar, percentil 25, percentil 50, percentil 97.5 para los Bootstraps de los residuales del modelo MRL-CI-MF-NA, para 10 réplicas.

A.3. Error absoluto medio (MAE), raíz del error cuadrático medio (RMSE) y error cuadrático medio (MSE) del modelo general para el estudio de simulación

n	RMSE	MAE	MSE
150	3.150	2.311	9.920
250	3.218	2.327	10.359
500	3.154	2.247	9.948
1000	2.863	2.101	8.197
2000	2.939	2.133	8.639

Cuadro A.6: (Censura al 8 %) Medidas de precisión del modelo MRL-CI-MF-NA para diferentes tamaños de muestra

n	RMSE	MAE	MSE
150	5.901	3.548	34.820
250	5.791	3.420	33.535
500	5.550	3.384	30.804
1000	5.538	3.413	30.669
2000	5.439	3.290	29.581

Cuadro A.7: (Censura al 20 %) Medidas de precisión del modelo MRL-CI-MF-NA para diferentes tamaños de muestra

n	RMSE	MAE	MSE
150	9.897	6.037	97.951
250	9.346	5.723	87.348
500	9.197	5.502	84.587
1000	9.125	5.484	83.274
2000	9.350	5.619	87.424

Cuadro A.8: (Censura al 35 %) Medidas de precisión del modelo MRL-CI-MF-NA para diferentes tamaños de muestra

Bibliografía

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**: 716–723.

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

Azzalini, A. (1985). A class of distributions wich includes the normal ones, *Scandinavian Journal of Statistics* **12**: 171–178.

Bartolucci, F. y Scaccia, L. (2005). The use of mixtures for dealing with non-normal regression errors. computational statistics and data analysis, *Statistics and its Interface* **48(4)**: 821–834.

Basford, K., Greenway, D., McLachlan, G. y Peel, D. (1997). Standard errors of fitted component means of normal mixtures, *Computational Statistics* **12**: 1–18.

Basso, R., Lachos, V., Cabral, C. y Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions, *Computational Statistics and Data Analysis* **-**: 2926–2941.

Benites, L., Lachos, V. H. y Bolfarine, H. (2018). Regression modeling of censored data based on mixtures of scale mixture of normal distributions, *under review* .

Benites, L., Maehara, R., Lachos, V. H. y Bolfarine, H. (2019). Linear regression models with finite mixtures of skew heavy-tailed errors, *Chilean Journal of Statistics* **10**: 21–41.

Bogaertz, K., Komárek, A. y Lessafre, E. (2017). *A Practical Approach with Examples in R, SAS, and BUGS*, Chapman and Hall Book.

Branco, M. D. y Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions, *Journal of Multivariate Analysis* **79**: 99–114.

Caudill, S. (2012). A partially adaptive estimator for the censored regression model based on a mixture of normal distributions, *Journal of the Italian Statistical Society* **21**: 121–137.

Delyon, B., Lavielle, M. y Moullines, E. (1999). Convergence of a stochastic approximation version of the em algorithm, *The Annals of Statistics* **27**(1): 94–128.

Dempster, A., Laird, N. y Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society* **39**(1): 1–38.

Fernandez, C. y Steel, M. F. J. (1999). Multivariate student-t regression models: Pitfalls and inference, *Biometrika* **86**(1): 153–167.

Galarza, C., Lachos, V. y Bandyopadhyay, D. (2015). Quantile regression for linear mixed models: A stochastic approximation em approach, --, Universidade Estadual de Campinas, -.

Galarza, C., Lachos, V. y Bandyopadhyay, D. (2017). Quantile regression for linear mixed models: A stochastic approximation em approach, *Stat Interface* **-**: 471–482.

Garay, A. M., Lachos, V. H., Bolfarine, H. y Cabral, C. R. (2017). Linear censored regression models with scale mixtures of normal distributions, *Statistical Papers* **58**: 247–278.

Garay, A. M., Lachos, V. H., Bolfarine, H. y Lin, T. (2016). Nonlinear censored regression models with heavy-tailed distributions, *Statistics and its Interface* **9**: 281–293.

Kaufman, L. y Rousseeuw, P. (1990). *Finding Groups in Data*, New Wiley, New York.

Klein, J. P. y Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, New York. Springer.

Kuhn, H. y Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure, *ESAIM Probability and Statistics* **8**: 115–131.

McLachlan, G. J. y Krishnan, T. (2008). *The EM algorithm and extensions*, John Wiley & Sons, New Jersey.

Meza, C. y De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions, *Statistics and Computing* **22**: 121–139.

Mroz, T. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, *Econometrica* **55**: 765–799.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**: 461–464.

Thalita, d. B. M. (2016). *Robust estimation in regression models for censored data*, Master's thesis, Universidade Estadual de Campinas.

Thalita, d. B. M., Garay, A. M. y Lachos, V. H. (2017). Likelihood-based inference for censored linear regression models with scale mixtures of skew-normal distributions, *Journal of Applied Statistics* pp. 2039–2066.

Tobin, J. (1958). Estimation of relationships for limited dependent variables, *Econometrica* **26**: 24–36.

Vaida, F. (2005). Parameter convergence for em and mm algorithms, *Statistica Sinica* **15**: 831–840.

Wei, C. G. y Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association* **85**: 699–704. `https://www.jstor.org/stable/2290005`.

Willmott, C.J y Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate Research* **30**: 79–82.

Wu, C. F. (1983). On the convergence properties of the em algorithm), *The Annals of Statistics* **11**: 95–103.

Zeller, C. B., Cabral, C. R. B., Lachos, V. H. y Benites, L. (2018). Finite mixture of regression models for censored data based on scale mixtures of normal distributions, *Advances in Data Analysis and Classification* pp. 1–28.