PONTIFICIA UNIVERSIDAD
CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MODELO DE REGRESIÓN BINARIA ROBUSTA BAJO UN ENFOQUE BAYESIANO

Tesis para optar el grado académico de Maestro en Estadística que presenta:

Augusto Elmer Racchumí Vela

Asesor:

Cristian Luis Bayes Rodríguez

Lima, 2025

Informe de Similitud

Yo, Cristian Luis Bayes Rodríguez, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulada MODELO DE REGRESIÓN BINARIA ROBUSTA BAJO UN ENFOQUE BAYESIANO, de el autor Augusto Elmer Racchumí Vela, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 12%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 10 de marzo de 2025.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de investigación, 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 marzo del 2025.

Apellidos y nombres del asesor: Bayes Rodríguez, Cristian Luis	
DNI: 40372640	Firma 
ORCID: 0000-0003-0474-7921	

Dedicatoria

A mi padre Augusto Racchumí, por cada consejo brindado y por el esfuerzo que realizó para cuidarme y apoyarme en mis estudios, y a mi madre Zadith Vela, que a pesar de la distancia, siempre está a mi lado cuidándome.

A mi tía Angélica, a quien considero mi segunda madre y que siempre está pendiente y orando por mi bienestar.

A mis abuelitos, Manuel Vela y Edita Melendez, por las enseñanzas de vida, el amor y los valores brindados a cada miembro de la familia.

Agradecimentos

En primer lugar, agradezco a Dios, porque sin su presencia en nuestras vidas, no somos nada.

A mi asesor, el Doctor Cristian Bayes, por sus enseñanzas en la maestría, su orientación, y seguimiento durante la formulación del presente trabajo de investigación.

A mis amigos y hermanos, Iván Ocampo, Karen Barturén, Antony Niño y Jefrey Niño, por su paciencia y apoyo durante mi etapa académica y en la convivencia en el hogar.

A mis buenos amigos que conocí en la etapa laboral y de quienes aprendí mucho estando a su lado: Dr. Marcos Espinola Sánchez, Dr. Jorge Segundo Paredes y Dra. María Medina Pfluker, por su liderazgo y guía; Econ. Raquel Huamaní, por sus consejos y constante motivación; Dr. Cender Quispe, Bióloga Leny Sánchez y Bióloga Catherine Apaza, por su amistad y lealtad en los momentos difíciles.

Resumen

El estudio se origina con el propósito de contribuir a la comunidad científica mediante la descripción y aplicación de modelos bayesianos capaces de mitigar el efecto distorsionador de las variables cuantitativas sobre una variable respuesta binaria, especialmente en presencia de valores extremos y debido a una mala especificación de la función de enlace en el predictor lineal. El objetivo de esta investigación fue estudiar dichas propiedades de los modelos de regresión bayesianos robustas, utilizando simulaciones y casos con datos reales para evaluar su rendimiento comparado con un modelo tradicional.

Los modelos bayesianos robustos analizados incluyen los modelos Robit con enlace t-Student y parámetro de forma fijo, así como una variante en la que el parámetro de forma se estima a partir de los datos (Robit v aleatorio). Además, se examinó un modelo Robit con función de enlace t-Student Generalizada (Robit tG) y su variante con estimación del parámetro de forma basada en los datos observados (Robit tG v aleatorio). Estos modelos se estimaron utilizando el método Markov Chain Monte Carlo (MCMC) y el algoritmo de Gibbs Sampling para obtener muestras representativas de la distribución a posteriori. Las simulaciones y estimaciones de parámetros se llevaron a cabo con R y JAGS (Just Another Gibbs Sampler).

Los resultados obtenidos de las simulaciones y los casos de aplicación demuestran que los modelos de regresión robusta presentan una mejor recuperación de parámetros y ajuste a los datos en escenarios de contaminación con datos atípicos, en comparación con el modelo Probit de referencia. Los modelos más efectivos fueron aquellos con una función de enlace t-Student Generalizada, especialmente en situaciones de contaminación con datos atípicos.

Palabras-clave: Modelos bayesianos, Regresión binaria robusta, Regresión Robit, Regresión Robit tG.

Abstract

The study aims to contribute to the scientific community by describing and applying Bayesian models capable of mitigating the distorting effect of quantitative variables on a binary response variable, especially in the presence of extreme values and due to poor specification of the link function in the linear predictor. The objective of this research was to study the properties of robust Bayesian regression models, using simulations and real data cases to evaluate their performance compared to a traditional model.

The robust Bayesian models analyzed include the Robit models with a t-Student link and fixed shape parameter, as well as a variant where the shape parameter is estimated from the data (random shape Robit). Additionally, a Robit model with a Generalized t-Student link function (Robit tG) and its variant with shape parameter estimation based on observed data (random shape Robit tG) were examined. These models were estimated using the Markov Chain Monte Carlo (MCMC) method and the Gibbs Sampling algorithm to obtain representative samples of the posterior distribution. Simulations and parameter estimations were carried out with R and JAGS (Just Another Gibbs Sampler).

The results obtained from the simulations and application cases show that robust regression models have better parameter recovery and data fit in scenarios with outlier contamination compared to the reference Probit model. The most effective models were those with a Generalized t-Student link function, especially in situations of outlier contamination.

Keywords: Bayesian models, Robust binary regression, Robit regression, Robit tG regression.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En diversos campos de estudio, muchos investigadores buscan explicar o predecir resultados binarios con un conjunto finito de predictores cuantitativos o cualitativos, utilizando el modelo de regresión logística tradicional. Sin embargo, las estimaciones de máxima verosimilitud (EMV) de los coeficientes de regresión podrían estar sesgadas debido a la mala especificación de la función de enlace asociada al predictor lineal y a la probabilidad de éxito (Czado y Santner, 1992). Esto puede ocurrir debido a la presencia de valores atípicos o a un fuerte desequilibrio de clases en la variable objetivo. Si los datos presentan estas características particulares, la mala especificación del modelo de regresión binaria, en base a funciones comúnmente utilizadas como la logit, probit, log-log complementario, entre otros, puede generar un sesgo sustancial en las estimaciones de la respuesta media (Prentice (1976) y Wang y Dey (2010)).

Debido al problema de falta de robustez en las EMV en modelos de respuesta binaria, cuando se utilizan funciones de enlace inadecuadas, diversos estudios sugieren el uso de funciones de enlace basadas en la distribución t y sus generalizaciones, ya que son más resistentes ante valores extremos y desbalance de clases en la variable objetivo (Peyhardi, 2020).

Liu (2004) propone un modelo de regresión binaria robusto que utiliza la distribución-t simétrica como función de enlace, en lugar de la distribución normal utilizada en el modelo probit. Este modelo, llamado Robit, puede tener grados de libertad conocidos o desconocidos. El modelo Robit es una generalización de los modelos de regresión logística y probit, el cual proporciona una familia de modelos que incluyen estos casos como particulares para el análisis de respuesta binaria.

Se han encontrado referencias a otros estudios que exploran el uso de la distribución t de

Student para la estimación robusta en modelos de regresión en diversos contextos de aplicación, ya sea bajo el enfoque frecuentista como bayesiano. Entre los autores que han abordado este tema, se destacan los siguientes:Rubin (1983),West (1984), Lange, Little y Taylor (1989) y Liu y Rubin (1995), quienes demostraron que las estimaciones de los coeficientes de regresión obtenidas con estos modelos son más confiables ante la presencia de valores atípicos.

En base a los antecedentes mencionados, la presente investigación tiene como propósito aplicar el modelo de regresión Robit para respuesta binaria bajo un enfoque bayesiano, con el fin de cuantificar los efectos de diversos factores de gran relevancia en el diagnóstico de eventos de interés clínico. Los casos analizados presentan particularidades que aumentan la probabilidad de que las variables cuantitativas muestren un comportamiento heterogéneo dado a la presencia de valores atípicos, los cuales pueden afectar las estimaciones de los coeficientes de regresión. Para esta investigación, se considerarán dos casos de aplicación. El primero se centra en estimar el efecto de una variable explicativa numérica sobre la presencia de complicaciones en pacientes pediátricos con quemaduras severas. El segundo caso aplicativo, también se busca estimar el efecto de una variable explicativa numérica sobre la presencia de mortalidad de menores. En ambas aplicaciones, se compararán los efectos estimados de las principales variables explicativas tanto en presencia como en ausencia de datos considerados atípicos.

El estudio se enmarca en un enfoque cuantitativo y se basa en un diseño analítico de cohorte transversal.

1.2. Objetivos

1.2.1. Objetivo General

El principal objetivo de investigación consiste en estudiar las propiedades de un modelo de regresión binaria robusta bajo inferencia bayesiana y realizar la estimación del modelo con un conjunto de datos reales.

1.2.2. Objetivos Específicos

- Revisar en la literatura las diferentes propiedades de un modelo de regresión binaria robusta bajo su enfoque bayesiano.
- Realizar la implementación computacional mediante simulación MCMC del modelo de regresión binaria robusta bajo inferencia bayesiana para diferentes escenarios que contengan datos atípicos.

- Realizar comparaciones de las estimaciones de los coeficientes de regresión a partir de los modelos propuestos contra un modelo de referencia.
- Aplicar los modelos de regresión binaria robusta a un conjunto de datos reales.

1.3. Organización del trabajo

En el Capítulo 2, se presenta una descripción concisa de los conceptos y propiedades fundamentales de la distribución t de Student, así como de algunas de sus generalizaciones. Este apartado ofrece un conocimiento básico que será de gran utilidad en el modelado de datos cuando se presenten valores atípicos. Además, se aborda los aspectos clave que permitirán comprender y aplicar adecuadamente esta distribución en diversas situaciones estadísticas.

En el capítulo 3 se presenta las propuestas de varios modelos de regresión robusta para el análisis de enlaces binarios, utilizando un enfoque bayesiano. Se detalla el algoritmo de simulación a utilizar, así como la función de verosimilitud, la distribución a priori y la distribución a posteriori correspondiente a los modelos propuestos. Además, se abordan los métodos para diagnosticar la convergencia de la muestra a posteriori y se analiza indicadores para la comparación de modelos. Esta sección brinda una comprensión completa y detallada de los modelos, así como las herramientas necesarias para evaluar su rendimiento.

En el capítulo 4, se presenta un estudio de simulación con el objetivo de examinar el impacto de los valores atípicos en la estimación de los coeficientes de regresión para los modelos bayesianos: probit, robit con enlace t-Student con parámetro de forma fijo y parámetro de forma aleatorio, y robit con enlace t-Generalizado con parámetro de forma fijo y parámetro de forma aleatorio. Se simuló varios conjuntos de datos con diferentes tamaños de muestra donde se introdujo una covariable contaminada con valores atípicos, y se analizó cómo estos valores afectan dichas estimaciones. Luego, se realizó una comparación de las estimaciones obtenidas mediante los modelos bayesianos propuestos respecto al modelo bayesiano probit, al que se consideró como el modelo de referencia. Este estudio permitió evaluar la robustez de los modelos bayesianos propuestos frente a la presencia de estos valores atípicos y determinar si estos modelos proporcionan estimaciones más precisas y confiables en comparación al modelo de referencia.

En el capítulo 5, se presenta la aplicación de los modelos propuestos utilizando datos reales de un contexto clínico. Se analizaron dos estudios de caso: el primero se centró en la estimación de los efectos asociados a la presencia de complicaciones en pacientes pediátricos que sufrieron quemaduras de segundo y tercer grado, mientras que el segundo se enfocó en

la estimación de los efectos asociados a la mortalidad de pacientes pediátricos. Los datos corresponden a pacientes atendidos en un Instituto de Salud de tercer nivel en Perú.

Por último, en el capítulo 6 se presentan y discuten los resultados obtenidos a partir de la investigación realizada, donde se examinan tanto los aspectos positivos como las limitaciones de los modelos propuestos.

Capítulo 2

Distribución t generalizada

La teoría clásica de inferencia estadística se basa en la suposición de normalidad e independencia de los errores. No obstante, en muchas aplicaciones del mundo real, se ha observado que este comportamiento no se cumple en diversas variables. Es más común encontrar distribuciones con colas más pesadas en comparación con la distribución normal, lo cual puede afectar las inferencias realizadas (Ahsanullah, Kibria y Shakil, 2014).

En ese sentido, este capítulo tiene como objetivo presentar de manera concisa las principales características de la distribución t de Student y la t de Student Generalizada (t-Generalizada), las cuales sirven como base para modelar respuestas binarias controlando el efecto de posibles perturbaciones en los datos cuantitativos. A continuación, se examinará la función de densidad de probabilidad, función de distribución acumulada y sus propiedades relacionadas con la esperanza y varianza.

2.1. Distribución t de Student

La distribución t de Student fue desarrollada por Willieam S. Gosset en 1908 en su trabajo titulado: "The probable error of a mean", bajo el seudónimo de Student. Esta distribución forma parte de una familia de distribuciones de probabilidad continua y se destaca por su forma de campana simétrica en su función de densidad de probabilidad, la cual varía según el tamaño de la muestra. La distribución t de Student estándar se caracteriza por tener una media $\mu = 0$ y una desviación estándar $\sigma > 1$ para grados de libertad mayores a 2, mientras que, no existe para grados de libertad de 1 y 2. Conforme el tamaño de muestra tiende hacia infinito ($n \to \infty$), la distribución t se aproxima a una distribución normal estándar (Ahsanullah et al., 2014).

Las propiedades de la distribución t ofrecen una extensión valiosa en comparación con la

distribución normal, especialmente en el modelado estadístico de datos que presentan errores con colas más largas a causa de los valores atípicos. Los grados de libertad (v) de la distribución t proporcionan una dimensión flexible para realizar inferencias estadísticas más robustas, aunque esto puede variar según la complejidad computacional requerida por los diferentes modelos utilizados (Lange, Little y Taylor, 1989).

2.1.1. Función de densidad de probabilidad

Sea X una variable aleatoria con distribución t de Student con v grados de libertad, se tiene que su función de densidad de probabilidad está dada por (Li y Nadarajah, 2020).

$$f_v(x) = \frac{\Gamma(\frac{v+1}{2})}{(\pi v)^{1/2}\Gamma(\frac{v}{2})}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}} \qquad x \in \mathbb{R},\ \ v > 0, \tag{2.1}$$

donde $\Gamma(.)$ denota a la función gamma definida de la siguiente forma:

$$\Gamma(a) = \int_0^\infty t^{a-1} exp(-t)\, dt.$$

Dada a las propiedades de la distribución t-Student, se conoce que a medida que los grados de libertad ($v \to \infty$), la función de densidad de la t-Student se aproxima a la distribución normal. En la Figura 2.1 se muestra la comparación de la densidad de una $t(v = 3)$, $t(v = 10)$ y $t(v = 15)$, respecto a una normal estándar $N(0, 1)$, en ella se muestra que la densidad de la distribución $t(v = 15)$ (con mayor grado de libertad), está más próxima a la normal.

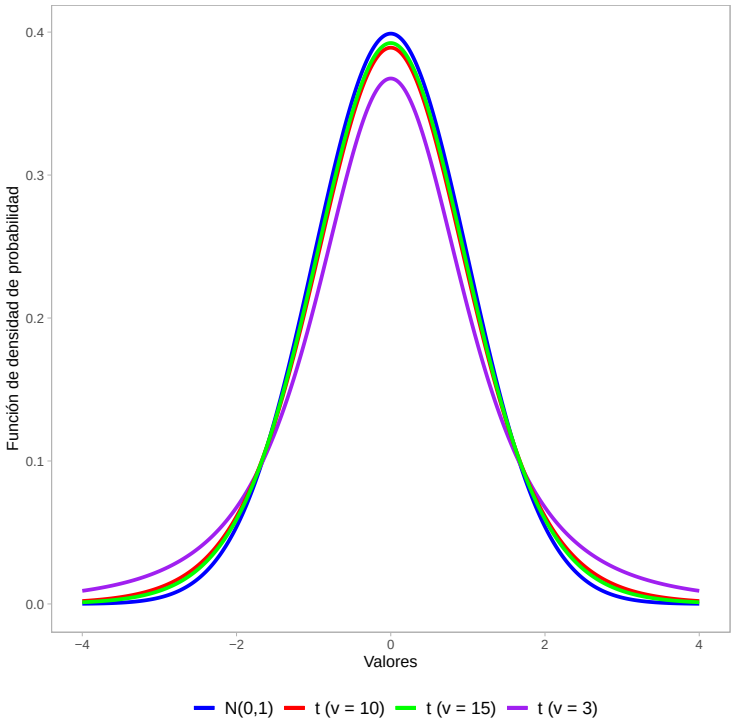

Figura 2.1: Función de densidad de la distribución t-Student.

Luego, su función de distribución acumulada es dada por:

$$F_v(x) = \frac{1}{2} + x\Gamma\left(\frac{v+1}{2}\right) \frac{{}_2F_1(\frac{1}{2}, \frac{v+1}{2}; \frac{3}{2}; \frac{-x^2}{v})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}, \tag{2.2}$$

donde ${}_2F_1$ denota la función hipergeométrica de Gauss definida por:

$${}_2F_1(a, b; c; x) = \sum_{k=0}^{\infty} \frac{(a)^k (b)^k}{(c)^k} \frac{x^k}{k!},$$

mientras que, la expresión $(e)^k$ puede ser expresado como un factorial ascendente $(e)^k = e(e+1)\dots(e+k-1)$. Sin embargo, a partir del estudio realizado por Schlüter y Fischer (2012), es posible realizar una aproximación de un quantil de la distribución t-Student $F(x_\alpha) = 1-\alpha$ a partir de la siguiente ecuación:

$$x_\alpha \approx \left[\frac{(v/2)^{v/2}\Gamma(\frac{v-1}{2})}{\alpha\sqrt{2\pi}2^{\frac{1-v}{2}}\Gamma(\frac{v}{2})} \right].$$

La Figura 2.2 muestra una comparación de las funciones de distribución acumulada de la t-Student con diferentes grados de libertad, en contraste con la normal estándar. Conforme aumenta el número de grados de libertad, como en $t(v = 15)$, la función de distribución acumulada de la t-Student tiende a asemejarse más a la de $N(0, 1)$.

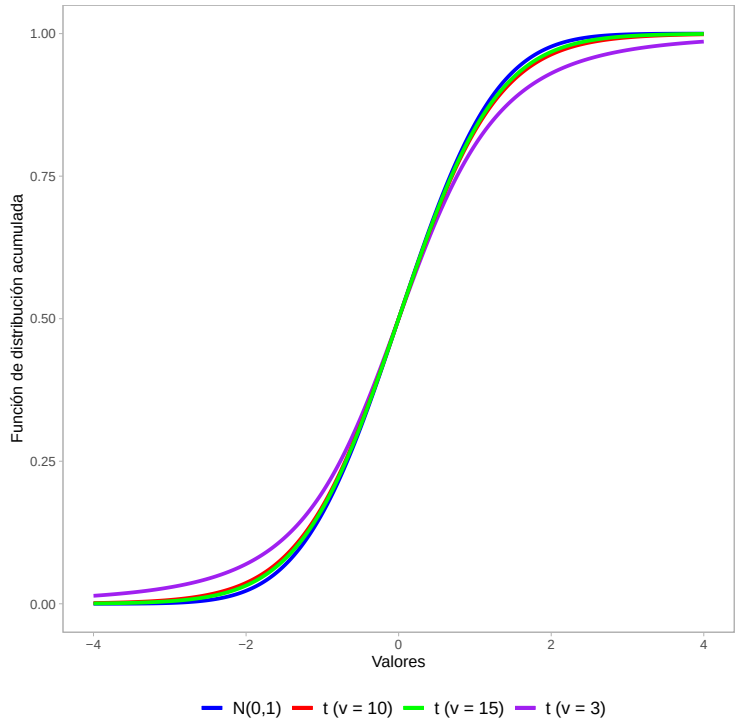

Figura 2.2: Función de distribución acumulada de la t-Student.

2.1.2. Propiedades

Sea X una variable aleatoria con distribución t-Student con v grados de libertad, entonces, la media, varianza, asimetría y curtosis se definen como:

$$E(X) = \begin{cases} 0, \text{ para } v > 1 \\ \text{no definido, otros casos.} \end{cases} \tag{2.3}$$

$$V(X) = \begin{cases} \frac{v}{v-2}, \text{ para } v > 2 \\ +\infty, \text{ para } 1 < v \leq 2. \end{cases} \tag{2.4}$$

$$Asimetria(X) = \begin{cases} 0, \text{ para } v > 3 \\ \text{no definido, otros casos.} \end{cases} \tag{2.5}$$

$$Curtosis(X) = \begin{cases} 3 + \frac{6}{v-4}, \text{ para } v > 4 \\ +\infty, \text{ para } 2 < v \leq 4. \end{cases} \tag{2.6}$$

2.2. Distribución t Generalizada

Por otro lado, la versión generalizada de la distribución t-Student (t-Generalizada), está compuesta por dos parámetros. El primer parámetro controla la pesadez de las colas, mientras

que, el segundo parámetro controla la escala. La distribución t-Generalizada, utilizada en la función de enlace para el predictor lineal en un modelo de regresión, ofrece dos ventajas significativas, en comparación con el enlace t-Student. En primer lugar, un enlace t-Generalizado simétrico, con un parámetro de forma desconocido es más fácil de identificar en comparación con un enlace t-Student que tiene grados de libertad desconocidos y un parámetro de escala conocido. En segundo lugar, los enlaces t-Generalizados asimétricos, que tienen tanto el parámetro de forma como el de escala desconocidos, proporcionan modelos de regresión con enlaces asimétricos que son más flexibles y mejores en comparación con los enlaces sesgados existentes (Kim, Chen y Dey, 2008).

En la literatura se han propuesto numerosas generalizaciones de la distribución t, debido a los resultados prometedores que ofrecen para el manejo de datos atípicos, especialmente en términos de control de la dispersión y la curtosis de la distribución (Li y Nadarajah, 2020). Es por esta razón que la distribución t y sus generalizaciones se han convertido en modelos ampliamente populares para el análisis de datos, por lo que es una propuesta interesante para su implementación y aplicación en modelos de respuesta binaria.

2.2.1. Función de densidad de probabilidad

Kim et al. (2008) hace referencia a la función de densidad de la distribución t-Generalizada introducido por Arellano-Valle y Bolfarine (1995), donde X es una variable aleatoria continua, v_1 es el parámetro de forma (grados de libertad) y v_2 es el parámetro de escala, y su función de densidad de probabilidad tiene la siguiente forma:

$$f_{v_1,v_2}(x) = \frac{\Gamma\left(\frac{v_1+1}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)} \frac{1}{\sqrt{v_2\pi}} \left(1 + \frac{x^2}{v_2}\right)^{-\frac{v_1+1}{2}} \qquad x \in \mathbb{R}, \quad v_1, v_2 > 0. \tag{2.7}$$

En la Figura 2.3 se presenta una comparación entre las funciones de densidad de la distribución t-Generalizada, con parámetros fijos y variables, respecto a una distribución normal estándar $N(0,1)$. En la Figura 2.3.A) se muestran diferentes densidades de la t-Generalizada con el parámetro de forma fijo en $v_1 = 1$, mientras que, el parámetro de escala varía en $v_2 \in (1,2,3)$. Se observa que a medida que aumenta el parámetro de escala, la curtosis tiende a disminuir, volviéndose más platicúrtica y alejándose de la distribución $N(0,1)$. Por otro lado, en la Figura 2.3.B), manteniendo el parámetro de escala fijo en $v_2 = 1$ y variando el parámetro de forma $v_1 \in (1,2,3)$, no se aprecia una variación significativa en la curtosis, con una tendencia a aproximar la densidad a la distribución $N(0,1)$.

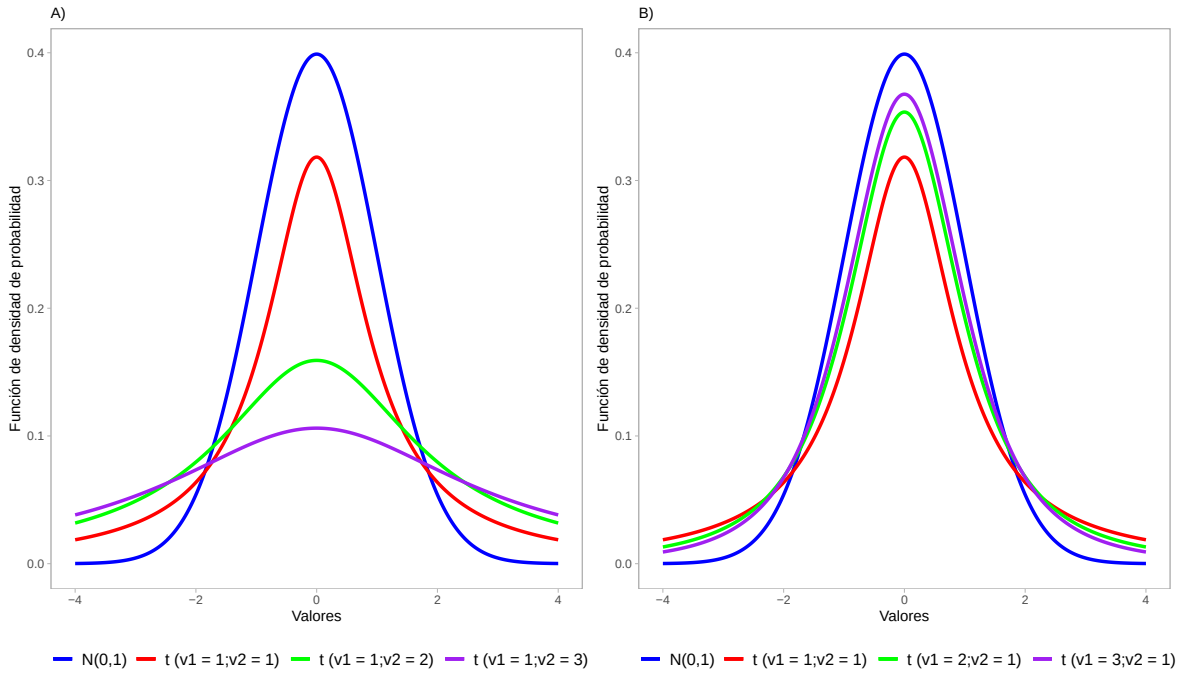

Figura 2.3: Función de densidad de probabilidad para la t-Generalizada con parámetro de forma fijo $v_1 = 1$ y parámetro de escala variable $v_2 \in (1, 2, 3)$ en A) y parámetro de forma variable $v_1 \in (1, 2, 3)$ y parámetro de escala fijo $v_2 = 1$ en B)

Mientras que, la función de distribución acumulada para la t-Generalizada está dada por:

$$F_{v_1,v_2}(x) = \int_{-\infty}^{x} f_{v_1,v_2}(u)du. \tag{2.8}$$

2.2.2. Propiedades

La distribución t-Generalizada presenta simetría alrededor de cero y una forma similar a una campana, al igual que la distribución t-Student estándar. Cuando el parámetro de forma, v_1, es pequeño, la distribución se caracteriza por tener colas pesadas. Luego, sea X una variable aleatoria continua con distribución t-Generalizada $X \sim tG(v_1, v_2)$, entonces, sus propiedades son las siguientes:

- $E(X) = 0$ para $v_1 > 1$.
- $Var(X) = \frac{v_2}{v_1 - 2}$ para $v_1 > 2$.
- Cuando $v_1 = v_2 = v$, la distribución se reduce a una distribución t-Student con v grados de libertad.
- Cuando $v \to \infty$, la distribución t-Student converge a la distribución normal.

- Para $\mu = 0$ y $\sigma = 1$, la distribución t-Generalizada se reduce a una distribución t-Student estándar.

Al realizar una comparación entre las principales propiedades de la función de distribución de probabilidad de la t-Generalizada y la distribución t-Student estándar, se observa una gran similitud en su forma de campana simétrica centrada en cero. Ambas exhiben colas más pesadas a medida que los grados de libertad (v_1) disminuyen. Sin embargo, al introducir el segundo parámetro, en este caso, el de escala (v_2), los efectos generados por los dos parámetros de la t-Generalizada, añade nuevas características beneficiosas para modelar respuestas binarias.

En el estudio de simulación realizado por Kim et al. (2008), se exploraron dos escenarios comparativos interesantes. En el primero, los parámetros de forma y escala tomaron valores de $v_1 = 1{,}2$ y $v_2 = 2$ respectivamente, lo que hizo que la función de densidad de la t-Generalizada intersectara con la densidad de una distribución $N(0, 1)$ en un punto, mientras que, la función de distribución acumulada de la t-Generalizada nunca se cruzó con la de la $N(0, 1)$. En el segundo escenario, los parámetros de forma y escala fueron $v_1 = 2$ y $v_2 = 1$, respectivamente, lo que resultó en que la función de densidad de la t-Generalizada intersectara la densidad de la $N(0, 1)$ en dos puntos, mientras que, su función de distribución acumulada intersectara en un punto con la de la $N(0, 1)$. Kim et al. (2008) demostró que el segundo escenario de modelado era superior, lo que resalta la influencia significativa de las diferentes formas que puede tomar la función de distribución acumulada de la t-Generalizada en la modelización de datos binarios.

Un punto importante mencionado por los autores es que la función de distribución acumulada de la t-Student nunca cruza la de la $N(0, 1)$. Esto significa que la velocidad a la que la probabilidad de una respuesta binaria con enlace t-Student se acerca a 0 o 1 nunca supera la del enlace probit. Por otro lado, la velocidad de la probabilidad en la distribución t-Generalizada con parámetros $v_1 = 2$ y $v_2 = 1$ es más rápida al principio para acercarse a 1, luego se ralentiza, en comparación con el enlace probit. En consecuencia, la ventaja de utilizar una función de enlace t-Generalizada en lugar de una t-Student, es que el enlace t nunca supera al enlace probit en términos de la velocidad con la que la probabilidad se acerca a 0 o 1, además de su mejor capacidad para manejar datos atípicos en conjuntos de datos reales. Según lo mencionado, esto implica que al estimar la probabilidad de éxito sin la presencia de casos atípicos, el modelo Probit, en teoría, tendría un mejor ajuste a los datos en comparación con los modelos con enlace tG y t (siendo este último el que presenta el menor ajuste a los datos). En un escenario de simulación, esto sugiere que el modelo Probit tendría un menor

sesgo al recuperar los parámetros. Sin embargo, al introducir perturbaciones en el predictor, las estimaciones obtenidas por los modelos tG serían superiores en términos de ajuste a los datos, en comparación con los demás modelos.

Según lo comentado previamente, en la Figura 2.4 se presentan los resultados de simulación obtenidos por Kim et al. (2008). En la Figura 2.4.A), se observa que la densidad de la distribución t-Generalizada con $v_1 = 2$ y $v_2 = 1$ intersecta la densidad de la distribución $N(0, 1)$ en dos puntos, a diferencia de la distribución $t(v = 2)$, que solo se intersecta en un punto. Por otro lado, en la Figura 2.4.B), la distribución acumulada de la t-Generalizada con $v_1 = 2$ y $v_2 = 1$ se intersecta con la distribución acumulada de $N(0, 1)$ en un punto, aproximándose de forma más rápida a 1, mientras que esto no ocurre con la distribución acumulada de $t(v = 2)$.

Con base en estos hallazgos, tanto en la fase de simulación como en la aplicación del presente estudio, se emplearán estos valores de los parámetros de forma y escala para los modelos bayesianos Robit y Robit tG. Estos valores se utilizarán para estimar los coeficientes de regresión y luego comparar los resultados con los obtenidos mediante el modelo de referencia Probit bayesiano.

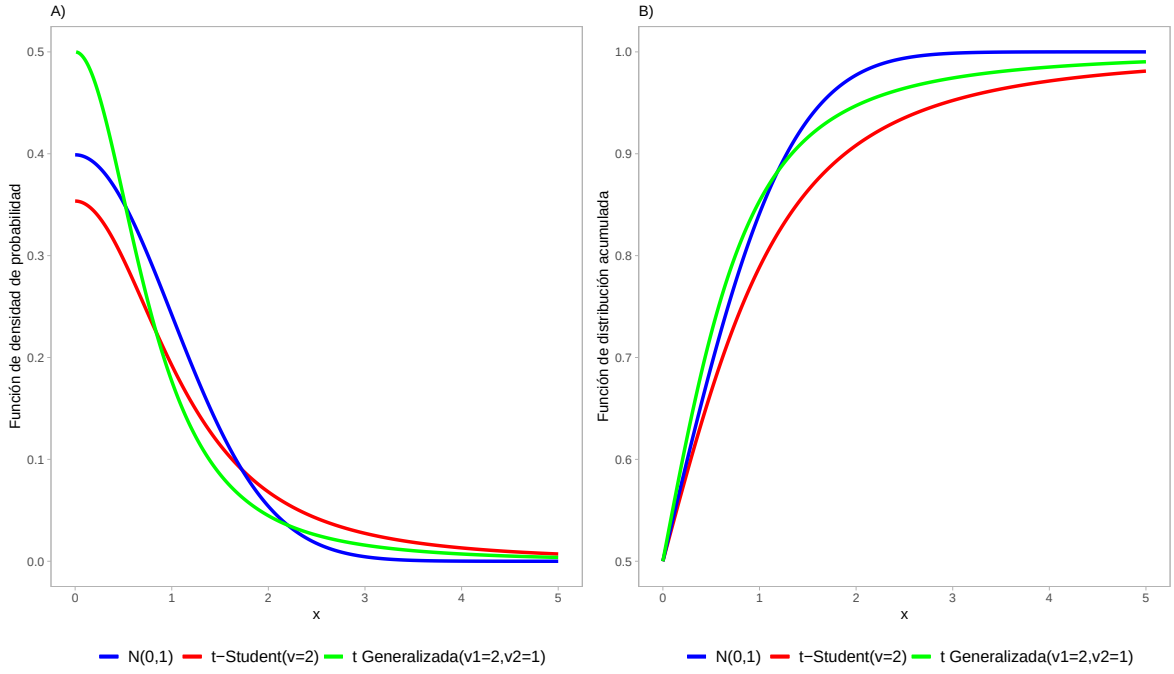

Figura 2.4: Función de densidad de probabilidad en A) y función de distribución acumulada en B) de la t-Student, t-Generalizada y N(0,1).

Capítulo 3

Modelo de Regresión Binaria Robusta

Los modelos de regresión binario robustos son técnicas utilizadas en el modelamiento estadístico para la explicación o predicción de una variable respuesta dicotómica a partir de p covariables. El objetivo de estos modelos es obtener estimaciones más precisas de los parámetros del modelo de regresión en presencia de valores atípicos u otras anomalías en los datos del mundo real, los cuales tienen un efecto distorsionador significativo en la media y la varianza. A diferencia de los modelos clásicos como logit y probit, los modelos robustos emplean estimadores alternativos que son menos sensibles a las distorsiones que puedan presentar los datos. Esto asegura que las estimaciones de los parámetros del modelo no se vean sesgadas por la influencia de valores atípicos, brindando resultados más confiables para la toma de decisiones (Li, Liao, Tang, Li, Li y Xu, 2023).

Este enfoque alternativo a los modelos tradicionales busca ajustar las observaciones atípicas a una distribución robusta. Estas distribuciones tienen colas más pesadas en su función de densidad, lo que significa que disminuyen lentamente hacia cero en comparación con las colas de una distribución normal. Algunas distribuciones paramétricas que presentan colas pesadas son la distribución de Weibull, la distribución log-normal, la distribución t de Student, entre otras. La distribución t de Student es especialmente adecuada para este problema debido a sus propiedades de cola pesada. Estas propiedades varían según los grados de libertad, lo que le confiere una mayor robustez frente a la presencia de valores atípicos (Li et al., 2023).

3.1. Modelo de regresión robusta t-Student

Con el objetivo de abordar el desafío de estimar parámetros en un modelo de regresión, especialmente cuando existe una mala especificación en la función de enlace de los modelos logit y probit debido a la presencia de valores atípicos, se propone un enfoque del modelo

bayesiano robusto para un resultado binario llamado robit. Este modelo aprovecha las propiedades de la distribución t-Student simétrica que presenta colas pesadas, el cual se incorpora en su función de enlace del predictor lineal. De esta manera, el modelo robit minimiza el impacto de los valores atípicos en la estimación de los parámetros, ofreciendo una solución más robusta y confiable.

Para el presente estudio, se considera utilizar la siguiente igualdad de los parámetros: $v_1 = v_2 = v$, a partir de las propiedades de la t-Generalizada mencionada por Kim et al. (2008). Debido a esto, la densidad de la t-Generalizada se reduce a una t-Student estándar descrita por Liu (2004).

En general, se supone que y_i es la variable respuesta $i = 1, 2, ..., n$, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ es el vector de covariables de p dimensiones, $y_i = (y_1, y_2, ..., y_n)^T$ es el vector de respuestas binarias observadas, y $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ es el vector p-dimensional de coeficientes de regresión. Luego, basado en un enfoque de variable latente, el modelo de regresión binario es dado por:

$$y_i = \begin{cases} 1, \text{ para } w_i > 0 \\ 0, \text{ para } w_i \leq 0, \end{cases} \tag{3.1}$$

donde la variable latente se define como:

$$w_i = x_i^T \beta + \epsilon_i, \tag{3.2}$$

y la variable aleatoria $\epsilon_i \sim F_v$, siendo F_v la función de distribución acumulada de la t-Student estándar. Entonces, para hacer referencia al modelo de regresión robit, se utilizará la notación $robit(v)$, el cual describe al modelo robusto con v grados de libertad.

$$\begin{aligned} P(y_i = 1|w_i > 0) &= 1 - P(y_i = 0|w_i \leq 0) = F_v(w_i) \\ P(y_i = 1|x_i^T, \beta) &= 1 - P(y_i = 0|x_i^T, \beta) = F_v(x_i^T \beta) \quad (i = 1, \ldots, n). \end{aligned} \tag{3.3}$$

Dada a las propiedades de este modelo robusto, es posible aproximar los modelos logit y probit utilizando el enlace robit con grados de libertad adecuados. Investigaciones previas han demostrado que el enlace probit puede aproximarse utilizando valores altos de los grados de libertad en un modelo $robit(v)$, mientras que, para lograr una buena aproximación del enlace logit, se requieren sólo siete grados de libertad (Liu (2004) y Kim et al. (2008)).

3.1.1. Función de verosimilitud

Para describir el modelo robit, supongamos que $Y = (Y_1, Y_2, ..., Y_n)^T$ es un vector de n variables aleatorias binarias independientes tales que: $P(Y_i = 1) = F_v(x_i^T\beta)$, donde $F_v(.)$ es la función de distribución acumulada de la t-Student univariante con grados de libertad conocidos y fijos v (2,2), mientras que, $x_i, i = 1, 2, 3, ..., n$ es el vector de dimensiones $p \times 1$ de covariables en el modelo, y β el vector de parámetros desconocidos del modelo. Entonces, la función de verosimilitud tiene la siguiente densidad de probabilidad (Roy, 2012):
Sea:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= F_v(x_i^T\beta), \end{aligned} \tag{3.4}$$

entonces, la verosimilitud está dada por:

$$\begin{aligned} p(y \mid \theta) &= \prod_{i=1}^{n} (\pi_i)^{y_i} (1-\pi_i)^{1-y_i} \\ p(y \mid \theta) &= \prod_{i=1}^{n} \left(F_v(x_i^T\beta)\right)^{y_i} \left(1 - F_v(x_i^T\beta)\right)^{1-y_i} \\ p(y \mid \theta) &= \prod_{i=1}^{n} \left(\frac{1}{2} + (x_i^T\beta) \times \Gamma\left(\frac{v+1}{2}\right) \frac{{}_2F_1(\frac{1}{2}, \frac{v+1}{2}; \frac{3}{2}; \frac{-(x_i^T\beta)^2}{v})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}\right)^{y_i} \\ &\quad \times \left(1 - \left(\frac{1}{2} + (x_i^T\beta) \times \Gamma\left(\frac{v+1}{2}\right) \frac{{}_2F_1(\frac{1}{2}, \frac{v+1}{2}; \frac{3}{2}; \frac{-(x_i^T\beta)^2}{v})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}\right)\right)^{1-y_i}, \end{aligned} \tag{3.5}$$

donde y son los datos observados y $\theta = (\beta^T, v)^T$.

3.1.2. Función de distribución a priori

Para el contexto de esta investigación, se asumirá que la distribución a priori de los coeficientes β en el modelo de regresión robit bayesiano, tendrán una distribución normal de la forma:

$$\beta_j \overset{ind}{\sim} N(0, \sigma_\beta^2), \tag{3.6}$$

donde $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ es el vector p-dimensional de coeficientes de regresión independientes, en consecuencia, la función de distribución a priori de los coeficientes de regresión también son independientes. Por otro lado, la función de distribución a priori para el pará-

metro de forma (v) en el modelo, se asumirá como una distribución exponencial desplazada, según la propuesta de Ross (2022a), con una modificación en la forma de esta distribución, dada a partir de la siguiente expresión:

$$\begin{aligned} v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\ u &\sim U(1, 100), \end{aligned} \tag{3.7}$$

donde U es la distribución uniforme, cuyos parámetros varían en un rango de 1 a 100, luego, la distribución a priori conjunta para todos los parámetros para el modelo robit(v) está dada por:

$$\pi(\theta) = \pi(\beta_1) \times \pi(\beta_2) \times \cdots \times \pi(\beta_p) \times \pi(v). \tag{3.8}$$

3.1.3. Función de distribución a posteriori

Entonces, a partir de las ecuaciones descritas en (3.5) y en (3.8), se tiene que la función de distribución a posteriori para el modelo $robit(v)$ con la t-Student estándar se define como:

$$\begin{aligned} \pi(\theta \mid y) &\propto p(y \mid \theta)\pi(\theta) \\ \pi(\theta \mid y) &\propto \prod_{i=1}^{n} (\pi_i)^{y_i} (1-\pi_i)^{1-y_i} \times \pi(\theta) \\ \pi(\theta \mid y) &\propto \prod_{i=1}^{n} \left(F_v(x_i^T\beta)\right)^{y_i} \left(1 - F_v(x_i^T\beta)\right)^{1-y_i} \times \pi(\theta), \end{aligned} \tag{3.9}$$

donde y son los datos observados y $\theta = (\beta^T, v)^T$.

Encontrar la solución analítica a la expresión en (3.9) resulta ser compleja, por lo tanto, se optará por emplear el método MCMC para obtener muestras que representen la distribución a posteriori de manera más efectiva.

3.2. Modelo de regresión robusta t-Student Generalizada

Siguiendo con las definiciones presentadas en la investigación realizada por Kim et al. (2008), el objetivo central de este trabajo es destacar que las particularidades de la distribución t-Generalizada superan a las de la t-Student convencional, principalmente debido a la influencia adicional proporcionada por el parámetro de escala. Esta elección se basa en

la necesidad de mejorar la precisión de los modelos, y como se ha señalado anteriormente, emplear la distribución t-Generalizada como función de enlace conlleva ventajas significativas para captar de manera más efectiva patrones complejos en la relación entre las variables predictoras y la variable respuesta.

La flexibilidad añadida por la elección de la distribución t-Generalizada permite una mejor adaptación a las relaciones subyacentes en los datos. Como resultado, esta elección contribuye de manera sustancial a mejorar la precisión del modelo.

Al igual que en el modelo robit, se considera la siguiente forma general para describir modelo de regresión binaria, que toma como enfoque de una variable latente a w_i. En este contexto, $y_i = (y_1, y_2, ..., y_n)^T$ denota el vector de respuestas binarias observadas para $i = 1, 2, ..., n$. Por otro lado, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ representa el vector de covariables de p dimensiones, mientras que $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ es el vector p-dimensional de coeficientes de regresión. El modelo más sencillo sería el siguiente:

$$y_i = \begin{cases} 1, \text{ para } w_i > 0 \\ 0, \text{ para } w_i \leq 0, \end{cases} \tag{3.10}$$

donde la variable latente se define como:

$$w_i = x_i^T \beta + \epsilon_i. \tag{3.11}$$

Bajo este nuevo escenario, se considera que la variable aleatoria ϵ_i tiene la siguiente distribución, $\epsilon_i \sim F_{v_1,v_2}$, siendo F_{v_1,v_2} la función de distribución acumulada de la distribución t-Generalizada definida como:

$$F_{v_1,v_2}(w) = \int_{-\infty}^{w} f_{v_1,v_2}(u)du, \tag{3.12}$$

donde f_{v_1,v_2} representa la función de densidad de la t-Generalizada con v_1 y v_2 grados de libertad, como se define en la ecuación (2.7). Cuando asumimos que el parámetro de forma es igual al parámetro de escala en la función de densidad de la t-Generalizada, entonces, la función de distribución acumulada de la t-Generalizada se iguala a la distribución acumulada de la t-Student estándar, y la ecuación definida en (3.10) se simplifica a los enlaces robit, tal como se describe en la ecuación (2.7).

Como se mencionó, el parámetro v_1 en el modelo robit controla la pesadez de la cola en la función de enlace y, al mismo tiempo, está relacionado con la escala de la variable latente w_i. Cuando v_1 toma valores grandes, los coeficientes del modelo en valor absoluto, se vuelven

pequeños. Sin embargo, la limitante es que v_1 carece de una interpretación sencilla y también provoca una convergencia lenta en el algoritmo de Gibbs utilizado para obtener las muestras a posteriori. Para resolver el problema de convergencia lenta, los autores proponen la una reparametrización, y tomando como parámetro de forma $v_1 = v$ y el parámetro de escala fijo en $v_2 = 1$ tal como se muestra a continuación:

$$y_i = \begin{cases} 1, \text{ si } w_i/\sqrt{v} > 0 \\ 0, \text{ si } w_i/\sqrt{v} \leq 0, \end{cases}$$

donde la variable latente está definida como:

$$w_i/\sqrt{v} = x_i^T(\beta/\sqrt{v}) + \epsilon_i/\sqrt{v}, \tag{3.13}$$

entonces, dada a la reparametrización propuesta por los autores, la nueva variable latente se define como: $w_i^* = w_i/\sqrt{v}$, $\beta^* = \beta/\sqrt{v}$ y $\epsilon_i^* = \epsilon_i/\sqrt{v}$, y en consecuencia, se obtiene un nuevo modelo de variable latente equivalente a:

$$y_i = \begin{cases} 1, \text{ si } w_i^* > 0 \\ 0, \text{ si } w_i^* \leq 0, \end{cases} \tag{3.14}$$

luego se tiene que la variable latente es $w_i^* = x_i^T\beta^* + \epsilon_i^*$, y el nuevo error sigue una distribución t-Generalizada según la ecuación definido en (2.7) dada la siguiente notación: $\epsilon_i^* \sim f_{v_1=v,v_2=1}$. El modelo presentado en la ecuación (3.14) se conoce como el modelo de enlace t-Generalizado simétrico, y su reparametrización contribuye a mejorar la eficiencia del algoritmo de Gibbs utilizado.

3.2.1. Función de verosimilitud

Para definir a la función de verosimilitud del modelo robit con enlace t-Generalizada, se tiene que, Y es una variable aleatoria que se distribuye mediante un ensayo de Bernoulli dada por la siguiente forma:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= F_v\left(x_i^T\beta\sqrt{v}\right), \end{aligned} \tag{3.15}$$

luego, la función de verosimilitud asociada al modelo robit con enlace t-Generalizado simétrico y a la variable latente definida en (3.14), es dada por:

$$
\begin{aligned}
p(y \mid \theta) &= \prod_{i=1}^{n} (\pi_i)^{y_i} (1-\pi_i)^{1-y_i} \\
p(y \mid \theta) &= \prod_{i=1}^{n} \left[F_v\left(x_i^T\beta \times \sqrt{v}\right)\right]^{y_i} \left[1 - F_v\left(x_i^T\beta \times \sqrt{v}\right)\right]^{1-y_i} \\
p(y \mid \theta) &= \prod_{i=1}^{n} \left[\frac{1}{2} + \left(x_i^T\beta \times \sqrt{v}\right) \times \Gamma\left(\frac{v+1}{2}\right) \frac{{}_2F_1(\frac{1}{2}, \frac{v+1}{2}; \frac{3}{2}; \frac{-\left(x_i^T\beta \times \sqrt{v}\right)^2}{v})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}\right]^{y_i} \\
&\quad \times \left[1 - \left(\frac{1}{2} + \left(x_i^T\beta \times \sqrt{v}\right) \times \Gamma\left(\frac{v+1}{2}\right) \frac{{}_2F_1(\frac{1}{2}, \frac{v+1}{2}; \frac{3}{2}; \frac{-\left(x_i^T\beta \times \sqrt{v}\right)^2}{v})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}\right)\right]^{1-y_i},
\end{aligned}
\tag{3.16}
$$

donde y son los datos observados y $\theta = (\beta^T, v)^T$.

3.2.2. Función de distribución a priori

Para la distribución a priori de los coeficientes β del modelo de regresión robit con enlace t-Generalizado simétrico, se asumirá para cada, una distribución normal dada por:

$$
\beta_j \overset{ind}{\sim} N(0, \sigma_\beta^2), \tag{3.17}
$$

donde $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ es el vector p-dimensional de coeficientes de regresión. Por otro lado, la distribución a priori para el parámetro de forma (v), al igual que lo propuesto en el modelo robit, se asumirá como una exponencial desplazada, dada por Ross (2022a):

$$
\begin{aligned}
v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\
u &\sim U(1, 100),
\end{aligned}
\tag{3.18}
$$

donde U es la distribución uniforme con parámetros que varían entre 1 y 100. Dada a la propiedad de independencia de los parámetros del modelo, se tiene que la forma de la distribución a priori conjunta, está dada por:

$$
\pi(\theta) = \pi(\beta_1) \times \pi(\beta_2) \times \cdots \times \pi(\beta_p) \times \pi(v). \tag{3.19}
$$

3.2.3. Función de distribución a posteriori

A partir de las ecuaciones en (3.16) y (3.19), la función de distribución a posteriori para el modelo robusto con función de enlace t-Generalizado se define como:

$$\begin{aligned}
\pi(\theta \mid y) &\propto p(y \mid \theta)\pi(\theta) \\
\pi(\theta \mid y) &\propto \prod_{i=1}^{n} (\pi_i)^{y_i} (1-\pi_i)^{1-y_i} \times \pi(\theta) \\
\pi(\theta \mid y) &\propto \prod_{i=1}^{n} \left[F_v\left(x_i^T\beta \times \sqrt{v}\right)\right]^{y_i} \left[1 - F_v\left(x_i^T\beta \times \sqrt{v}\right)\right]^{1-y_i} \times \pi(\theta).
\end{aligned} \tag{3.20}$$

Nuevamente, para encontrar la solución analítica a la expresión definida en (3.20) resulta ser compleja, por lo tanto, se empleará el método MCMC con el algoritmo de Gibbs Sampling para obtener muestras que representen la distribución a posteriori de manera más efectiva.

3.3. Inferencia bayesiana

Previo a proporcionar las notaciones estocásticas de los modelos estadísticos bayesianos, es esencial presentar de manera breve los conceptos fundamentales asociados a este enfoque de inferencia estadística.

La inferencia bayesiana es un método de inferencia probabilística que ha ganado mucha popularidad en las últimas dos décadas debido a los avances en capacidad computacional asequible y por los métodos de Monte Carlo de Cadena de Markov (MCMC) que permiten aproximar integrales de alta dimensión. Su origen se remonta a Thomas Bayes (1764), quien realizó la derivada de la probabilidad inversa de éxito θ en una secuencia de ensayos de Bernoulli independientes. En este contexto, θ se consideraba desconocido y se suponía que seguía una distribución uniforme en el intervalo [0,1] (Liang, Liu y Carroll, 2011).

La estimación bayesiana en comparación a métodos frecuentistas, ofrece un mayor control sobre la incertidumbre al incorporar conocimientos previos en el proceso de estimación de parámetros. Mediante la inferencia bayesiana, se busca obtener la distribución de la muestra a posteriori de los parámetros basada en los datos empíricos y la distribución a priori definida previamente. Este enfoque, que combina información previa e incertidumbre, conduce a inferencias más sólidas, incluso en casos de muestras muy pequeñas y con presencia de valores atípicos. Al integrar tanto la información previa como los datos observados, la estimación bayesiana proporciona una mayor robustez en los resultados obtenidos en comparación a las EMV (Li et al., 2023).

3.3.1. Algoritmo Markov Chain Monte Carlo (MCMC)

Simular muestras independientes e idénticamente distribuidas para una distribución objetivo puede resultar complicado. En su lugar, es más sencillo utilizar muestras dependientes, siempre y cuando la media muestral converja a una velocidad adecuada. El método de las cadenas de Markov es una poderosa herramienta para simular secuencias de variables aleatorias dependientes, cuya distribución se aproxime a la distribución objetivo. Una característica fundamental de este método es que, dado un estado presente, los estados pasados y futuros son independientes entre sí (Liang et al. (2011) y Ross (2022b)).

3.3.2. Especificación del modelo bayesiano

El objetivo de la inferencia estadística es estimar un conjunto de parámetros θ (vector de parámetros desconocidos) a partir de los datos observados en problemas del mundo real. En el contexto de modelos bayesianos, se puede expresar el análisis bayesiano en los siguiente puntos:

- Especificación del modelo de muestreo para los datos observados $X_i = (x_{i1}, x_{i2}, ..., x_{in})^T$, condicionado a una cantidad desconocida de parámetros $\theta_i = (\theta_1, \theta_2, ..., \theta_p)^T$.

$$X \sim f(X|\theta) \quad (X \in \chi,\ \theta \in \Theta), \tag{3.21}$$

  donde $f(X|\theta)$ es la función de densidad de probabilidad.

- Especificación de la distribución marginal o distribución a priori para θ.

$$X \sim \pi(\theta) \quad (\theta \in \Theta). \tag{3.22}$$

Para obtener resultados inferenciales de interés, es necesario calcular las integrales con respecto a la distribución a posteriori usando el teorema de Bayes, a partir de las ecuaciones 3.21 y 3.22.

$$\pi(\theta \mid X) = \frac{\pi(\theta)L(\theta \mid X)}{\int \pi(\theta)L(\theta \mid X)d\theta} \quad (\theta \in \Theta), \tag{3.23}$$

donde la expresión $L(\theta \mid X)$ representa la función de verosimilitud en su forma general, y $\pi(\theta \mid X)$ es la distribución a posteriori.

La ecuación descrito en 3.23 puede ser expresado bajo la condición de proporcionalidad de la siguiente manera:

$$\pi(\theta \mid X) \propto \pi(\theta) L(\theta \mid X). \tag{3.24}$$

3.4. Análisis de convergencia

Una etapa crucial en la inferencia bayesiana implica evaluar si el modelo se adapta apropiadamente a los datos (es decir, si la distribución a posteriori de los parámetros llega a converger). No realizar este paso podría llevar a inferencias engañosas cuando el modelo presenta deficiencias y no resulta plausible para su aplicación en situaciones de la vida real. Si el modelo se ajusta correctamente, los datos simulados deberían ser coherentes con los datos observados (Gelman et al., 2014).

Sin embargo, la convergencia exacta de la mezcla de cadenas nunca se logra en realidad, lo que nos lleva a depender de la estimación estadística de los resultados de la simulación MCMC para evaluar la distancia entre las simulaciones actuales y una mezcla perfecta. En la práctica, se monitoriza la convergencia de cada parámetro y de otras magnitudes de interés de manera individual. Esto implica calcular la varianza de las simulaciones de cada cadena, descartando las primeras mitades de cada una (período de adaptación o burn-in), promediando las varianzas dentro de cada cadena y luego comparándolas con las varianzas de todas las cadenas mezcladas conjuntamente. En el momento de la convergencia, las cadenas habrán logrado mezclarse de tal manera que las distribuciones de las simulaciones entre cadenas y dentro de ellas serán idénticas. Para verificar esto, también se puede recurrir al análisis gráfico de las series temporales de las cadenas simuladas para identificar dónde ocurre la mezcla de manera eficiente (Brooks et al., 2011).

3.4.1. Estadístico de Geweke

La estadística propuesta por Geweke (1991) emplea un Z-score normalizado, basado en una prueba de igualdad de medias, que contrasta la igualdad de la media de la primera mitad inicial de la cadena que representa un 10 %, en comparación con la segunda mitad de la cadena (50 %). Si la prueba demuestra que estas diferencias no son relevantes, se llega a la conclusión de que la distribución objetivo ha logrado converger en algún punto dentro de la parte inicial de la cadena.

El estadístico es el siguiente:

$$Z = \frac{\bar{\theta}_{j1} - \bar{\theta}_{j2}}{\sqrt{se_{j1}^2 + se_{j2}^2}}. \tag{3.25}$$

Se realiza el cálculo del estadístico individualmente para cada parámetro. Cuando el valor absoluto del Z supera el umbral de 2, esto señala una convergencia deficiente.

3.4.2. Estadístico de Gelman y Rubin

El método introducido por Gelman y Rubin (1992) se fundamenta en llevar a cabo inferencias utilizando múltiples cadenas independientes ($m \geq 2$) de la distribución a posteriori, simuladas de manera iterativa, para cada uno de los parámetros de interés. Esta técnica evalúa la similitud entre las cadenas mediante un análisis numérico que se basa en la comparación de la variabilidad entre las cadenas y la variabilidad dentro de cada cadena.

La estadística se define como:

$$R = \sqrt{(df)\hat{V}/((df-2)W)}, \tag{3.26}$$

donde: $\hat{V}$ es la varianza estimada

$$\hat{V} = \hat{\sigma}^2 + B/(mn), \tag{3.27}$$

y definido a partir de:

- B/n es la varianza empírica entre las cadenas definido como $\Sigma_{i=1}^{m}(\bar{x}_{i.} - \bar{x}_{..})^2/(m-1)$
- σ^2 es la varianza objetivo, estimado mediante un promedio ponderado de W y B. W se define como el promedio de las varianzas, S_i^2, dentro de las secuencias con $n-1$ grados de libertad.

$$W = \sum_{i=1}^{m} s_i^2/m.$$

luego:

$$\hat{\sigma}^2 = B/n + (n-1)/nW,$$

y *df* los grados de libertad definido como:

$$df = \frac{2\hat{V}}{\widehat{Var}(\hat{V})},$$

donde $\widehat{Var}(\hat{V})$ las varianzas y covarianzas son obtenidos a partir de las m muestras de $\hat{x}_{i.}$ y s_i^2

$$\widehat{Var}(\hat{V}) = \left(\frac{n-1}{n}\right)^2 \frac{1}{m}\widehat{Var}(s_i^2) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1}B^2 + 2\frac{(m+1)(n-1)}{mn^2}\frac{n}{m}[\widehat{cov}(s_i^2, \bar{x}_{i.}^2) - 2\bar{x}_{..}\widehat{cov}(s_i^2, \bar{x}i.)]. \tag{3.28}$$

Por último, se estima R, que consiste en la relación entre la estimación actual de la varianza, V, y la varianza dentro de las secuencias, W, con un factor que considera la varianza adicional presente en la distribución t de Student. Si la posibilidad de reducción en la escala es alta, ello nos brinda motivos para suponer que la continuación de más simulaciones podría fortalecer la inferencia acerca de la distribución objetivo. Cuando el valor de R se acerca a 1 en relación con cada una de las estimaciones de los parámetros, entonces se señala que las cadenas han alcanzado la convergencia de la distribución a posteriori.

3.5. Métodos de comparación de modelos

En el marco de esta investigación sobre inferencia bayesiana, la comparación y selección del modelo más plausible es de suma importancia para el modelado de los datos observados. En este proceso de comparación entre los enfoques de función de enlace clásico y enlace simétrico t generalizado, se utilizará las métricas de Criterio de Información de Devianza (DIC) y el Criterio de Información de Watanabe-Akaike (WAIC).

3.5.1. Criterio de Información de Devianza (DIC)

Spiegelhalter, Best, Carlin y Van Der Linde (2002) proponen el criterio DIC como una herramienta para la comparación de modelos jerárquicos en inferencia bayesiana. Este criterio combina el grado de ajuste con una penalización por la complejidad del modelo, basándose en la medida de información llamada devianza (D). En esencia, el DIC puede verse como una extensión bayesiana del AIC. La formulación de esta medida es la siguiente:

$$\widehat{DIC} = D(\hat{\theta}) + 2\hat{p}_D, \tag{3.29}$$

donde: $\bar{D} = \frac{1}{S}\sum_{s=1}^{S} D(\theta^{(s)})$ es la media a posteriori de la devianza, $D(\hat{\theta}) = -2\sum_{i=1}^{n} log(y_i \mid \hat{\theta})$ es la devianza bayesiana, $\hat{\theta}$ es la media de los valores simulados de la distribución a posteriori y $\hat{p}_D \approx p$ es el número efectivo de parámetros definido como $\bar{D} - D(\hat{\theta})$. DIC puede calcularse a partir de las simulaciones obtenidas por MCMC.

Dado que el DIC se comporta similar al AIC, entonces, para encontrar al mejor modelo

que minimice la pérdida de información, la idea es preferir a los modelos con menor DIC en comparación a los modelos con DIC más grandes.

3.5.2. Criterio de Información de Watanabe - Akaike (WAIC)

El WAIC es una estrategia ampliamente utilizada en el análisis bayesiano y, al igual que el DIC, también evalúa el número efectivo de parámetros para mitigar el riesgo de sobreajuste. En la literatura, se distinguen dos tipos de WAIC: pWAIC1, similar al DIC, y pWAIC2, que se calcula a partir de la varianza. En este contexto, Gelman et al. (2014) recomienda utilizar pWAIC2, dado que sus resultados se asemejan a los obtenidos mediante la validación cruzada de dejar uno afuera (Leave-One-Out Cross-Validation, LOOCV).

El primer enfoque es el siguiente:

$$\mathcal{P}WAIC1 = 2\sum_{i=1}^{n}\left(\log(E_{post}p(y_i|\theta)) - E_{post}(\log(p(y_i|\theta))\right). \tag{3.30}$$

El cual puede ser calculado a partir de las simulaciones MCMC, considerando el promedio de las S muestras a posteriores de los parámetros θ.

$$\widehat{\mathcal{P}WAIC}1 = 2\sum_{i=1}^{n}\left(\log(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\theta^s)) - \frac{1}{S}\sum_{s=1}^{S}\log p(y_i|\theta^s)\right). \tag{3.31}$$

El segundo enfoque que utiliza la varianza es:

$$\mathcal{P}WAIC2 = \sum_{i=1}^{n}\text{var}_{post}(\log p(y_i|\theta)). \tag{3.32}$$

Se sabe que: $V_{s=1}^{S}a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$ es la varianza de la muestra y para poder estimar la ecuación 3.32, se debe de calcular la varianza a posteriori de la densidad de predicción logarítmica para cada punto de datos y_i, entonces se tiene:

$$\widehat{\mathcal{P}WAIC}2 = \sum_{i=1}^{n}V_{s=1}^{S}(\log p(y_i|\theta^S)), \tag{3.33}$$

Luego, a partir de las sugerencias realizadas por Gelman et al. (2014) y Watanabe y Opper (2010), se utiliza pWAIC2 dado a su aproximación a tiene que:

$$\widehat{WAIC} = -2(\widehat{lppd} - \widehat{\mathcal{P}WAIC}), \tag{3.34}$$

donde: $\widehat{lppd} = \sum_{i=1}^{n}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\theta^s)\right)$ es la densidad predictiva logarítmica.

Capítulo 4

Estudio de Simulación

En este capítulo, se llevó a cabo la simulación de datos, estimación y comparación de los coeficientes de regresión obtenidos mediante cinco modelos bayesianos propuestos. Estos modelos se evaluaron utilizando una base de datos simulada con tamaños de muestra de 50, 100 y 200 observaciones, que incluye una variable predictora cuantitativa y una variable respuesta binaria.

Los modelos bayesianos considerados fueron los siguientes: el modelo con enlace probit, designado como el modelo de referencia para la comparación; un modelo robit con función de enlace a partir de la distribución t-Student, con un parámetro de forma fijo en $v = 2$; otro modelo robit con el parámetro de forma tratado como una variable aleatoria, es decir, sin una especificación previa, permitiendo que sea estimado a partir de los datos simulados; un modelo robit con función de enlace a partir de la distribución t-Generalizado, donde los parámetros de forma y escala se fijaron en $v_1 = 2$ y $v_2 = 1$ respectivamente; y por último, un modelo robit t-Generalizado con el parámetro de forma tratado como variable aleatoria, y el parámetro de escala fijo en 1.

Se estimaron las probabilidades de ocurrencia de la variable respuesta utilizando los diferentes modelos bayesianos a partir de datos simulados con diferentes tamaños de muestra y niveles de contaminación con datos atípicos: 0 %, 2 %, 4 % y 6 %. El objetivo principal es determinar la eficacia de los modelos robustos en comparación con el modelo probit de referencia en términos de su capacidad para estimar con el menor sesgo posible la probabilidad real de los datos simulados. Los resultados de la estimación de las probabilidades corresponden a una réplica de análisis.

La simulación de los datos se realizó con R, y los coeficientes de regresión se estimaron mediante el algoritmo de Gibbs Sampling utilizando JAGS (Just Another Gibbs Sampler). Ambos programas estadísticos son de acceso gratuito.

4.1. Modelos bayesianos en estudio

Aquí se presentan las notaciones de los cinco modelos bayesianos que serán empleados: el primero, considerado como la base de comparación, es el modelo de regresión binaria con enlace probit. Por otro lado, para la estimación de los modelos robustos con funciones de enlace t-Student y t-Generalizado, se adoptó el enfoque sugerido por Kim et al. (2008).

4.1.1. Modelo 1: Regresión Probit

Se describen los parámetros del modelo para la estimación de los coeficientes de regresión, utilizando el modelo binario de referencia con enlace Probit, mediante la siguiente notación:

$$Y_i \sim Bernoulli(\pi_i)$$
$$\pi_i = \Phi(\beta_0 + \beta_1 x_i),$$

donde $i = 1, 2, \ldots, 100$ representa el tamaño de muestra simulada, y las a prioris para los parámetros β del modelo de regresión se considera como una $N(0, \sigma_\beta^2 = 100^2)$.

4.1.2. Modelo 2: Regresión Robit y parámetro de forma fijo

El primer modelo robusto propuesto, que emplea una función de enlace t-Student con un parámetro de forma fijo en 2, se detalla mediante la ecuación definida en (3.4) del capítulo 3 de modelos. Para la estimación de los coeficientes de regresión a partir de este modelo, se tiene en cuenta la siguiente notación:

$$Y_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_{(v=2)}(\beta_0 + \beta_1 x_i),$$

donde $i = 1, 2, \ldots, 100$ representa al tamaño de muestra simulada, y las a prioris para los parámetros β del modelo de regresión, se describe al modelo especificado en (3.6) con $\sigma_\beta^2 = 100^2$.

4.1.3. Modelo 3: Regresión Robit y parámetro de forma aleatorio

El segundo modelo robusto propuesto, que emplea una función de enlace t-Student con un parámetro de forma considerado como aleatorio, se detalla mediante la ecuación definida

en (3.4). Para la estimación de los coeficientes de regresión a partir de este modelo, se tiene en cuenta la siguiente notación:

$$Y_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v(\beta_0 + \beta_1 x_i),$$

donde $i = 1, 2, \ldots, 100$ representa al tamaño de muestra simulada, y las a prioris para los parámetros β del modelo de regresión, se considera según lo especificado en (3.6) con $\sigma_\beta^2 = 100^2$ y la a priori para el parámetro de forma v, se asume una distribución exponencial desplazada según lo definido en (3.7).

4.1.4. Modelo 4: Regresión Robit t-Generalizada y parámetro de forma fijo

El tercer modelo robusto propuesto, que emplea una función de enlace t-Generalizado con un parámetro de forma y escala fijo en $v_1 = 2$ y $v_2 = 1$ respectivamente, se detalla mediante la ecuación definida en (3.15). Para la estimación de los coeficientes de regresión a partir de este modelo, se tiene en cuenta la siguiente notación:

$$Y_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v\left[(\beta_0 + \beta_1 \times x_i) \times \sqrt{v}\right],$$

donde $i = 1, 2, \ldots, 100$ representa al tamaño de muestra simulada y las a prioris para los parámetros β del modelo de regresión, se considera según lo especificado en (3.17) con $\sigma_\beta^2 = 100^2$.

4.1.5. Modelo 5: Regresión Robit t-Generalizada y parámetro de forma aleatorio

El cuarto robusto propuesto, que emplea una función de enlace t-Generalizado con un parámetro de forma considerado como aleatorio y parámetro de escala fijo en 1, se detalla mediante la ecuación definida en (3.4). Para la estimación de los coeficientes de regresión a partir de este modelo, se tiene en cuenta la siguiente notación:

$$
\begin{aligned}
Y_i &\sim Bernoulli(\pi_i) \\
\pi_i &= F_v\left[(\beta_0 + \beta_1 \times x_i) \times \sqrt{v}\right]
\end{aligned}
$$

donde $i = 1, 2, \ldots, 100$ y las a prioris para los parámetros β del modelo de regresión se considera según lo especificado en (3.17) con $\sigma_\beta^2 = 100^2$ y la a priori para el parámetro de forma v se asume una distribución exponencial desplazada definido en (3.18).

4.2. Criterio de comparación para estimadores

Para llevar a cabo la comparación entre las estimaciones de los parámetros provenientes de cada modelo, y en cada uno de los diversos escenarios de contaminación por datos atípicos, en relación con los valores reales de dichos parámetros, se consideró la aplicación de la siguiente métrica:

1. **Sesgo**: Definido de la siguiente forma:

$$Sesgo(\hat{\beta}_j) = E(\hat{\beta}_j \mid D) - \beta_j, \tag{4.1}$$

   donde D, representa a los datos observados. De acuerdo con la teoría, se espera que los modelos de regresión robustos con enlaces t y t-Generalizado, deberían lograr una recuperación más precisa de los parámetros reales, lo que a su vez resultaría en una estimación mínima del sesgo para estos modelos en comparación del modelo probit.

4.3. Simulación de datos

Para la simulación de datos, se consideró a la variable respuesta Y_i con una distribución Bernoulli y dada por:

$$
\begin{aligned}
Y_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \Phi(\beta_0 + \beta_1 \times x_i) \\
x_i &\sim N(0, 1),
\end{aligned} \tag{4.2}
$$

donde $i = 1, \ldots, 50$ representa el primer tamaño de muestra a simular, luego se simuló otros dos conjuntos de datos con tamaños de muestra de 100 y 200 casos. Para estimar la probabi-

lidad real π_i del modelo Bernoulli, se asumió calcularlo utilizando la función de distribución acumulada Φ, que corresponde a la distribución normal estándar. Para llevar a cabo esta estimación, se asignaron valores específicos a los parámetros de regresión: $\beta_0 = 2$ y $\beta_1 = 3$. La covariable cuantitativa se generó aleatoriamente a partir de una distribución normal estándar. Es importante destacar que en este primer escenario (i) de datos simulados no se consideró la presencia de valores atípicos. Posteriormente, se crearon tres escenarios adicionales que incluían niveles de valores atípicos, con tasas de ocurrencia de $r \in \{2\,\%, 4\,\% \text{ y } 6\,\%\}$.

Para generar la data contaminada con los datos atípicos en los escenarios restantes se realizó lo siguiente:

- Escenario (ii): Para agregar una contaminación del $2\,\%$ de datos atípicos a una data simulada de $n \in \{50, 100, 200\}$ observaciones, se realizó reemplazando el valor de y_i bajo la siguiente condición: $y^* = y[x \leq min(x)] = 1 - y$ y $y^* = y[x \geq max(x)] = 1 - y$.

- Escenario (iii): Para agregar una contaminación del $4\,\%$ de datos atípicos a una data simulada de $n \in \{50, 100, 200\}$ observaciones, se realizó reemplazando el valor de y_i bajo la siguiente condición: $y^* = y[x \leq P_{0,02}(x)] = 1 - y$ y $y^* = y[x \geq P_{0,98}(x)] = 1 - y$.

- Escenario (iv): Para agregar una contaminación del $6\,\%$ de datos atípicos a una data simulada de $n \in \{50, 100, 200\}$ observaciones, se realizó reemplazando el valor de y_i bajo la siguiente condición: $y^* = y[x \leq P_{0,03}(x)] = 1 - y$ y $y^* = y[x \geq P_{0,97}(x)] = 1 - y$.

4.4. Resultados de simulación

La Figura 4.1 muestra cómo la presencia de datos atípicos (representados por círculos dorados) genera un impacto negativo y significativo en los resultados debido a una especificación deficiente de la función de enlace en el predictor lineal, lo que conduce a una estimación imprecisa de la probabilidad real del modelo. A medida que aumenta el nivel de contaminación con datos atípicos, el sesgo en la estimación de la probabilidad real a partir del modelo de regresión Probit (de referencia) tiende a incrementarse, resultando en menor precisión en comparación con los modelos robustos, ya que la probabilidad estimada tiende a desviarse más de la probabilidad real a medida que aumenta el nivel de contaminación con datos atípicos. Este efecto es observable en todos los escenarios de tamaño de muestra.

Es importante precisar que estas estimaciones de probabilidad fueron obtenidas a partir de una única réplica de análisis. Las configuraciones utilizadas para cada modelo se detallaron en la sección 4.1. Para las estimaciones de coeficientes de regresión y probabilidades, se utilizó

el software JAGS, que emplea el algoritmo de Gibbs. Todos los modelos se configuraron con un total de 10,000 iteraciones en cuatro cadenas de Markov. Se incluyó un período de burn-in de 1,000 iteraciones para asegurar la convergencia adecuada de las cadenas. Además, para mitigar el efecto de la autocorrelación en la muestra posterior, se aplicó un salto (thin=2).

Al analizar los resultados de los coeficientes de regresión estimados para cada modelo bayesiano, se observa que, para un tamaño de muestra de $n = 50$ y para los diferentes escenarios de contaminación con datos atípicos, las estimaciones obtenidas por los modelos robustos tienden a sobreestimar en mayor medida los verdaderos parámetros utilizados en la simulación, en comparación con el modelo de referencia (para algunos casos). Además, se aprecia que la estimación puntual del parámetro de forma para los modelos Robit y Robit tG, cuando no hay presencia de datos atípicos, es considerablemente alta, así como el intervalo de credibilidad. Sin embargo, en los otros escenarios, estos intervalos tienden a tener una menor amplitud (ver Tablas del 4.1 al 4.4).

Tabla 4.1: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 0 % y para un tamaño de muestra n=50.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	1.898	0.007	(0.920 ; 3.193)	0.102
	$\hat{\beta}_1$	3.667	0.013	(1.921 ; 6.047)	-0.667
Robit	$\hat{\beta}_0$	5.240	0.097	(1.547 ; 14.31)	-3.240
	$\hat{\beta}_1$	9.497	0.163	(3.235 ; 24.44)	-6.497
Robit v aleatorio	$\hat{\beta}_0$	4.589	0.843	(1.010 ; 29.16)	-1.759
	$\hat{\beta}_1$	8.489	1.756	(2.103 ; 51.20)	-5.489
	$\hat{v}$	37.764	2.039	(1.052 ; 213.24)	-
Robit tG	$\hat{\beta}_0$	3.759	0.079	(1.074 ; 10.56)	-1.759
	$\hat{\beta}_1$	6.784	0.134	(2.289 ; 17.68)	-3.784
Robit tG v aleatorio	$\hat{\beta}_0$	7.350	0.660	(0.231 ; 35.72)	-5.350
	$\hat{\beta}_1$	13.219	1.166	(0.454 ; 64.07)	-10.219
	$\hat{v}$	6.195	1.875	(1.014 ; 61.42)	-

Cuando el tamaño de muestra va en aumento ($n = 100$ y $n = 200$), las estimaciones de los coeficientes de regresión de los modelos robustos tienden a recuperar mejor los parámetros utilizados. Al analizar los escenarios con datos atípicos, se observa que estos modelos presentan un menor sesgo en comparación con el modelo de referencia, y a su vez, la estimación de la incertidumbre para el parámetro de forma en los modelos Robit y Robit tG, tienen una menor amplitud en los escenarios con atípicos respecto al escenario sin atípicos.

En conclusión, al analizar tanto los resultados gráficos como los resultados analíticos, se espera que las estimaciones obtenidas de los modelos robustos que estiman el parámetro de forma (Robit v aleatorio y Robit tG v aleatorio) tengan un mejor ajuste a los datos, logrando así una mejor recuperación de los parámetros reales del modelo en comparación con el modelo

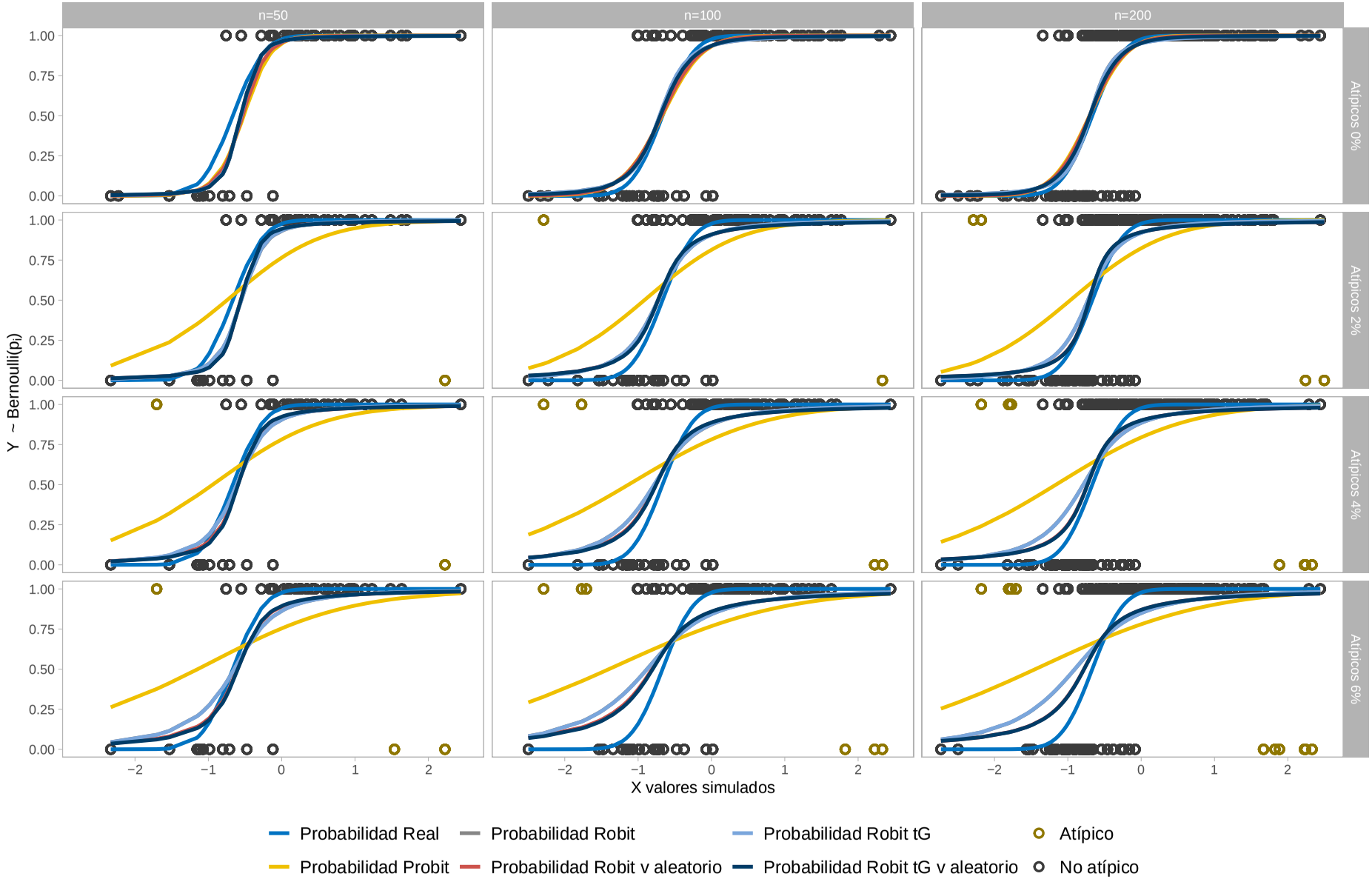

Figura 4.1: Estimación de la probabilidad real del desenlace según el nivel de contaminación con valores atípicos al 0 %, 2 %, 4 % y 6 %, en tamaños de muestra de 50, 100 y 200, para los diferentes modelos de regresión bayesianos propuestos.

Tabla 4.2: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 2 % y para un tamaño de muestra n=50.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.743	0.002	(0.309 ; 1.210)	1.258
	$\hat{\beta}_1$	0.999	0.002	(0.510 ; 1.544)	2.002
Robit	$\hat{\beta}_0$	3.181	0.038	(1.066 ; 7.803)	-1.181
	$\hat{\beta}_1$	5.837	0.068	(2.182 ; 13.513)	-2.837
Robit v aleatorio	$\hat{\beta}_0$	7.045	0.429	(1.305 ; 27.419)	-5.045
	$\hat{\beta}_1$	12.754	0.723	(2.618 ; 48.225)	-9.754
	$\hat{v}$	1.430	0.011	(1.008 ; 2.944)	-
Robit tG	$\hat{\beta}_0$	2.279	0.027	(0.769 ; 5.651)	-0.279
	$\hat{\beta}_1$	4.161	0.048	(1.549 ; 9.739)	-1.161
Robit tG v aleatorio	$\hat{\beta}_0$	7.171	0.541	(1.072 ; 30.474)	-5.171
	$\hat{\beta}_1$	13.024	1.034	(2.111 ; 53.906)	-10.024
	$\hat{v}$	1.311	0.011	(1.006 ; 2.396)	-

Tabla 4.3: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 4 % y para un tamaño de muestra n=50.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.795	0.002	(0.364 ; 1.243)	1.205
	$\hat{\beta}_1$	0.850	0.002	(0.397 ; 1.336)	2.150
Robit	$\hat{\beta}_0$	2.317	0.018	(0.887 ; 5.030)	-0.317
	$\hat{\beta}_1$	3.821	0.030	(1.498 ; 8.231)	-0.821
Robit v aleatorio	$\hat{\beta}_0$	4.507	0.125	(0.981 ; 15.384)	-2.507
	$\hat{\beta}_1$	7.644	0.218	(1.546 ; 25.459)	-4.644
	$\hat{v}$	1.553	0.071	(1.007 ; 3.762)	-
Robit tG	$\hat{\beta}_0$	1.663	0.012	(0.628 ; 3.580)	0.337
	$\hat{\beta}_1$	2.737	0.021	(1.048 ; 5.805)	0.263
Robit tG v aleatorio	$\hat{\beta}_0$	4.636	0.212	(0.891 ; 15.633)	-2.636
	$\hat{\beta}_1$	7.913	0.440	(1.532 ; 26.048)	-4.913
	$\hat{v}$	1.294	0.011	(1.005 ; 2.328)	-

Tabla 4.4: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 6 % y para un tamaño de muestra n=50.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.699	0.002	(0.297 ; 1.112)	1.301
	$\hat{\beta}_1$	0.615	0.002	(0.201 ; 1.049)	2.385
Robit	$\hat{\beta}_0$	1.703	0.010	(0.676 ; 3.452)	0.297
	$\hat{\beta}_1$	2.656	0.017	(0.915 ; 5.496)	0.344
Robit v aleatorio	$\hat{\beta}_0$	2.942	0.036	(0.928 ; 6.241)	-0.942
	$\hat{\beta}_1$	5.521	0.060	(1.690 ; 11.661)	-2.521
	$\hat{v}$	1.221	0.006	(1.005 ; 1.724)	-
Robit tG	$\hat{\beta}_0$	1.200	0.007	(0.526 ; 2.308)	0.800
	$\hat{\beta}_1$	2.085	0.011	(0.928 ; 3.924)	0.915
Robit tG v aleatorio	$\hat{\beta}_0$	3.030	0.058	(0.970 ; 6.345)	-1.030
	$\hat{\beta}_1$	5.708	0.096	(1.747 ; 12.180)	-2.708
	$\hat{v}$	1.218	0.007	(1.006 ; 1.728)	-

de referencia (Probit), el modelo Robit con parámetro de forma fijo en $v = 2$ y el modelo Robit tG con parámetros de forma y escala fijos en $v_1 = 2$ y $v_2 = 1$, respectivamente.

Tabla 4.5: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 0 % y para un tamaño de muestra n=100.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	1.573	0.004	(1.000 ; 2.261)	0.427
	$\hat{\beta}_1$	2.342	0.006	(1.528 ; 3.308)	0.658
Robit	$\hat{\beta}_0$	2.825	0.018	(1.494 ; 4.860)	-0.825
	$\hat{\beta}_1$	4.003	0.023	(2.235 ; 6.658)	-1.003
Robit v aleatorio	$\hat{\beta}_0$	1.800	0.020	(1.047 ; 3.427)	0.200
	$\hat{\beta}_1$	2.638	0.026	(1.604 ; 4.826)	0.362
	$\hat{v}$	48.621	1.045	(1.620 ; 221.071)	-
Robit tG	$\hat{\beta}_0$	2.014	0.014	(1.069 ; 3.559)	-0.014
	$\hat{\beta}_1$	2.852	0.018	(1.604 ; 4.851)	0.148
Robit tG v aleatorio	$\hat{\beta}_0$	1.392	0.078	(0.198 ; 4.662)	0.608
	$\hat{\beta}_1$	1.983	0.109	(0.294 ; 6.421)	1.017
	$\hat{v}$	9.748	1.199	(1.092 ; 60.012)	-

Tabla 4.6: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 2 % y para un tamaño de muestra n=100.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.907	0.001	(0.577 ; 1.255)	1.093
	$\hat{\beta}_1$	0.991	0.002	(0.635 ; 1.373)	2.009
Robit	$\hat{\beta}_0$	2.167	0.012	(1.160 ; 3.674)	-0.167
	$\hat{\beta}_1$	2.975	0.015	(1.624 ; 4.912)	0.025
Robit v aleatorio	$\hat{\beta}_0$	2.800	0.035	(1.151 ; 5.820)	-0.800
	$\hat{\beta}_1$	3.849	0.049	(1.565 ; 7.876)	-0.849
	$\hat{v}$	1.736	0.030	(1.015 ; 4.409)	-
Robit tG	$\hat{\beta}_0$	1.514	0.008	(0.827 ; 2.539)	0.486
	$\hat{\beta}_1$	2.086	0.010	(1.156 ; 3.449)	0.914
Robit tG v aleatorio	$\hat{\beta}_0$	2.641	0.044	(0.771 ; 5.989)	-0.641
	$\hat{\beta}_1$	3.637	0.061	(1.047 ; 8.082)	-0.637
	$\hat{v}$	1.535	0.101	(1.009 ; 3.256)	-

Tabla 4.7: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 4 % y para un tamaño de muestra n=100.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.790	0.001	(0.494 ; 1.096)	1.210
	$\hat{\beta}_1$	0.696	0.001	(0.390 ; 1.010)	2.304
Robit	$\hat{\beta}_0$	1.695	0.007	(0.940 ; 2.762)	0.305
	$\hat{\beta}_1$	2.178	0.009	(1.144 ; 3.609)	0.822
Robit v aleatorio	$\hat{\beta}_0$	2.390	0.023	(1.026 ; 4.747)	-0.390
	$\hat{\beta}_1$	3.172	0.032	(1.247 ; 6.348)	-0.172
	$\hat{v}$	1.448	0.022	(1.009 ; 3.173)	-
Robit tG	$\hat{\beta}_0$	1.210	0.005	(0.675 ; 2.001)	0.790
	$\hat{\beta}_1$	1.552	0.007	(0.815 ; 2.579)	1.448
Robit tG v aleatorio	$\hat{\beta}_0$	2.277	0.026	(0.835 ; 4.743)	-0.277
	$\hat{\beta}_1$	3.032	0.035	(1.052 ; 6.275)	-0.032
	$\hat{v}$	1.305	0.010	(1.007 ; 2.330)	-

Tabla 4.8: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 6 % y para un tamaño de muestra n=100.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.739	0.001	(0.457 ; 1.037)	1.261
	$\hat{\beta}_1$	0.530	0.001	(0.244 ; 0.820)	2.471
Robit	$\hat{\beta}_0$	1.390	0.005	(0.788 ; 2.237)	0.610
	$\hat{\beta}_1$	1.581	0.006	(0.752 ; 2.725)	1.419
Robit v aleatorio	$\hat{\beta}_0$	1.953	0.018	(0.835 ; 3.775)	0.047
	$\hat{\beta}_1$	2.427	0.025	(0.781 ; 4.854)	0.573
	$\hat{v}$	1.522	0.071	(1.007 ; 4.153)	-
Robit tG	$\hat{\beta}_0$	0.986	0.003	(0.557 ; 1.586)	1.014
	$\hat{\beta}_1$	1.122	0.005	(0.534 ; 1.927)	1.878
Robit tG v aleatorio	$\hat{\beta}_0$	1.898	0.019	(0.718 ; 3.837)	0.102
	$\hat{\beta}_1$	2.389	0.025	(0.767 ; 4.977)	0.611
	$\hat{v}$	1.271	0.008	(1.006 ; 2.230)	-

Tabla 4.9: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 0 % y para un tamaño de muestra n=200.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	1.755	0.004	(1.274 ; 2.311)	0.245
	$\hat{\beta}_1$	2.492	0.005	(1.828 ; 3.246)	0.508
Robit	$\hat{\beta}_0$	3.130	0.017	(2.005 ; 4.774)	-1.130
	$\hat{\beta}_1$	4.489	0.024	(2.910 ; 6.758)	-1.489
Robit v aleatorio	$\hat{\beta}_0$	1.938	0.018	(1.331 ; 3.190)	0.062
	$\hat{\beta}_1$	2.759	0.025	(1.901 ; 4.536)	0.241
	$\hat{v}$	53.544	1.299	(2.039 ; 235.959)	-
Robit tG	$\hat{\beta}_0$	2.215	0.012	(1.403 ; 3.369)	-0.215
	$\hat{\beta}_1$	3.175	0.017	(2.029 ; 4.791)	-0.175
Robit tG v aleatorio	$\hat{\beta}_0$	1.797	0.264	(0.258 ; 6.156)	0.203
	$\hat{\beta}_1$	2.573	0.372	(0.365 ; 8.776)	0.427
	$\hat{v}$	8.702	1.406	(1.062 ; 45.088)	-

Tabla 4.10: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 2 % y para un tamaño de muestra n=200.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.937	0.001	(0.704 ; 1.180)	1.063
	$\hat{\beta}_1$	0.971	0.001	(0.710 ; 1.239)	2.029
Robit	$\hat{\beta}_0$	2.293	0.009	(1.495 ; 3.337)	-0.293
	$\hat{\beta}_1$	3.157	0.012	(2.037 ; 4.616)	-0.157
Robit v aleatorio	$\hat{\beta}_0$	3.532	0.038	(1.834 ; 6.209)	-0.532
	$\hat{\beta}_1$	4.981	0.054	(2.536 ; 8.771)	-1.981
	$\hat{v}$	1.280	0.007	(1.006 ; 2.131)	-
Robit tG	$\hat{\beta}_0$	1.605	0.006	(1.049 ; 2.360)	0.395
	$\hat{\beta}_1$	2.213	0.009	(1.425 ; 3.269)	0.787
Robit tG v aleatorio	$\hat{\beta}_0$	3.513	0.051	(1.470 ; 6.467)	-1.513
	$\hat{\beta}_1$	4.961	0.075	(2.028 ; 9.135)	-1.961
	$\hat{v}$	1.216	0.011	(1.004 ; 1.899)	-

Tabla 4.11: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 4 % y para un tamaño de muestra n=200.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.828	0.001	(0.610 ; 1.048)	1.172
	$\hat{\beta}_1$	0.712	0.001	(0.487 ; 0.944)	2.289
Robit	$\hat{\beta}_0$	1.773	0.005	(1.190 ; 2.521)	0.227
	$\hat{\beta}_1$	2.256	0.007	(1.434 ; 3.300)	0.744
Robit v aleatorio	$\hat{\beta}_0$	2.833	0.023	(1.574 ; 4.670)	-0.833
	$\hat{\beta}_1$	3.870	0.032	(2.016 ; 6.481)	-0.870
	$\hat{v}$	1.183	0.004	(1.004 ; 1.743)	-
Robit tG	$\hat{\beta}_0$	1.252	0.004	(0.883 ; 1.818)	0.748
	$\hat{\beta}_1$	1.590	0.006	(1.000 ; 2.378)	1.410
Robit tG v aleatorio	$\hat{\beta}_0$	2.772	0.026	(1.371 ; 4.734)	-0.772
	$\hat{\beta}_1$	3.794	0.037	(1.771 ; 6.561)	-0.794
	$\hat{v}$	1.152	0.004	(1.003 ; 1.636)	-

Tabla 4.12: Modelos de regresión bayesiano con parámetros de simulación: β_0=2 y β_1=3, datos atípicos al 6 % y para un tamaño de muestra n=200.

Modelos	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	**Sesgo**
Probit	$\hat{\beta}_0$	0.774	0.001	(0.570 ; 0.981)	1.226
	$\hat{\beta}_1$	0.537	0.001	(0.329 ; 0.751)	2.463
Robit	$\hat{\beta}_0$	1.411	0.003	(0.960 ; 1.980)	0.589
	$\hat{\beta}_1$	1.557	0.005	(0.919 ; 2.351)	1.443
Robit v aleatorio	$\hat{\beta}_0$	2.220	0.015	(1.241 ; 3.632)	-0.220
	$\hat{\beta}_1$	2.855	0.022	(1.380 ; 4.915)	0.145
	$\hat{v}$	1.174	0.005	(1.003 ; 1.766)	-
Robit tG	$\hat{\beta}_0$	0.995	0.002	(0.677 ; 1.397)	1.005
	$\hat{\beta}_1$	1.099	0.003	(0.654 ; 1.672)	1.901
Robit tG v aleatorio	$\hat{\beta}_0$	2.127	0.016	(1.097 ; 3.571)	-0.127
	$\hat{\beta}_1$	2.745	0.023	(1.248 ; 4.819)	0.255
	$\hat{v}$	1.144	0.004	(1.003 ; 1.615)	-

Capítulo 5

Aplicación

En este capítulo, se presentan los resultados del análisis de datos de dos aplicaciones relacionadas al campo clínico. Se mostrarán dos escenarios para cada aplicación. En el primer escenario, se aplicaron los cinco modelos de regresión bayesianos propuestos: Probit (modelo de referencia), Robit con función de enlace t-Student y parámetro de forma fijo en 2, Robit con parámetro de forma v aleatorio, Robit con función t-Generalizada (Robit tG) con parámetro de forma fijo en 2 y escala en 1, y Robit tG con parámetro de forma v aleatorio y escala fija en 1. Estos modelos se utilizaron con los datos originales que contengan alguna perturbación. En el segundo escenario, se realizaron las estimaciones considerando la ausencia de dichas perturbaciones en los datos.

5.1. Aplicación 1

5.1.1. Descripción de los datos

Para esta primera aplicación se considera un conjunto de datos conformado por 258 pacientes pediátricos menores de 5 años que sufrieron quemaduras de segundo y tercer grado, quienes recibieron atención en una institución de salud de tercer nivel en Perú. Los pacientes fueron seleccionados mediante criterios de conveniencia, asegurándose de que contaran con mediciones completas en las principales variables explicativas. El objetivo de la investigación fue estimar el efecto de la variable explicativa sobre la variable respuesta Complicación experimentada por el paciente.

A continuación se describen las variables de interés para la aplicación:

1. Variable respuesta: Se trata de la condición del paciente, que representa a la complicación que podría surgir a raíz de la quemadura grave. Esta variable es de naturaleza

categórica binaria, donde el valor 0 indica que no hay presencia de complicación, mientras que, el valor 1 indica que sí la hay.

2. Variable explicativa: La principal variable explicativa considerada para la aplicación de los modelos bayesianos es la albúmina (continua y cuantitativa), que es una proteína hepática vital para la evaluación y el tratamiento de las quemaduras. Otras variables analizadas de forma descriptiva son el género del paciente (variable binaria), la edad (discreta y cuantitativa).

5.1.2. Análisis exploratorio de los datos

Se analizaron las características de 258 niños que sufrieron quemaduras de segundo y tercer grado, encontrándose que el 24.4 % experimentó algún tipo de complicación. En la Figura 5.1, se representa el patrón de relaciones entre la variable explicativa y la variable respuesta. Por otro lado, en la Tabla 5.1, se presenta la distribución bivariada entre las variables explicativas de interés y su asociación con la variable respuesta. Se observa una homogeneidad en la distribución de la edad de los pacientes respecto a los grupos de complicaciones. En contraste, las demás variables muestran diferencias estadísticamente significativas ($p < 0{,}05$) en relación a la variable complicaciones.

Tabla 5.1: Análisis descriptivo bivariado de las características de los pacientes menores de 5 años con quemaduras que sufrieron algún tipo de complicación.

Características	**N = 258**	**Complicaciones**		**Valor p**
		No = 195	**Si = 63**	
Edad del paciente				0.200
Menor igual a 1 año	55.0 (21.3 %)	45.0 (23.1 %)	10.0 (15.9 %)	
2 a 5 años	203.0 (78.7 %)	150.0 (76.9 %)	53.0 (84.1 %)	
Sexo del paciente				0.014
Femenino	109.0 (42.2 %)	74.0 (37.9 %)	35.0 (55.6 %)	
Masculino	149.0 (57.8 %)	121.0 (62.1 %)	28.0 (44.4 %)	
Albúmina				0.001
Mediana (RIC)	2.7 (2.3 ; 3.3)	2.8 (2.4 ; 3.4)	2.3 (2.0 ; 2.8)	

5.1.3. Estimación de efectos con datos atípicos

En esta primera etapa del análisis, se pretende evaluar el efecto individual de cada variable explicativa en relación con la variable de respuesta que indica complicaciones en niños que han sufrido quemaduras. Para lograr esto, se estimaron los efectos de cada variable explicativa utilizando los cinco modelos bayesianos previamente definidos.

Figura 5.1: Análisis exploratorio de datos sobre las características de pacientes pediátricos menores de 5 años con quemaduras.

Cada modelo bayesiano bivariado se configuró con las siguientes especificaciones para generar muestras a posteriori mediante el método de Monte Carlo Markov Chain (MCMC). Se realizaron 40,000 iteraciones distribuidas en 4 cadenas de Markov. Con el fin de reducir la autocorrelación, se aplicaron saltos de 30 (thin=30). Además, se descartaron las primeras 1000 observaciones de cada cadena muestreada. Para verificar la convergencia adecuada de las cadenas para cada parámetro estimado, se utilizó la inspección visual de las series temporales y el estadístico de Geweke.

Los resultados de cada modelo se muestran a continuación:

Modelo 1: Regresión Probit con datos atípicos

Para este primer modelo, se estimó el efecto de la variable predictora albúmina sobre la ocurrencia de complicaciones en niños, para lo cual se utilizó un modelo probit bayesiano definido de la siguiente forma:

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = \Phi(\beta_0 + \beta_1 \times \text{Albúmina}_i),$$

donde $i = 1, 2, \ldots, 258$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.2 se muestra que el efecto estimado para la variable albúmina fue de $\hat{\beta}_1 = -0{,}539$ con un intervalo de credibilidad del 95 % comprendido entre -0.802 y -0.285, siendo este efecto estadísticamente significativo.

Tabla 5.2: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión probit bayesiano en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Probit	$\hat{\beta}_0$	0.766	0.007	0.058	1.477
	$\hat{\beta}_1$	-0.539	0.002	-0.802	-0.285
DIC		271.2			
WAIC		273.7			

Modelo 2: Regresión Robit con datos atípicos

En este modelo, se estimó el efecto de la variable predictora albúmina en la ocurrencia de complicaciones en niños. Para este modelo se utilizó como función de enlace a la distribución acumulada de la t-Student, con un parámetro de forma constante fijado en $v = 2$. El modelo fue definido de la siguiente forma:

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v(\beta_0 + \beta_1 \times \text{Albúmina}_i),$$

donde $i = 1, 2, \ldots, 258$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.3 se muestra que el efecto estimado para la variable lactato fue de $\hat{\beta}_1 = -1{,}059$ con un intervalo de credibilidad del 95 % comprendido entre -1.647 y -0.553, siendo este efecto estadísticamente significativo.

Tabla 5.3: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma $v = 2$ en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit	$\hat{\beta}_0$	1.861	0.016	0.610	3.226
	$\hat{\beta}_1$	-1.059	0.007	-1.647	-0.553
DIC		267.7			
WAIC		270.2			

Modelo 3: Regresión Robit con parámetro de forma aleatorio con datos atípicos

Para este análisis, se estimó el efecto de la variable predictora albúmina en la ocurrencia de complicaciones en niños. Siguiendo la estructura del modelo de regresión Robit, el cual considera al parámetro de forma como una variable aleatoria, el modelo tiene la siguiente forma:

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v(\beta_0 + \beta_1 \times \text{Albúmina}_i)$$
$$v \mid u \sim Exp\left(\frac{1}{u}, 1\right)$$
$$u \sim U(1, 100),$$

donde $i = 1, 2, \ldots, 258$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.4 se muestra que el efecto estimado para la variable albúmina fue de $\hat{\beta}_1 = -1{,}222$ con un intervalo de credibilidad del 95 % comprendido entre -2.449 y -0.409, siendo este efecto estadísticamente significativo.

Tabla 5.4: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma aleatorio en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit v aleatorio	$\hat{\beta}_0$	2.180	0.056	0.351	4.790
	$\hat{\beta}_1$	-1.222	0.027	-2.449	-0.409
	$\hat{v}$	8.607	0.865	1.012	63.126
DIC		267.3			
WAIC		270.3			

Modelo 4: Regresión Robit tG con datos atípicos

En este modelo, se estimó el efecto de la variable predictora albúmina en la ocurrencia de complicaciones en niños. La función de enlace utilizada fue de la familia Robit tG, en donde se fijó el parámetro de forma en $v_1 = v = 2$ y al parámetro de escala en $v_2 = 1$. La especificación de este modelo tiene la siguiente forma:

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v[(\beta_0 + \beta_1 \times \text{Albúmina}_i) \times \sqrt{v})],$$

donde $i = 1, 2, \ldots, 258$ y las distribuciones a prioris no informativas consideradas para los

coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.5 se muestra que el efecto estimado para la variable albúmina fue de $\hat{\beta}_1 = -0{,}741$ con un intervalo de credibilidad del 95 % comprendido entre -1.149 y -0.394, siendo este efecto estadísticamente significativo.

Tabla 5.5: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión robit tG bayesiano con parámetro de forma $v_1 = 2$ y escala $v_2 = 1$ en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG	$\hat{\beta}_0$	1.300	0.011	0.427	2.258
	$\hat{\beta}_1$	-0.741	0.005	-1.149	-0.394
DIC		267.7			
WAIC		270.1			

Modelo 5: Regresión Robit tG con parámetro de forma aleatorio con datos atípicos

Para esta aplicación, se estimó el efecto de la variable predictora albúmina en la ocurrencia de complicaciones en niños, utilizando la familia de la distribución t-Generalizada como función de enlace. En este modelo, el parámetro de forma se considera una variable aleatoria que se estimará en función de los datos, mientras que el parámetro de escala se mantiene fijo en $v_2 = 1$. Este modelo tiene la siguiente forma:

$$\begin{aligned}
\text{Complicación}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= F_v[(\beta_0 + \beta_1 \times \text{Albúmina}_i) \times \sqrt{v})] \\
v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\
u &\sim U(1, 100),
\end{aligned}$$

donde $i = 1, 2, \ldots, 258$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.6 se muestra que el efecto estimado para la variable albúmina fue de $\hat{\beta}_1 = -1{,}360$ con un intervalo de credibilidad del 95 % comprendido entre -2.563 y -0.286, siendo este efecto estadísticamente significativo.

Tabla 5.6: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión robit tG bayesiano con parámetro de forma aleatorio y escala $v_2 = 1$ en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG v aleatorio	$\hat{\beta}_0$	2.514	0.051	0.385	5.032
	$\hat{\beta}_1$	-1.360	0.025	-2.563	-0.286
	$\hat{v}$	1.578	0.083	1.007	5.048
DIC		265.7			
WAIC		268.2			

5.1.4. Estimación de efectos sin datos atípicos

Para la segunda parte de la aplicación 1, se realizó un nuevo modelamiento para estimar el efecto de la variable explicativa albúmina sobre la variable respuesta, complicaciones a causa de quemaduras en pacientes pediátricos. Estas nuevas estimaciones, obtenidas mediante cada uno de los cinco modelos bayesianos, consideran la exclusión de datos atípicos. La identificación de casos atípicos se realizó mediante un análisis gráfico entre la probabilidad estimada por cada modelo y el evento de interés, y mediante la aplicación de un modelo de regresión binaria con enlace probit frecuentista, analizando los valores de leverage.

Los resultados de las nuevas estimaciones sin datos atípicos se muestran a continuación:

Modelo 6: Regresión Probit sin datos atípicos

El modelo probit bayesiano tiene la siguiente forma:

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = \Phi(\beta_0 + \beta_1 \times \text{Albúmina}_i),$$

donde $i = 1, 2, \ldots, 253$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.7 se muestra que el efecto estimado para la variable albúmina sin la presencia de datos atípicos fue de $\hat{\beta}_1 = -0{,}894$ con un intervalo de credibilidad del 95 % comprendido entre -1.223 y -0.585, siendo este efecto estadísticamente significativo.

Modelo 7: Regresión Robit sin datos atípicos

El modelo robit bayesiano con parámetro de forma fijo fue definido de la siguiente forma:

Tabla 5.7: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión probit bayesiano sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Probit	$\hat{\beta}_0$	1.596	0.009	0.793	2.453
	$\hat{\beta}_1$	-0.894	0.003	-1.223	-0.585
DIC		240.1			
WAIC		242.1			

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v(\beta_0 + \beta_1 \times \text{Albúmina}_i),$$

donde $i = 1, 2, \ldots, 253$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.8 se muestra que el efecto estimado para la variable albúmina sin la presencia de datos atípicos fue de $\hat{\beta}_1 = -1{,}526$ con un intervalo de credibilidad del 95 % comprendido entre -2.253 y -0.923, siendo este efecto estadísticamente significativo.

Tabla 5.8: Estimación del efecto de la albúmina sobre las complicaciones por quemaduras en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma $v = 2$ sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit	$\hat{\beta}_0$	2.881	0.020	1.461	4.515
	$\hat{\beta}_1$	-1.526	0.009	-2.253	-0.923
DIC		239.0			
WAIC		241.1			

Modelo 8: Regresión Robit con parámetro de forma aleatorio sin datos atípicos

El modelo robit con parámetro de forma aleatorio tiene la siguiente forma:

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v(\beta_0 + \beta_1 \times \text{Albúmina}_i)$$
$$v \mid u \sim Exp\left(\frac{1}{u}, 1\right)$$
$$u \sim U(1, 100),$$

donde $i = 1, 2, \ldots, 253$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.9 se muestra que el efecto estimado para la variable albúmina sin datos atípicos fue de $\hat{\beta}_1 = -1{,}257$ con un intervalo de credibilidad del 95 % comprendido entre -2.691 y -0.636, siendo este efecto estadísticamente significativo.

Tabla 5.9: Estimación del efecto de la albúmina sobre la compliación por quemaduras en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma aleatorio sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit v aleatorio	$\hat{\beta}_0$	2.329	0.061	0.917	5.324
	$\hat{\beta}_1$	-1.257	0.029	-2.691	-0.636
	$\hat{v}$	31.200	2.456	1.041	168.244
DIC		239.6			
WAIC		241.9			

Modelo 9: Regresión Robit tG sin datos atípicos

El modelo robit con función de enlace t Generalizado tiene la siguiente forma:

$$\text{Complicación}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v[(\beta_0 + \beta_1 \times \text{Albúmina}_i) \times \sqrt{v})],$$

donde $i = 1, 2, \ldots, 253$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.10 se muestra que el efecto estimado para la variable albúmina sin datos atípicos fue de $\hat{\beta}_1 = -1{,}083$ con un intervalo de credibilidad del 95 % comprendido entre

-1.591 y -0.659, siendo este efecto estadísticamente significativo.

Tabla 5.10: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión robit tG bayesiano con parámetro de forma $v_1 = 2$ y escala $v_2 = 1$ sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG	$\hat{\beta}_0$	2.047	0.015	1.051	3.187
	$\hat{\beta}_1$	-1.083	0.006	-1.591	-0.659
DIC		239.0			
WAIC		241.2			

Modelo 10: Regresión Robit tG con parámetro de forma aleatorio sin datos atípicos

El modelo robit con función de enlace t Generalizado y con parámetro de forma aleatorio tiene la siguiente forma:

$$\begin{aligned}
\text{Complicación}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= F_v[(\beta_0 + \beta_1 \times \text{Albúmina}_i) \times \sqrt{v})] \\
v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\
u &\sim U(1, 100),
\end{aligned}$$

donde $i = 1, 2, \ldots, 253$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.11 se muestra que el efecto estimado para la variable albúmina sin datos atípicos fue de $\hat{\beta}_1 = -1{,}349$ con un intervalo de credibilidad del 95 % comprendido entre -2.795 y -0.148, siendo este efecto estadísticamente significativo.

5.1.5. Comparación de modelos

En la Figura 5.2, se presenta una comparativa de las probabilidades estimadas a partir de los cinco modelos bayesianos, en escenarios con y sin datos atípicos. Las curvas de probabilidad estimadas en presencia de datos atípicos muestran comportamientos distintos. La estimación del modelo probit tiene un comportamiento casi lineal, en contraste con los modelos robustos.

Tabla 5.11: Estimación del efecto de la albúmina sobre la complicación por quemaduras en niños, mediante un modelo de regresión robit tG bayesiano con parámetro de forma aleatorio y escala $v_2 = 1$ sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG v aleatorio	$\hat{\beta}_0$	2.561	0.120	0.239	5.484
	$\hat{\beta}_1$	-1.349	0.061	-2.795	-0.148
	$\hat{v}$	4.543	1.030	1.013	31.744
DIC		238.8			
WAIC		240.9			

Las estimaciones de probabilidad a partir de los modelos Robit, Robit con parámetro de forma v aleatorio y Robit tG son muy similares, ya que sus valores estimados se superponen. Sin embargo, las estimaciones del modelo Robit tG con parámetro de forma v aleatorio presentan una mayor curvatura, acercándose más a la presencia del evento (presencia de complicaciones = 1).

En el segundo escenario, cuando se excluyen los valores considerados como atípicos, las estimaciones de probabilidad tienden a parecerse en gran medida, aunque persiste un pequeño sesgo entre ellas. Se espera que, idealmente, las curvas de probabilidad estimadas de todos los modelos bayesianos se superpongan, reflejando estimaciones muy similares en un escenario sin perturbación.

Los resultados descritos en la Tabla 5.12 muestran que, en el escenario con los datos originales, los modelos robustos tienen un mejor rendimiento en comparación con el modelo de referencia. En este escenario, el mejor modelo es el Robit con enlace t Generalizado, con parámetro de forma v aleatorio y parámetro de escala fijo en 1, con un valor de $WAIC =$ 268,200.

Por otro lado, en el escenario sin la presencia de datos atípicos, los valores de WAIC para los modelos robustos son menores en comparación con el modelo de referencia, aunque la diferencia no se considera significativo, siendo aproximadamente de 2 puntos en el WAIC. Según el sentido común para la aplicación de estos modelos, se esperaría que el valor de WAIC para el modelo de referencia fuera menor. Sin embargo, se considera que estos resultados son aceptables y que las diferencias son suficientemente pequeñas para ser consideradas similares.

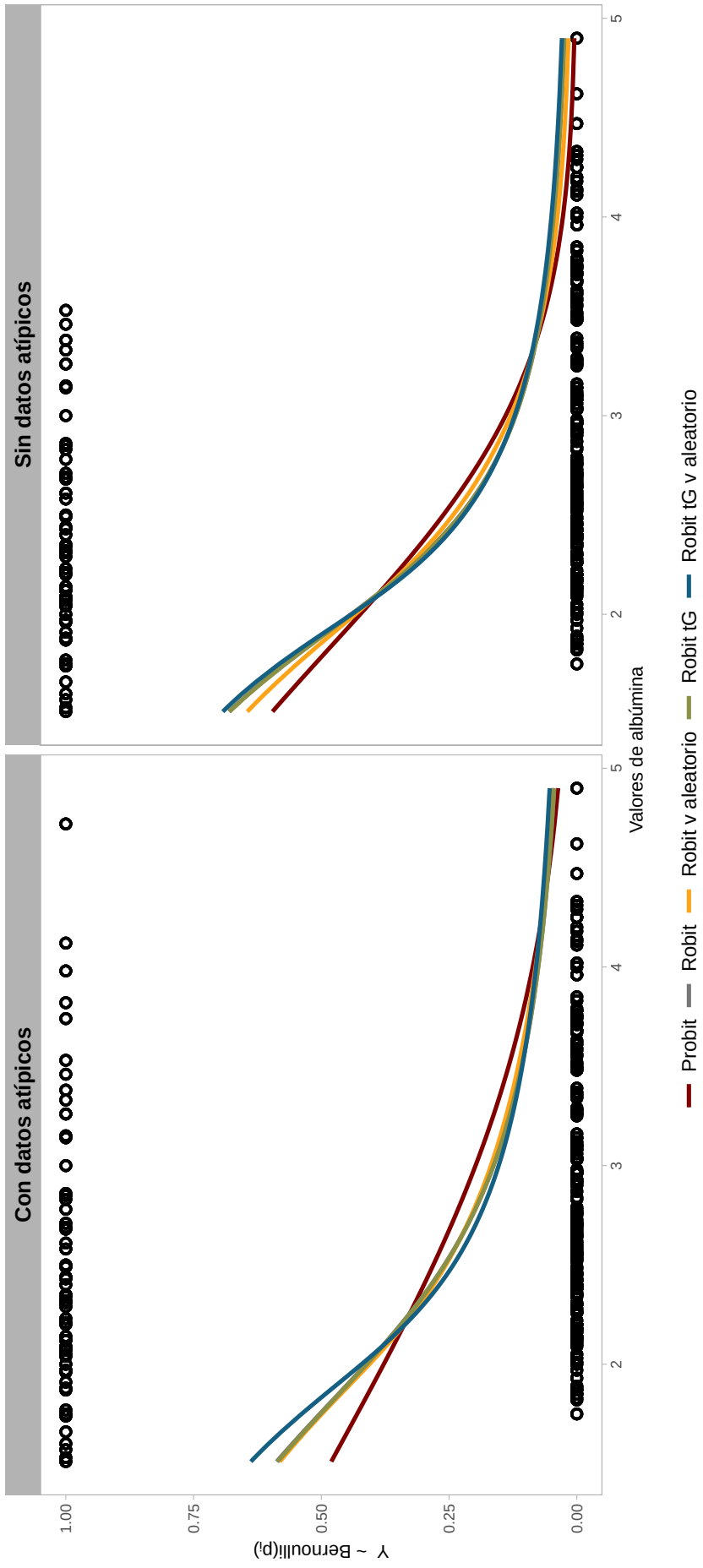

Figura 5.2: Comparación de las estimaciones de probabilidades con y sin datos atípicos para la presencia de complicaciones por quemaduras en pacientes pediátricos.

Tabla 5.12: Comparación de modelos bayesianos en presencia y ausencia de datos atípicos para la aplicación 1.

Datos originales					
Modelos	$\hat{\beta}_1$	**SE**	**IC95 %**	**DIC**	**WAIC**
Probit	-0.539	0.002	(-0.802 ; -0.285)	271.200	273.700
Robit	-1.059	0.007	(-1.647 ; -0.553)	267.700	270.200
Robit v aleatorio	-1.222	0.027	(-2.449 ; -0.409)	267.300	270.300
Robit tG	-0.741	0.005	(-1.149 ; -0.394)	267.700	270.100
Robit tG v aleatorio	-1.360	0.025	(-2.563 ; -0.286)	265.700	268.200
Datos sin presencia de atípicos					
Modelos	$\hat{\beta}_1$	**SE**	**IC95 %**	**DIC**	**WAIC**
Probit	-0.894	0.003	(-1.223 ; -0.585)	240.100	242.111
Robit	-1.526	0.009	(-2.253 ; -0.923)	239.000	241.114
Robit v aleatorio	-1.257	0.029	(-2.691 ; -0.636)	239.600	241.906
Robit tG	-1.083	0.007	(-1.591 ; -0.659)	239.000	241.152
Robit tG v aleatorio	-1.349	0.061	(-2.795 ; -0.148)	238.800	240.916

5.2. Aplicación 2

Para la segunda aplicación, se consideró un conjunto de datos conformado por 210 pacientes pediátricos hospitalizados en la Unidad de Cuidados Intensivos (UCI) de una institución de salud de tercer nivel. Los pacientes fueron seleccionados por conveniencia, asegurándose de que contaran con mediciones completas en las variables de interés. El objetivo de esta aplicación fue estimar el efecto de la variable lactato, considerada como el principal factor asociado a la mortalidad del paciente.

5.2.1. Descripción de los datos

A continuación se describen las variables de interés para la aplicación:

1. Variable respuesta: Mortalidad: 1 indica que el paciente falleció, mientras que 0 indica que el paciente sobrevivió.

2. Variable explicativa: El lactato, siendo esta un marcador importante para el desenlace clínico de mortalidad. Además, se considera a otras variables como la edad del paciente, su sexo y si si tiene alguna comorbilidad.

5.2.2. Análisis exploratorio de los datos

En la Tabla 5.13 se presentan los resultados de las características de los 210 pacientes estudiados. Se estimó que el 27.1 % de los pacientes fallecieron. Se evidenció que las características

analizadas en la muestra tienen una asociación estadísticamente significativa ($p < 0{,}05$) con la variable de mortalidad. La variable lactato muestra un valor mediano mayor en el grupo de pacientes que tuvieron un desenlace fatal. En la Figura 5.3 se presentan las distribuciones de las características analizadas respecto a la mortalidad, destacando que la distribución de la variable lactato es asimétrica en ambos grupos, con aparentemente, un mayor número de valores extremos en el grupo de pacientes que no fallecieron.

Tabla 5.13: Análisis descriptivo bivariado de las características clínicas y su relación con la mortalidad en pacientes pediátricos de UCI.

Características	**N = 210**	**Mortalidad**		**Valor p**
		No = 153	**Si = 57**	
Lactato				0.001
Mediana (RIC)	3.4 (2.1 - 5.7)	2.8 (1.9 - 4.3)	6.8 (3.4 - 11.3)	
Sexo del paciente				0.001
Femenino	95.0 (45.2 %)	81.0 (52.9 %)	14.0 (24.6 %)	
Masculino	115.0 (54.8 %)	72.0 (47.1 %)	43.0 (75.4 %)	
Edad del paciente				0.011
Infantes	146.0 (69.5 %)	98.0 (64.1 %)	48.0 (84.2 %)	
Niños	50.0 (23.8 %)	44.0 (28.8 %)	6.0 (10.5 %)	
Adolescentes	14.0 (6.7 %)	11.0 (7.2 %)	3.0 (5.3 %)	
Comorbilidad				0.031
No	107.0 (51.0 %)	71.0 (46.4 %)	36.0 (63.2 %)	
Si	103.0 (49.0 %)	82.0 (53.6 %)	21.0 (36.8 %)	

5.2.3. Estimación de efectos con datos atípicos

Al igual que en la primera aplicación, en esta primera etapa de análisis se pretende estimar solo el efecto sólo de la variable lactato sobre la mortalidad en los pacientes. Se utilizarán todos los modelos propuestos, manteniendo las mismas configuraciones bayesianas que en la primera aplicación. Las muestras a posteriori se generarán mediante el método MCMC, ejecutando 40,000 iteraciones distribuidas en 4 cadenas de Markov. Para reducir la autocorrelación, se aplicarán saltos de 30 (thin = 30) y se descartarán las primeras 1,000 observaciones de cada cadena muestreada. La convergencia adecuada de las cadenas para cada parámetro estimado se verificará mediante la inspección visual de las series temporales y el estadístico de Geweke.

A continuación, se presentan los resultados de la estimación del efecto del lactato sobre la mortalidad, tanto en presencia como en ausencia de datos atípicos:

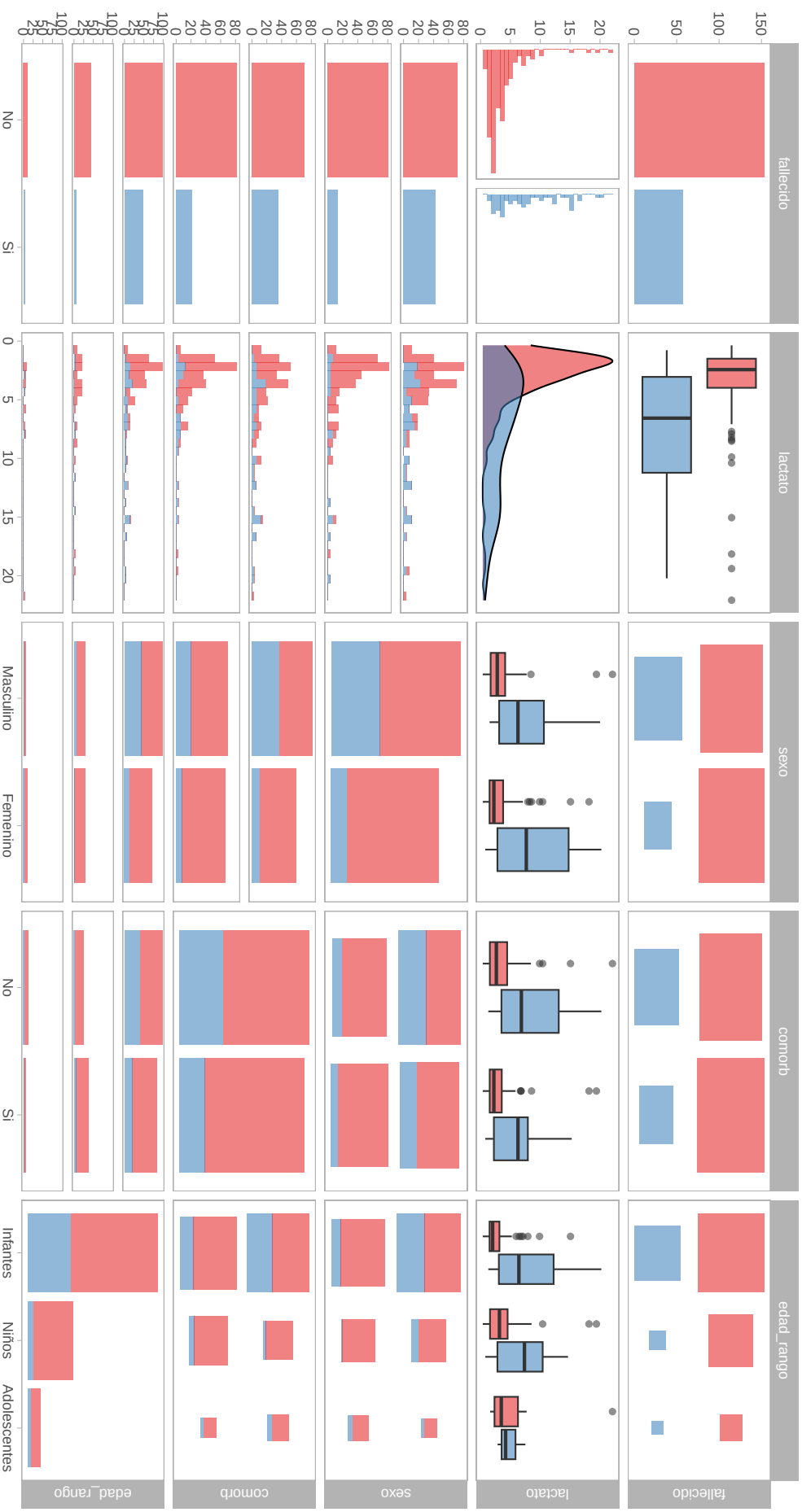

Figura 5.3: Análisis exploratorio de datos sobre la mortalidad y características clínicas de pacientes pediátricos.

Modelo 11: Regresión Probit con datos atípicos

Para el primer modelo de la segunda aplicación, se estimó el efecto de la variable predictora lactato sobre la mortalidad en niños utilizando un modelo probit considerado como el modelo de referencia. La especificación de este modelo es la siguiente:

$$\text{Mortalidad}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = \Phi(\beta_0 + \beta_1 \times \text{Lactato}_i),$$

donde $i = 1, 2, \ldots, 210$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.14 se muestra que el efecto estimado para la variable lactato fue de $\hat{\beta}_1 =$ 0,133 con un intervalo de credibilidad del 95 % comprendido entre 0.088 y 0.179, siendo este efecto estadísticamente significativo.

Tabla 5.14: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión probit bayesiano en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Probit	$\hat{\beta}_0$	-1.304	0.002	-1.606	-1.007
	$\hat{\beta}_1$	0.133	0.0003	0.088	0.179
DIC		211.5			
WAIC		214.4			

Modelo 12: Regresión Robit con datos atípicos

Se estimó el efecto de la variable predictora lactato sobre la mortalidad en niños utilizando el modelo Robit que emplea como función de enlace la distribución t-Student estándar. En este modelo, el parámetro de forma está fijo en 2. La especificación de este modelo es la siguiente:

$$\text{Mortalidad}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v(\beta_0 + \beta_1 \times \text{Lactato}_i),$$

donde $i = 1, 2, \ldots, 210$ y las distribuciones a prioris no informativas consideradas para los

coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.15 se muestra que el efecto estimado para la variable lactato fue de $\hat{\beta}_1 =$ 0,225 con un intervalo de credibilidad del 95 % comprendido entre 0.134 y 0.333, siendo este efecto estadísticamente significativo.

Tabla 5.15: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma $v = 2$ en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit	$\hat{\beta}_0$	-1.900	0.004	-2.565	-1.334
	$\hat{\beta}_1$	0.225	0.001	0.134	0.333
DIC		209.9			
WAIC		212.5			

Modelo 13: Regresión Robit con parámetro de forma aleatorio con datos atípicos

Se estimó el efecto de la variable predictora lactato sobre la mortalidad en niños utilizando un modelo que emplea como función de enlace la distribución t-Student estándar. En este modelo, el parámetro de forma se considera una variable aleatoria. La especificación de este modelo es la siguiente:

$$\begin{aligned} \text{Mortalidad}_i &\sim Bernoulli(\pi_i) \\ \pi_i &= F_v(\beta_0 + \beta_1 \times \text{Lactato}_i) \\ v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\ u &\sim U(1, 100), \end{aligned}$$

donde $i = 1, 2, \ldots, 210$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.16 se muestra que el efecto estimado para la variable lactato fue de $\hat{\beta}_1 =$ 0,198 con un intervalo de credibilidad del 95 % comprendido entre 0.099 y 0.411, siendo este efecto estadísticamente significativo.

Tabla 5.16: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma aleatorio en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit v aleatorio	$\hat{\beta}_0$	-1.732	0.012	-3.110	-1.080
	$\hat{\beta}_1$	0.198	0.002	0.099	0.411
	$\hat{v}$	26.989	1.725	1.034	185.445
DIC		210.5			
WAIC		213.5			

Modelo 14: Regresión Robit tG con datos atípicos

Se estimó el efecto de la variable predictora lactato sobre la mortalidad en niños utilizando un modelo que emplea como función de enlace la distribución t-Generalizada. En este modelo, el parámetro de forma está fijo en 2, mientras que el parámetro de escala se fija en 1. La especificación de este modelo es la siguiente:

$$\text{Mortalidad}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v[(\beta_0 + \beta_1 \times \text{Lactato}_i) \times \sqrt{v})],$$

donde $i = 1, 2, \ldots, 210$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.17 se muestra que el efecto estimado para la variable lactato fue de $\hat{\beta}_1 =$ 0,159 con un intervalo de credibilidad del 95 % comprendido entre 0.094 y 0.237, siendo este efecto estadísticamente significativo.

Tabla 5.17: Estimación del efecto de lactato sobre la mortalidad en niños mediante un modelo de regresión robit tG bayesiano con parámetro de forma $v_1 = 2$ y escala $v_2 = 1$ en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG	$\hat{\beta}_0$	-1.344	0.003	-1.811	-0.945
	$\hat{\beta}_1$	0.159	0.001	0.094	0.237
DIC		209.9			
WAIC		212.4			

Modelo 15: Regresión Robit tG con parámetro de forma aleatorio con datos atípicos

Se estimó el efecto de la variable predictora lactato sobre la mortalidad en niños utilizando un modelo que emplea como función de enlace la distribución t-Generalizada. En este modelo, el parámetro de forma se considera una variable aleatoria, mientras que el parámetro de escala se fija en 1. La especificación de este modelo es la siguiente:

$$\begin{aligned}
\text{Mortalidad}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= F_v[(\beta_0 + \beta_1 \times \text{Lactato}_i) \times \sqrt{v})] \\
v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\
u &\sim U(1, 100),
\end{aligned}$$

donde $i = 1, 2, \ldots, 210$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.18 se muestra que el efecto estimado para la variable lactato fue de $\hat{\beta}_1 =$ 0,221 con un intervalo de credibilidad del 95 % comprendido entre 0.024 y 0.446, siendo este efecto estadísticamente significativo.

Tabla 5.18: Estimación del efecto del lactato sobre la mortalidad en niños mediante un modelo de regresión robit tG bayesiano con parámetro de forma aleatorio y escala $v_2 = 1$ en presencia de datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG v aleatorio	$\hat{\beta}_0$	-1.799	0.036	-3.381	-0.235
	$\hat{\beta}_1$	0.221	0.005	0.024	0.446
	$\hat{v}$	3.472	1.069	1.010	33.740
DIC		209.4			
WAIC		212.1			

5.2.4. Estimación de efectos sin datos atípicos

Para la segunda parte de la aplicación 2, se llevó a cabo un nuevo modelamiento para estimar el efecto de la variable explicativa lactato sobre la variable respuesta mortalidad en niños. La metodología utilizada para la detección de datos atípicos fue la misma que se

describió en la primera aplicación.

Los resultados de las nuevas estimaciones sin datos atípicos se muestran a continuación:

Modelo 16: Regresión Probit sin datos atípicos

Se estimó el nuevo efecto de la variable predictora lactato sobre la mortalidad en niños, sin la presencia de datos atípicos. La especificación del modelo probit es la siguiente:

$$\text{Mortalidad}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = \Phi(\beta_0 + \beta_1 \times \text{Lactato}_i),$$

donde $i = 1, 2, \ldots, 206$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.19 se muestra que el efecto estimado para la variable lactato sin la presencia de datos atípicos fue de $\hat{\beta}_1 = 0{,}234$ con un intervalo de credibilidad del 95 % comprendido entre 0.164 y 0.310, siendo este efecto estadísticamente significativo.

Tabla 5.19: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión probit bayesiano sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Probit	$\hat{\beta}_0$	-1.693	0.003	-2.078	-1.322
	$\hat{\beta}_1$	0.234	0.001	0.164	0.310
DIC		185.8			
WAIC		187.6			

Modelo 17: Regresión Robit sin datos atípicos

Se estimó el nuevo efecto de la variable predictora lactato sobre la mortalidad en niños, sin la presencia de datos atípicos. La especificación del modelo robit con parámetro de forma fijo en 2 es la siguiente:

$$\text{Mortalidad}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v(\beta_0 + \beta_1 \times \text{Lactato}_i),$$

donde $i = 1, 2, \ldots, 206$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.20 se muestra que el efecto estimado para la variable lactato sin la presencia de datos atípicos fue de $\hat{\beta}_1 = 0{,}323$ con un intervalo de credibilidad del 95 % comprendido entre 0.212 y 0.456, siendo este efecto estadísticamente significativo.

Tabla 5.20: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma $v = 2$ sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit	$\hat{\beta}_0$	-2.354	0.005	-3.139	-1.703
	$\hat{\beta}_1$	0.323	0.001	0.212	0.456
DIC		187.1			
WAIC		189.1			

Modelo 18: Regresión Robit con parámetro de forma aleatorio sin datos atípicos

Se estimó el nuevo efecto de la variable predictora lactato sobre la mortalidad en niños, sin la presencia de datos atípicos. La especificación del modelo robit con parámetro de forma aleatorio es la siguiente:

$$\begin{aligned} \text{Mortalidad}_i &\sim Bernoulli(\pi_i) \\ \pi_i &= F_v(\beta_0 + \beta_1 \times \text{Lactato}_i) \\ v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\ u &\sim U(1, 100), \end{aligned}$$

donde $i = 1, 2, \ldots, 206$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron las siguientes: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.21 se muestra que el efecto estimado para la variable lactato sin datos atípicos fue de $\hat{\beta}_1 = 0{,}254$ con un intervalo de credibilidad del 95 % comprendido entre 0.169 y 0.408, siendo este efecto estadísticamente significativo.

Tabla 5.21: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión robit bayesiano con parámetro de forma aleatorio sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit v aleatorio	$\hat{\beta}_0$	-1.847	0.006	-2.890	-1.361
	$\hat{\beta}_1$	0.254	0.001	0.169	0.408
	$\hat{v}$	50.369	0.940	1.308	245.625
DIC		186.1			
WAIC		187.9			

Modelo 19: Regresión Robit tG sin datos atípicos

Se estimó el nuevo efecto de la variable predictora lactato sobre la mortalidad en niños, sin la presencia de datos atípicos. La especificación del modelo robit tG con parámetro de forma fijo en 2 y escala fijo en 1 es la siguiente:

$$\text{Mortalidad}_i \sim Bernoulli(\pi_i)$$
$$\pi_i = F_v[(\beta_0 + \beta_1 \times \text{Lactato}_i) \times \sqrt{v})],$$

donde $i = 1, 2, \ldots, 206$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$.

En la Tabla 5.22 se muestra que el efecto estimado para la variable lactato sin datos atípicos fue de $\hat{\beta}_1 = 0{,}228$ con un intervalo de credibilidad del 95 % comprendido entre 0.149 y 0.324, siendo este efecto estadísticamente significativo.

Tabla 5.22: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión robit tG bayesiano con parámetro de forma $v_1 = 2$ y escala $v_2 = 1$ sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG	$\hat{\beta}_0$	-1.661	0.004	-2.222	-1.203
	$\hat{\beta}_1$	0.228	0.001	0.149	0.324
DIC		187.1			
WAIC		189.1			

Modelo 20: Regresión Robit tG con parámetro de forma aleatorio sin datos atípicos

Se estimó el nuevo efecto de la variable predictora lactato sobre la mortalidad en niños, sin la presencia de datos atípicos. La especificación del modelo robit tG con parámetro de forma aleatorio y escala fijo en 1 es la siguiente:

$$\begin{aligned}
\text{Mortalidad}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= F_v[(\beta_0 + \beta_1 \times \text{Lactato}_i) \times \sqrt{v})] \\
v \mid u &\sim Exp\left(\frac{1}{u}, 1\right) \\
u &\sim U(1, 100),
\end{aligned}$$

donde $i = 1, 2, \ldots, 206$ y las distribuciones a prioris no informativas consideradas para los coeficientes de regresión fueron: $\beta_0 \sim N(0, 100^2)$ y $\beta_1 \sim N(0, 100^2)$, mientras que, para el parámetro de forma v se consideró como una exponencial desplazada en 1.

En la Tabla 5.23 se muestra que el efecto estimado para la variable lactato sin datos atípicos fue de $\hat{\beta}_1 = -1{,}349$ con un intervalo de credibilidad del 95 % comprendido entre -2.795 y -0.148, siendo este efecto estadísticamente significativo.

Tabla 5.23: Estimación del efecto del lactato sobre la mortalidad en niños, mediante un modelo de regresión robit tG bayesiano con parámetro de forma aleatorio y escala $v_2 = 1$ sin datos atípicos.

Modelo	**Parámetros**	**Estimación**	**Error estándar**	**IC95 %**	
				LI	**LS**
Robit tG v aleatorio	$\hat{\beta}_0$	-1.427	0.042	-3.519	-0.192
	$\hat{\beta}_1$	0.196	0.006	0.025	0.499
	$\hat{v}$	9.834	2.794	1.041	72.118
DIC		186.9			
WAIC		188.9			

5.2.5. Comparación de modelos

Para esta nueva aplicación, la Figura 5.4 presenta una comparación de las probabilidades estimadas a partir de los cinco modelos bayesianos, en escenarios con y sin datos atípicos en la variable explicativa lactato. En el primer escenario, se observa que las curvas de probabilidad estimada son distintas para cada modelo. Sin embargo, al analizar el conjunto de datos

excluyendo los casos más probables de ser atípicos, las estimaciones de probabilidad son muy similares. Este comportamiento es el esperado según la teoría.

Luego, en la Tabla 5.24, se muestran las estimaciones del WAIC para los datos originales (sin eliminar datos atípicos). Los modelos robustos presentan un mejor ajuste a los datos, destacándose el modelo Robit tG con parámetro de forma aleatorio, que obtuvo un valor de $WAIC = 212{,}120$. Por otro lado, en el escenario sin la presencia de datos atípicos, se observa que el modelo Probit tiene un menor valor de WAIC en comparación con los modelos robustos.

Tabla 5.24: Comparación de modelos bayesianos en presencia y ausencia de datos atípicos para la aplicación 2.

Datos originales					
Modelos	$\hat{\beta}_1$	**SE**	**IC95 %**	**DIC**	**WAIC**
Probit	0.133	0.000	(0.088 ; 0.179)	211.500	214.400
Robit	0.225	0.001	(0.134 ; 0.333)	209.900	212.530
Robit v aleatorio	0.198	0.002	(0.100 ; 0.411)	210.500	213.482
Robit tG	0.159	0.001	(0.094 ; 0.237)	209.900	212.434
Robit tG v aleatorio	0.221	0.005	(0.024 ; 0.446)	209.400	212.120
Datos sin presencia de atípicos					
Modelos	$\hat{\beta}_1$	**SE**	**IC95 %**	**DIC**	**WAIC**
Probit	0.234	0.000	(0.164 ; 0.310)	185.800	187.641
Robit	0.323	0.001	(0.212 ; 0.456)	187.100	189.132
Robit v aleatorio	0.254	0.001	(0.169 ; 0.408)	186.100	187.931
Robit tG	0.228	0.001	(0.149 ; 0.324)	187.100	189.135
Robit tG v aleatorio	0.196	0.006	(0.025 ; 0.499)	186.900	188.920

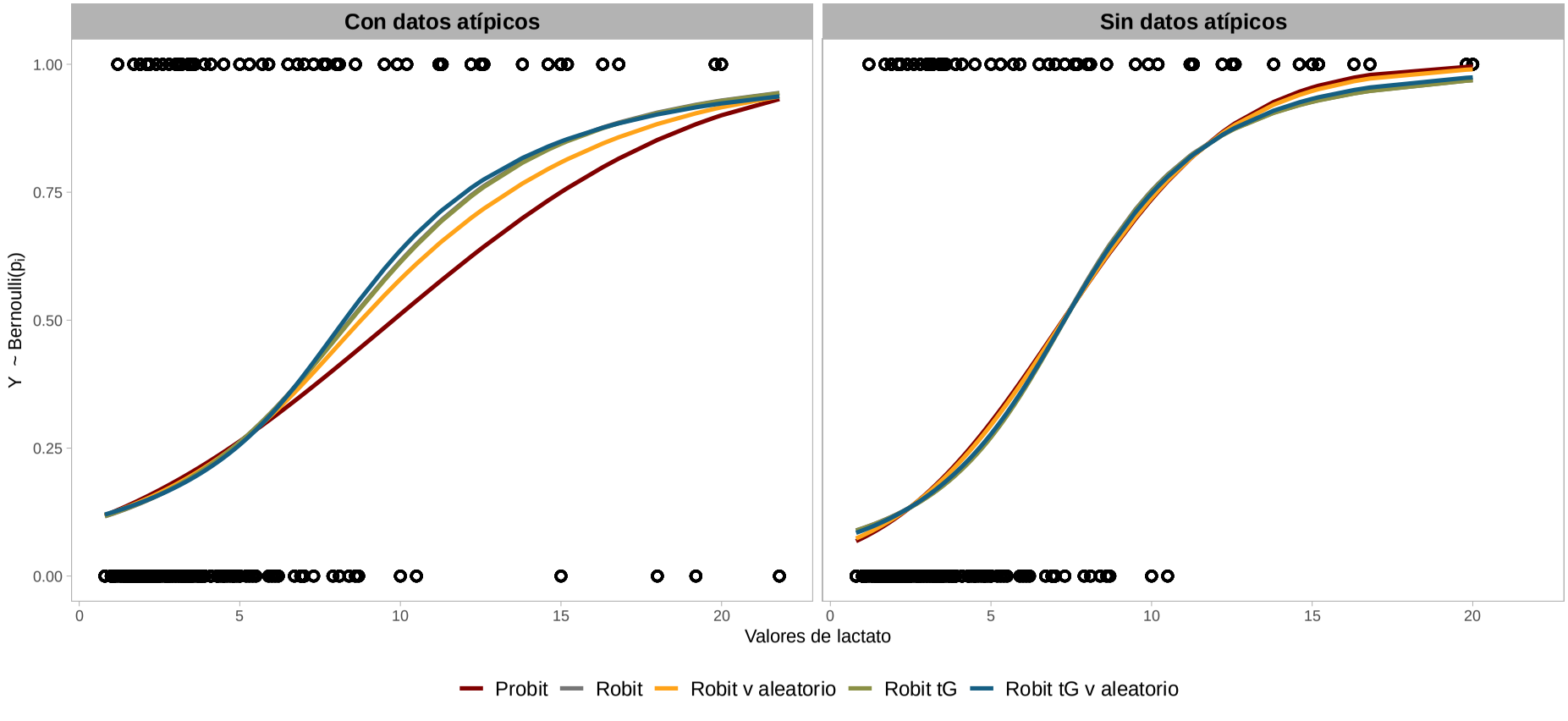

Figura 5.4: Comparación de las estimaciones de probabilidades con y sin datos atípicos para la presencia de mortalidad en pacientes pediátricos.

Capítulo 6

Conclusiones

Las principales conclusiones obtenidas a partir de los modelos bayesianos Probit, Robit, Robit con parámetro de forma aleatorio, Robit tG y Robit tG con parámetro de forma aleatorio, estudiados en esta investigación son las siguientes:

Conclusiones obtenidas para el estudio de simulación:

- Se estimó el efecto de una variable cuantitativa sobre una variable binaria, los cuales fueron simulados a partir de un modelo Probit. El efecto de esta variable fue estimado a partir de 5 modelos bayesianos, siendo el modelo Probit, el modelo considerado como el de referencia para la comparación de las estimaciones de los coeficientes de regresión y de la probabilidad de éxito. Los modelos que se utilizaron para comprobar su eficacia sobre el modelo de referencia fue el modelo Robit con parámetro de forma fijo en 2, el modelo Robit con parámetro de forma considerado como una variable aleatoria el cual fue estimado en función a los datos, el modelo Robit tG el cual utiliza como función de enlace a la distribución acumulada de la t Student Generalizada con parámetro de forma fijo en 2 y parámetro de escala fijo en 1 y el último modelo fue el Robit tG con parámetro de forma considerado como variable aleatoria que también será estimado a partir de los datos y con parámetro de escala fijo en 1.

- El modelo Probit tiende a ser más eficaz en escenarios sin datos atípicos; sin embargo, comienza a perder eficiencia en la estimación de los coeficientes de regresión y de la probabilidad de éxito a partir de un 2 % de contaminación con datos atípicos. En algunos casos, los modelos robustos tienden a sobreestimar los verdaderos parámetros utilizados en la simulación.

- En cuanto a la recuperación de los parámetros de regresión en presencia de datos atípicos, los modelos robustos muestran estimaciones menos precisas con tamaños de muestra

pequeños ($n = 50$) en comparación con tamaños de muestra mayores. Sin embargo, incluso en tamaños de muestra pequeños, las estimaciones de los modelos robustos son mejores que las del modelo Probit. Las estimaciones de la probabilidad de éxito en los diferentes escenarios de contaminación y de tamaños de muestra, tienen menor sesgo en comparación al modelo Probit, al tratar de recuperar la verdadera forma de la probabilidad real utilizada en la simulación de los datos.

- En escenarios con una baja proporción de datos atípicos, los modelos robustos logran una buena recuperación de parámetros reales. No obstante, a medida que la proporción de datos atípicos supera el 6 % de contaminación, el sesgo en la estimación de los coeficientes de regresión y de la probabilidad de éxito aumenta. A pesar de esto, las estimaciones de los modelos robustos siguen siendo mucho mejores que las del modelo Probit, que pierde eficiencia a partir de una contaminación del 2 %, haciendo que las estimaciones puntuales de los coeficientes de regresión sean subestimadas y que el comportamiento de la estimación de probabilidad sea muy lineal.

- Las estimaciones puntuales y de intervalos de credibilidad para el parámetro de forma en los modelos Robit y Robit tG son mayores en escenarios sin datos atípicos (0 %) para los distintos tamaños de muestra, en comparación a escenarios con datos atípicos. A medida que la contaminación de datos atípicos va en aumento, estos modelos tienen un mejor control de la incertidumbre (menor amplitud en los intervalos de credibilidad).

Conclusiones obtenidas para los casos de aplicación:

- El objetivo de las aplicaciones fue estimar el efecto de una variable explicativa cuantitativa sobre una variable binaria utilizando modelos bayesianos robustos y comparándolos con el modelo Probit bayesiano de referencia. Los resultados mostraron un cambio en la estimación de los efectos puntuales de la variable explicativa en ambos escenarios, con y sin datos atípicos. Estas diferencias en las estimaciones también varían según el modelo bayesiano utilizado.

- En los dos casos de aplicaciones con datos reales, se observó que las estimaciones obtenidas por los modelos bayesianos robustos son superiores (menores valores de WAIC) al modelo de referencia cuando hay presencia de datos atípicos. El modelo bayesiano robusto con mejor ajuste a los datos fue el Robit tG con parámetro de forma aleatorio, que tuvo un valor de WAIC menor que los demás modelos.

- En ausencia de datos atípicos, el modelo de referencia tiende a tener un mejor o igual ajuste (valores de WAIC menores o muy cercanos a los de los modelos robustos) en comparación con los modelos robustos.

- Gráficamente, se evidenció un cambio en la forma de la estimación de la probabilidad de éxito en presencia y ausencia de datos atípicos para los diferentes modelos bayesianos utilizados. Cuando hay datos atípicos, las estimaciones obtenidas por los modelos robustos tienden a ser muy similares entre sí y diferentes del modelo de referencia. Al eliminar los datos potencialmente anómalos, las estimaciones de la probabilidad obtenidas por cada modelo tienen la misma forma.

Apéndice A

Código R para datos de simulación

A.1. Simulación de datos

Dado que se están manejando múltiples escenarios de simulación con distintos tamaños de muestra y niveles de contaminación por datos atípicos para cada uno de los modelos propuestos, en este apartado se considerará únicamente un escenario. Específicamente, se analizará un tamaño de muestra de 200 con un nivel de contaminación del 6 %. Se adjuntará el código en R y JAGS necesario para la estimación de los coeficientes de regresión en cada modelo bayesiano.

Librerías

```
rm(list=ls())
graphics.off()
library(tidyverse)
library(recipes)
library(bayesplot)
library(rjags)
library(coda)
library(doMC)
library(ggsci)
library(rio)

# Simulando datos a partir de una probit
simulacion <- function(n){
beta  <- c(2,3)
x     <- rnorm(n)
X     <- cbind(rep(1,n),x)
```

```
    pi    <- pnorm(X%*%beta)
    y     <- rbinom(n,1,pi)
    datos <- data.frame(x=x,y=y)}
```

Generando datos contaminados (6 %)

```
n=200
set.seed(321)
datos<-simulacion(n)
datos<-datos%>%
    mutate(x2 =
       case_when(
         y==1 & x>=quantile(x,probs=0.97) ~ -1*x,
         y==0 & x<=quantile(x,probs=0.03) ~ abs(x),
         TRUE ~ x))%>%
    mutate(atipico=
       case_when(
         y==1 & x>=quantile(x,probs=0.97) ~ 1,
         y==0 & x<=quantile(x,probs=0.03) ~ 1,
         TRUE ~ 2)) %>%
    mutate(atipico =
       factor(atipico,
              levels = c(1,2),
              labels = c("Atípico","No atípico")))
```

A.2. Modelos de regresión binario bayesiano con JAGS

A.2.1. Modelo Probit Bayesiano

Especificación de datos

```
N <-nrow(datos)
data_train <- list(y=datos$y,
                   N=N,
                   x=datos$x2,
                   mu.beta=0,
                   tau.beta=0.0001)
```

Especificación del modelo

```
modelo_probit <- textConnection("model{
    # Función de verosimilitud:
    for (i in 1:N) {
        y[i]~dbern(p[i])
        probit(p[i]) <- beta0 + beta1*x[i]
        verosimilitud[i] <- dbin(y[i],p[i],1)
    }
    # A Prioris:
    beta0 ~ dnorm(mu.beta,tau.beta)
    beta1 ~ dnorm(mu.beta,tau.beta)}")
```

Compilación del modelo

```
model1 <- jags.model(modelo_probit,
                     data = data_train,
                     n.chains=4,
                     quiet=TRUE,
                     n.adapt = 0)
```

Actualización y resumen del modelo

```
registerDoMC(cores = 6)
update(model1,1000,progress.bar="none")
covariables <- c("beta0","beta1")
samples_probit <- coda.samples(model1,
                               variable.names=covariables,
                               n.iter=10000,
                               progress.bar="none",
                               thin = 2)
summary(samples_probit)
```

A.2.2. Modelo Robit Bayesiano

Especificación de datos

```
N <-nrow(datos)
data_train <- list(y=datos$y,
                   N=N,
                   df=2,
                   tau=1,
                   mu=0,
                   x=datos$x2,
                   mu.beta=0,
                   tau.beta=0.0001)
```

Especificación del modelo

```
modelo_robit <- textConnection("model{
# Función de verosimilitud:
      for (i in 1:N) {
          y[i]~dbern(p[i])
          p[i] <- pt(beta0 + beta1*x[i],mu,tau,df)
          verosimilitud[i] <- dbin(y[i],p[i],1)
         }
# A Prioris:
    beta0 ~ dnorm(mu.beta,tau.beta)
    beta1 ~ dnorm(mu.beta,tau.beta)}")
```

Compilación del modelo

```
model2 <- jags.model(modelo_robit,
                     data = data_train,
                     n.chains=4,
                     quiet=TRUE,
                     n.adapt = 0)
```

Actualización y resumen del modelo

```
registerDoMC(cores = 6)
update(model2,1000,progress.bar="none")
covariables <- c("beta0","beta1")
samples_robit <- coda.samples(model2,
                              variable.names=covariables,
                              n.iter=10000,
                              progress.bar="none",
                              thin = 2)
summary(samples_robit)
```

A.2.3. Modelo Robit Bayesiano v aleatorio

Especificación de datos

```
N <-nrow(datos)
data_train <- list(y=datos$y,
                   N=N,
                   tau=1,
                   mu=0,
                   x=datos$x2,
                   mu.beta=0,
                   tau.beta=0.0001)
```

Especificación del modelo

```
modelo_robit <- textConnection("model{

   # Función de verosimilitud:
     for (i in 1:N) {
     y[i]~dbern(p[i])
     p[i] <- pt(beta0 + beta1*x[i],mu,tau,df)
     verosimilitud[i] <- dbin(y[i],p[i],1)
     }

   # A Prioris:
     beta0 ~ dnorm(mu.beta,tau.beta)
```

```
        beta1 ~ dnorm(mu.beta,tau.beta)
        u     ~ dunif(1,100)
        k0    ~ dexp(1/u)
        df    <- 1+k0
        }")
```

Compilación del modelo

```
model1 <- jags.model(modelo_robit,
                     data = data_train,
                     n.chains=4,
                     quiet=TRUE,
                     n.adapt = 0)
```

Actualización y resumen del modelo

```
registerDoMC(cores = 6)
update(model1,1000,progress.bar="none")
covariables <- c("beta0","beta1","df")
samples_robit1 <- coda.samples(model1,
                               variable.names=covariables,
                               n.iter=10000,
                               progress.bar="none",
                               thin = 2)
summary(samples_robit1)
```

A.2.4. Modelo Robit tG Bayesiano

Especificación de datos

```
N <-nrow(datos)
data_train <- list(y=datos$y,
                   N=N,
                   k=2,      # Parámetro de forma
                   tau=1,    # Parámetro de escala
                   mu=0,
```

```
                        x=datos$x2,
                        mu.beta=0,
                        tau.beta=0.0001)
```

Especificación del modelo

```
modelo_tG <- textConnection("model{
  # Función de verosimilitud:
    for (i in 1:N) {
      y[i]~dbern(p[i])
      p[i] <- pt((beta0 + beta1*x[i])*sqrt(k),mu,tau,k)
      verosimilitud[i] <- dbin(y[i],p[i],1)
      }
  # A Prioris:
  beta0 ~ dnorm(mu.beta,tau.beta)
  beta1 ~ dnorm(mu.beta,tau.beta)
  }")
```

Compilación del modelo

```
model3 <- jags.model(modelo_tG,
                     data = data_train,
                     n.chains=4,
                     quiet=TRUE,
                     n.adapt = 0)
```

Actualización y resumen del modelo

```
registerDoMC(cores = 6)
update(model3,1000,progress.bar="none")
covariables <- c("beta0","beta1")
samples_tG <- coda.samples(model3,
                           variable.names=covariables,
                           n.iter=10000,
                           progress.bar="none",
```

```
                                thin = 2)
summary(samples_tG)
```

A.2.5. Modelo Robit tG Bayesiano v aleatorio

Especificación de datos

```
N <-nrow(datos)
data_train <- list(y=datos$y,
                   N=N,
                   tau=1,
                   mu=0,
                   x=datos$x2,
                   mu.beta=0,
                   tau.beta=0.0001)
```

Especificación del modelo

```
modelo_robit <- textConnection("model{
   # Función de verosimilitud:
     for (i in 1:N) {
     y[i]~dbern(p[i])
     p[i] <- pt(beta0 + beta1*x[i],mu,tau,df)
     verosimilitud[i] <- dbin(y[i],p[i],1)
     }
  # A Prioris:
    beta0 ~ dnorm(mu.beta,tau.beta)
    beta1 ~ dnorm(mu.beta,tau.beta)
    u     ~ dunif(1,100)
    k0    ~ dexp(1/u)
    df    <- 1+k0
}")
```

Compilación del modelo

```
model1 <- jags.model(modelo_robit,
                     data = data_train,
                     n.chains=4,
                     quiet=TRUE,
                     n.adapt = 0)
```

Actualización y resumen del modelo

```
registerDoMC(cores = 6)
update(model1,1000,progress.bar="none")
covariables <- c("beta0","beta1","df")
samples_robit1 <- coda.samples(model1,
                               variable.names=covariables,
                               n.iter=10000,
                               progress.bar="none",
                               thin = 2)
summary(samples_robit1)
```

Bibliografía

Ahsanullah, M., Kibria, B. G. y Shakil, M. (2014). *Normal and student's t distributions and their applications*, Vol. 4, Springer.

Arellano-Valle, R. B. y Bolfarine, H. (1995). On some characterizations of the t-distribution, *Statistics & Probability Letters* **25**(1): 79–85.

Brooks, S., Gelman, A., Jones, G. y Meng, X.-L. (2011). *Handbook of markov chain monte carlo*, CRC press.

Czado, C. y Santner, T. J. (1992). The effect of link misspecification on binary regression inference, *Journal of Statistical Planning and Inference* **33**(2): 213–231.
URL: *https://www.sciencedirect.com/science/article/pii/0378375892900695*

Gelman, A., Carlin, J. B., Stern, H. S. y Rubin, D. B. (2014). Bayesian data analysis (vol. 2).

Gelman, A. y Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical science* **7**(4): 457–472.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *Technical report*, Federal Reserve Bank of Minneapolis.

Kim, S., Chen, M.-H. y Dey, D. K. (2008). Flexible generalized t-link models for binary response data, *Biometrika* **95**: 93–106.

Lange, K. L., Little, R. J. A. y Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution, *Journal of the American Statistical Association* **84**(408): 881–896.
URL: *http://www.jstor.org/stable/2290063*

Li, R. y Nadarajah, S. (2020). A review of student'st distribution and its generalizations, *Empirical Economics* **58**: 1461–1490.

Li, Z., Liao, H., Tang, R., Li, G., Li, Y. y Xu, C. (2023). Mitigating the impact of outliers in traffic crash analysis: A robust bayesian regression approach with application to tunnel crash data, *Accident Analysis & Prevention* **185**: 107019.
URL: *https://www.sciencedirect.com/science/article/pii/S0001457523000660*

Liang, F., Liu, C. y Carroll, R. (2011). *Advanced Markov chain Monte Carlo methods: learning from past samples*, Vol. 714, John Wiley & Sons.

Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression, *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives* pp. 227–238.

Liu, C. y Rubin, D. (1995). Ml estimation of the multivariate t distribution with unknown degrees of freedom, *Statistica Sinica* **5**: 19–39.

Peyhardi, J. (2020). Robustness of student link function in multinomial choice models, *Journal of Choice Modelling* **36**: 100228.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves, *Biometrics* pp. 761–768.

Ross, K. (2022a). An introduction to bayesian reasoning and methods, pp. 268–275.

Ross, S. M. (2022b). *Simulation*, Academic Press.

Roy, V. (2012). Convergence rates for mcmc algorithms for a robust bayesian binary regression model, *Electronic Journal of Statistics* **6**: 2463 – 2485.

Rubin, D. B. (1983). Iteratively reweighted least squares, entry in encyclopedia of the statistical sciences, *Encyclopedia of statistical sciences* .

Schlüter, S. y Fischer, M. (2012). A tail quantile approximation for the student t distribution, *Communications in Statistics-Theory and Methods* **41**(15): 2617–2625.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. y Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**(4): 583–639.

Wang, X. y Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption.

Watanabe, S. y Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory., *Journal of machine learning research* **11**(12).

West, M. (1984). Outlier models and prior distributions in bayesian linear regression, *Journal of the Royal Statistical Society: Series B (Methodological)* **46**(3): 431–439.