PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MODELOS DE REGRESIÓN A LA MEDIA CON EFECTOS MIXTOS PARA VARIABLE RESPUESTA SEMICONTINUA

Tesis para optar el grado de Magíster en Estadística

AUTOR

Luis Alberto Bautista Bautista

ASESOR

Dr. Luis Hilmar Valdivieso Serrano

JURADO

Dr. Cristian Luis Bayes Rodríguez
Dra. Rocío Paola Maehara Aliaga de Benites
Dr. Luis Hilmar Valdivieso Serrano

LIMA - PERÚ

2020

Dedicatoria

A mi madre por la dedicación, paciencia y apoyo que me brinda día tras día para seguir adelante en este largo camino de la vida académica.

A mi hermanita María Lourdes, por ser la razón, el motivo y la mayor inspiración para creer en los sueños.

A mis amigos y compañeros de trabajo que me dieron aliento y ánimo durante los estudios de la maestria.

Agradecimentos

Quiero expresar mi más sincero agradecimiento a mi asesor Dr. Luis Hilmar Valdivieso. Por su apoyo, voluntad, paciencia y dedicación en el desarrollo de la presente tesis, estaré eternamente agradecido.

A todos mis profesores de la maestria por sus enseñanzas, consejos y sobre todo por la motivación hacia la investigación.

A mis amigos y compañeros de vida: Sandra Cabrera, Jessica Chavez y Daniel Gavidia, gracias por sus consejos y sus enseñanzas, gracias por todo.

Resumen

En muchas situaciones se dispone de una variable aleatoria continua no negativa con asimetría positiva que eventualmente podriá tomar el valor cero. Datos de esta naturaleza son llamados semicontinuos o cero-inflacionados y fueron tradicionalmente modelados usando el modelo de regresión de dos partes propuesto por Duan et al. (1983). En este modelo la variable respuesta sigue una distribución mixta de probabilidades conformada por una distribución de Bernoulli y una distribución continua no negativa. Una versión longitudinal de este modelo de regresión, pero que apunta a explicar la media de la variable de respuesta, fue propuesto por Smith et al. (2017). Este modelo planteaba, para su componente continua de respuesta, una distribución Log Skew Normal.

El objetivo de este trabajo es estudiar un modelo alternativo al de Smith et al. (2017), que llamaremos, en general, un modelo de regresión a la media con efectos mixtos para respuestas semicontinuas, pues plantea una parametrización que permite estimar e interpretar los efectos de un conjunto de covariables sobre la media de las respuestas y no sobre la media condicionada a valores positivos. A diferencia del modelo de Smith et al. (2017), que hace uso de la distribución Log Skew Normal cero-inflacionada, nosotros modelaremos la respuesta con una distribución Gamma Generalizada cero-inflacionada. Este modelamiento, como se muestra, permite capturar de manera flexible ciertas características de los datos de respuesta, tales como, la asimetría y el comportamiento de las colas.

Los resultados del estudio de simulación para el nuevo modelo mostraron un adecuado desempeño en la recuperación de sus parámetros, donde para la estimación de estos utilizamos un enfoque bayesiano y el uso de métodos MCMC Hamiltonianos. Por último, los resultados de su aplicación en el estudio longitudinal del efecto que ciertas variables podrían ejercer sobre la media de los gastos en educación de los hogares en el Perú, mostraron un mejor ajuste a los datos respecto al modelo de Smith et al. (2017), en base a los criterios de información ampliamente aplicado y de validación cruzada de Leave-one-out.

Palabras clave: variable semicontinua o cero-inflacionada, distribución gamma generalizada, regresión de dos partes, enfoque bayesiano, distribución Log Skew Normal, MCMC, gasto en educación.

Índice general

Lista de Abreviaturas

MRMS	Modelo de regresión a la media con efectos mixtos para respuestas semicontinuas.
MRMS-GCI	MRMS con distribución gamma cero-inflacionada.
MRMS-GGCI	MRMS con distribución gamma generalizada cero-inflacionada.
MRMS-LSNCI	MRMS con distribución Log Skew Normal cero-inflacionada.
MRCI	Modelo de regresión a la media cero-inflacionado.
LSN	Log Skew Normal.
GG	Distribución gamma generalizada.
GGCI	Distribución gamma generalizada cero-inflacionado.
RMSE	Raíz del error cuadrático medio.
DIC	Criterio de Información de Desvío.
WAIC	Criterio de Información de Watanabe.
LOO-CV	Validación cruzada de Leave-One-Out.
MCMC	Métodos de cadenas de Markov Monte Carlo.
HMC	Monte Carlo Hamiltoniano.
MV	Máxima verosimilitud
IW	Distribución Inversa de Wishart
ENAHO	Encuesta Nacional de Hogares.
ENDES	Encuesta Demográfica y de Salud Familiar.
INEI	Instituto Nacional de Estadística e Informática.
JAGS	Just Another Gibbs Sampler.
INLA	Integrated Nested Laplace Approximation.

Lista de Símbolos

Y_{ij}	Respuesta del i-ésimo sujeto para la j-ésima medición.
$\mathbf{b}_i$	Vector de efectos aleatorios.
$\mathbf{\Sigma}$	Matriz de varianza y covarianza.
κ	Kappa.
σ	Sigma.
α	Alpha.
β	Beta.
η	Eta.
λ	Lambda.
ω	Omega.
δ	Delta.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En muchas situaciones los investigadores cuentan con datos no negativos de asimetría positiva, que eventualmente podrián contener al valor cero. Datos de esta naturaleza son llamados semicontinuos o cero-inflacionados y se presentan en muchos campos como la Economía, la Biología, etc. Se pueden encontrar, por ejemplo al modelar el gasto de los hogares destinados a educación o salud, la proporción de deuda sobre el monto del préstamo en clientes de un banco, la biomasa en una zona de pesca, la cantidad de lluvia en una zona, etc. Si bien las variables en estos ejemplos toman valores continuos positivos, es posible también encontrar hogares, clientes o zonas en las que la variable de interés tome el valor cero. Ante esta situación existe la necesidad de explicar o predecir variables con estas características.

La distribución de una variable aleatoria como la anterior esta formalmente constituida por la mixtura de una distribución de Bernoulli que explica si la respuesta toma el valor de cero o no y una distribución continua positiva.

La posibilidad de observar el valor de cero tiene como consecuencia que distribuciones continuas como, por ejemplo, la gamma, log-normal o Weibull, presenten limitaciones para modelar una variable semicontinua, pues el soporte de estas distribuciones esta limitado al intervalo abierto $(0, \infty)$.

La regresión sobre una respuesta semicontinua se ha tradicionalmente analizado mediante el modelo de dos partes, propuesto por Duan et al. (1983). Este modelo asume que la variable respuesta Y presenta una distribución de mixtura de probabilidades conformada por una distribución de Bernoulli, que explique si Y toma el valor de cero o no, y una distribución continua positiva para cuando esta última es estrictamente positiva. Este modelo plantea además ecuaciones de regresión para explicar la probabilidad $\pi = P(Y = 0)$ de una respuesta cero y la media de la respuesta condicionada a valores positivos $E(Y|Y > 0)$.

Si bien los modelos de regresión de dos partes parecen ser una opción ideal para modelar datos semicontinuos, pues incluyen un conjunto de parámetros para la respuesta binaria y un segundo conjunto para la parte continua condicionada a una respuesta positiva, el problema radica en combinar estas dos partes para medir el efecto de las covariables sobre la

media total, lo cual nos lleva a una dificultad en la interpretación directa de los efectos de las covariables sobre la media global.

En muchos casos, efectivamente, el principal interés de los investigadores radica en examinar tales efectos en la media global para extraer conclusiones sobre el impacto de los predictores en la población. Por ejemplo, en los estudios económicos de los gastos de atención en salud en todo el Perú, los investigadores y los políticos tal vez deseen comprender el efecto promedio en los gastos médicos de aumentar los precios de atención especializada en toda la población afectada. El modelo de dos partes estima por separado los efectos para la probabilidad de incurrir en gastos y el nivel de gastos dado que se incurre en ellos. En particular, una intervención puede tener un efecto sobre la probabilidad de ocurrencia pero el efecto opuesto sobre la ocurrencia de intensidad dada. En tales casos, los responsables de las políticas pueden quedar sin una verdadera comprensión del efecto general a nivel de la población de dicha intervención.

Para superar estas limitaciones Smith et al. (2014) propuso el modelo marginalizado de dos partes, el cual permite obtener estimaciones más interpretables de los efectos al parametrizar el modelo en términos de la media global. Este modelo mantiene muchas de las características importantes de los modelos convencionales de dos partes, como la captura de los valores cero-inflacionados y la asimetría. Además permite a los investigadores examinar los efectos de las covariables en la media general, que es un objetivo de interés primario en muchos aplicaciones.

Tanto Duan et al. (1983) como Smith et al. (2014) aplicaron su modelo a datos de corte transversal. Estos modelos se han extendido sin embargo para aplicaciones con datos longitudinales y agrupados, siendo uno de los primeros en proponerlo Olsen y Schafer (2001) mediante la introducción de coeficientes aleatorios en la parte binaria y continua. Estos modelos fueron luego extendidos por Smith et al. (2017) quienes propusieron un modelo de dos partes marginalizado para datos longitudinales que permita a los investigadores obtener el efecto de las covariables sobre la media general de la población. Este último modelo proporciona además estimaciones de la media general en la escala original y brinda una interpretación específica para cada sujeto.

Para el presente proyecto de tesis proponemos un modelo similar al modelo propuesto por Smith et al. (2017), que denominaremos modelo de regresión a la media con efectos mixtos para variables semicontinuas y que denotaremos en adelante por MRMS. Este modelo nos permite estimar en un solo proceso de optimización los efectos de un conjunto de covariables sobre la media total de la respuesta, $E(Y)$, contrario al modelo de dos partes que mide la influencia sobre la media condicional, $E(Y|Y > 0)$ y la probabilidad $P(Y = 0) = \pi$. Para el desarrollo de este modelo consideraremos en la componente continua una distribución gamma generalizada (GG), tomando como referencia el trabajo desarrollado por Vásquez (2018) para datos de tipo transversal.

Para estimar directamente la media de la variable respuesta se propone utilizar una parametrización similar a la dada por Bayes y Valdivieso (2016) para respuestas acotadas, parametrización que fue también utilizada por Smith et al. (2017) en su modelo.

La diferencia entre el modelo propuesto por Smith et al. (2017) con el que desarrollaremos en esta tesis radica esencialmente en la distribución de la componente continua del modelo, la cual será Gamma Generalizada para nuestro trabajo a diferencia de la distribución Log Skew Normal (LSN) desarrollada por Smith et al. (2017). En tal sentido, denotaremos de aquí en adelante por MRMS-LSNCI al modelo de Smith et al. (2017) y por MRMS-GGCI al modelo porpuesto en este trabajo. Al igual que en el MRMS-LSNCI, adoptaremos para la estimación de parámetros de nuestro modelo un enfoque Bayesiano, sin embargo, nosotros buscaremos realizar este proceso por medio de un algoritmo Hamiltoniano MCMC implementado en Stan a diferencia de Smith et al. (2017) que uso la rutina PROC MCMC del SAS.

El aporte fundamental de este proyecto, consiste en extender el modelo de regresión a la media para datos semicontinuos, con componente continua de distribución gamma generalizada (GG) aplicado a datos longitudinales o por conglomerados, incorporando efectos aleatorios y analizando su estimación mediante técnicas bayesianas de mayor eficiencia.

1.2. Objetivos de la tesis

El objetivo general de la tesis es presentar el modelo de regresión a la media con efectos mixtos para respuestas semicontinuas de distribución gamma generalizada cero-inflacionada (MRMS-GGCI), estudiando sus propiedades; estimando sus parámetros, mediante métodos bayesianos, y aplicándolo a un conjunto de datos reales de tipo longitudinal.

- Revisar la literatura acerca de las diferentes propuestas de modelos de regresión usados para explicar una variable semicontinua, identificando sus ventajas y limitaciones.
- Estudiar las propiedades e implementar la estimación del modelo de regresión a la media con efectos mixtos semicontinuos, utilizando un algoritmo de Monte Carlo Hamiltoniano que se encuentra implementado en el Stan.
- Realizar estudios de simulación bajo diferentes escenarios para el modelo propuesto y así evaluar el desempeño de sus estimadores.
- Aplicar el modelo a un conjunto de datos reales.

1.3. Organización de la tesis

La tesis se ha organizado en 6 capítulos. En el capítulo 2 introducimos los conceptos básicos para entender el MRMS-GGCI. En el capítulo 3 se presenta el MRMS-GGCI, el método de estimación de los parámetros y los métodos de selección de modelos. En el capítulo 4 se hace un estudio de simulación para distintos escenarios. En el capítulo 5 mostramos una aplicación del modelo propuesto para datos de la encuesta nacional de hogares y comparamos sus resultados con las del modelo MRMS-LSNCI. Finalmente, en el capítulo 6 mostramos las conclusiones de esta tesis y algunas sugerencias para futuras investigaciones.

Capítulo 2

Conceptos básicos

En el presente capítulo se hace un desarrollo previo de la distribución gamma generalizada, distribuciones cero-inflacionadas y la distribución gamma generalizada cero-inflacionada (GGCI). Esta última será la distribución que se tomará en cuenta para el modelamiento de los datos de carácter continuo en nuestro modelo. Los conceptos antes mencionados nos permitirán estudiar el modelo de regresión a la media con efectos mixtos semicontinuos, el cuál se desarrollará en el capítulo 3.

2.1. Distribución gamma

La distribución gamma es una distribución de probabilidad muy útil para modelar variables aleatorias continuas positivas, con asimetría positiva. La distribución gamma podría describir, por ejemplo, el tiempo que transcurre hasta que falle un dispositivo electrónico. Más aún, si la mayoría de estos dispositivos electrónicos tienden a fallar tempranamente, o muy cerca de su tiempo esperado de vida, pero unos pocos tardan más en fallar, esta distribución podría ser ideal debido a que posee colas ligeras.

La distribución gamma se define por sus parámetros de forma $\alpha > 0$ y escala $\beta > 0$. Una variable aleatoria continua Z tiene una distribución gamma de parámetros α y β, que se denota por $Z \sim G(\alpha, \beta)$, si su función de densidad está dada por:

$$f_Z(z; \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \left(\frac{1}{\beta}\right)^{\alpha} z^{\alpha-1} \exp\left(-\frac{z}{\beta}\right), \quad z > 0,$$

donde $\Gamma(\cdot)$ es la función gamma definida como $\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} \exp(-z)\, dz$. El parámetro α tiene una mayor influencia en el apuntamiento de la densidad; en cambio, el parámetro β tiene mayor infuencia en la propagación o alcance de la distribución tal y como se puede observar en la figura 2.1. La media y varianza de esta variable vienen dadas respectivamente por:

$$E(Z) = \alpha\beta,$$

$$V(Z) = \alpha\beta^2.$$

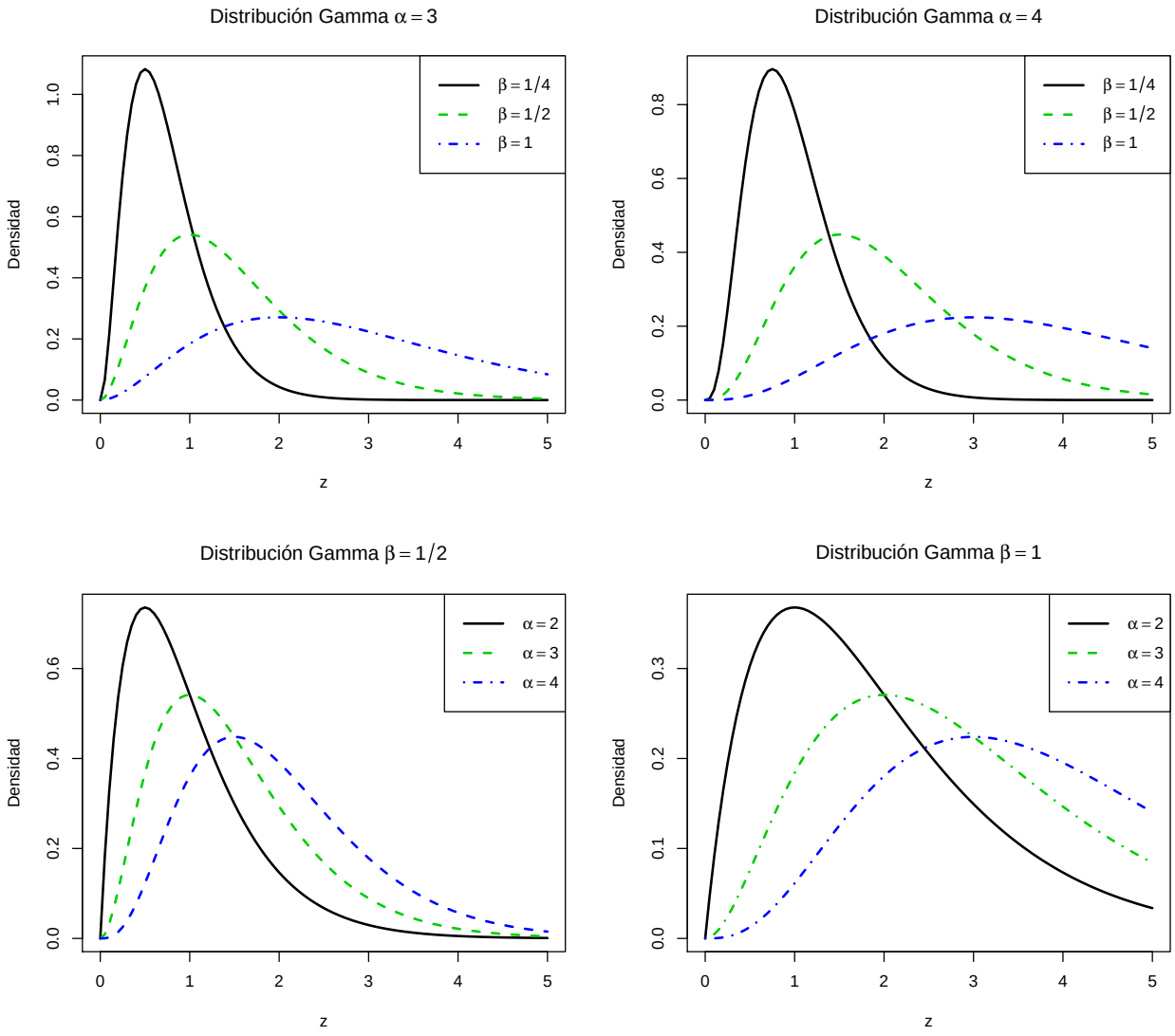

Figura 2.1: Función de densidad de probabilidad de la distribución gamma para diferentes valores de los parámetros α y β.

2.2. Distribución gamma generalizada

La distribución gamma generalizada (GG) extiende la distribución gamma a una de tres parámetros y se aplica, al igual que ella, sobre variables continuas no negativas. Esta es una distribución muy flexible dado que incluye como casos particulares a un número considerable de distribuciones, muchas de las cuales son potencialmente útiles en el modelamiento de variables semicontinuas. Destacan entre ellas, por ejemplo, la distribución de Weibull, log-normal, exponencial, gamma, gamma inversa, entre otras.

Una de sus formulaciones iniciales para una variable aleatoria Z bajo esta distribución fue propuesta por Stacy (1962) y consiste en incluir en la distribución gamma un segundo parámetro de forma $\rho > 0$ de la siguiente manera:

$$f_Z(z;\alpha,\beta,\rho) = \frac{1}{\Gamma(\alpha)}\left(\frac{1}{\beta}\right)^{\alpha\rho} \rho z^{\alpha\rho-1}\exp\left(-\left(\frac{z}{\beta}\right)^{\rho}\right), \quad z \geq 0. \tag{2.1}$$

En Stacy y Mihram (1965) se propone extender el espacio paramétrico de ρ hacia valores negativos, siendo ahora la única restricción que $\rho \neq 0$. En la función de densidad de probabilidad dada en (2.1) se reemplaza el factor $\rho z^{\alpha\rho-1}$ por $|\rho| z^{\alpha\rho-1}$ y se conserva los espacios de los demás parámetros.

$$f_Z(z;\alpha,\beta,\rho) = \frac{1}{\Gamma(\alpha)}\left(\frac{1}{\beta}\right)^{\alpha\rho} |\rho| z^{\alpha\rho-1}\exp\left(-\left(\frac{z}{\beta}\right)^{\rho}\right), \quad z \geq 0. \tag{2.2}$$

Una reparámetrización alternativa es presentada por Manning et al. (2005) que, para estimar un modelo de regresión, establece que α, β y ρ sean expresados en función de nuevos parámetros λ, κ y σ de la siguiente manera:

$$f_Z(z;\lambda,\kappa,\sigma) = \frac{\gamma^{\gamma}}{\sigma z\Gamma(\gamma)\sqrt{\gamma}}\exp(w\sqrt{\gamma}-\lambda), \quad z \geq 0. \tag{2.3}$$

donde $\gamma = |\kappa|^{-2}$, $w = \text{sign}(\kappa)(\ln(z)-\lambda)/\sigma$ y $\lambda = \gamma\exp(|\kappa|w)$. Reemplazando estas notaciones en la expresión anterior se obtienen la equivalencia buscada.

$$\begin{aligned}
f_Z(z;\lambda,\kappa,\sigma) &= \frac{\gamma^{\gamma}}{\sigma z\sqrt{\gamma}\Gamma(\gamma)}\exp\left(\frac{\text{sign}(\kappa)(\ln(z)-\lambda)}{\sigma/\sqrt{\gamma}} - \gamma\exp\left\{\frac{\text{sign}(\kappa)(\ln(z)-\lambda)}{\sigma/\sqrt{\gamma}}\right\}\right) \\
&= \frac{\gamma^{\gamma}}{\sigma z\sqrt{\gamma}\Gamma(\gamma)}\left[\frac{z}{\exp(\lambda)}\right]^{\text{sign}(\kappa)\gamma/\sigma\sqrt{\gamma}} \exp\left[-\gamma\left\{\frac{z}{\exp(\lambda)}\right\}^{\text{sign}(\kappa)1/\sigma\sqrt{\gamma}}\right] \\
&= \frac{(1/\sigma\sqrt{\gamma})[z]^{((\text{sign}(\kappa)1/\sigma\sqrt{\gamma})\gamma-1)}\exp\left(-(z/\exp(\lambda))/\gamma^{\text{sign}(\kappa)\sigma\sqrt{\gamma}}\right)^{\text{sign}(\kappa)1/\sigma\sqrt{\gamma}}}{\Gamma(\gamma)\left(\exp(\lambda)/\gamma^{(\text{sign}(\gamma)\sigma\sqrt{\gamma})}\right)^{((\text{sign}(\kappa)/\sigma\sqrt{\gamma})\gamma)}} \\
&= \frac{|\rho|(z)^{\rho\alpha-1}\exp\left(-(z/\beta)^{\rho}\right)}{\Gamma(\alpha)\beta^{\rho\alpha}}.
\end{aligned}$$

Comparando esta expresión final con la expresión definida en (2.2) los parámetros λ, κ y σ pueden ser expresados como:

$$\alpha = \frac{1}{|\kappa|^2},$$

$$\beta = \frac{\exp(\lambda)}{|\kappa|^{-2\text{sign}(\kappa)\sigma/|\kappa|}},$$

$$\rho = \frac{\text{sign}(\kappa)|\kappa|}{\sigma}.$$

Esto lleva a afirmar que una variable aleatoria Z presenta distribución gamma generalizada de parámetros $-\infty < \lambda < \infty$, $\kappa \neq 0$ y $\sigma > 0$, si su función de densidad es dada por:

$$f_Z(z;\lambda,\kappa,\sigma) = \frac{|\kappa|^{-\frac{2\text{sign}(\kappa)^2}{|\kappa|^2}}}{\sigma z\Gamma(1/|\kappa|^2)|\kappa|^{-1}}\exp\left(\text{sign}(\kappa)\frac{(\ln z-\lambda)}{|\kappa|\sigma} - \frac{1}{|\kappa|^{2\text{sign}(\kappa)^2}}\exp(\text{sign}(\kappa)\frac{(\ln z-\lambda)}{|\kappa|^{-1}\sigma})\right), \; z \geq 0,$$

donde λ es un parámetro de localización, κ un parámetro de forma y σ un parámetro de escala.

El cuadro 2.1 muestra cuáles son los valores que deben asumir los parámetros κ y σ para verificar algunos casos especiales de la distribución gamma generalizada. Como se aprecia, la distribución gamma estándar es un caso particular de la distribución gamma generalizada cuando $\kappa = \sigma$.

Cuadro 2.1: Casos especiales de $GG(\lambda, \kappa, \sigma)$

Distribución	κ	σ	λ
Exponecial	$\kappa = 1$	$\sigma = 1$	$-\infty < \lambda < \infty$
Gamma estándar	$\kappa = \sigma$	$\sigma > 0$	$-\infty < \lambda < \infty$
Log-normal	$\kappa \to 0$	$\sigma > 0$	$-\infty < \lambda < \infty$
Weibull	$\kappa = 1$	$\sigma > 0$	$-\infty < \lambda < \infty$
Gamma inversa	$\kappa = -\sigma$	$\sigma > 0$	$-\infty < \lambda < \infty$

En esta tesis consideraremos para el modelo principal a la gamma generalizada en terminos de la parametrización de Manning et al. (2005) pero restringido al caso donde κ toma valores positivos, $\kappa > 0$. Con esta restricción, la distribución es aplicable a variables estrictamente positivas. A continuación se muestra el detalle de la nueva expresión bajo la restricción mencionada anteriormente. Haciendo uso de las propiedades de la función signo, tenemos que:

$$\begin{aligned}
f_Z(z;\lambda,\kappa,\sigma) &= \frac{1}{\Gamma(1/\kappa^2)}\frac{\kappa^{-2/\kappa^2}}{\exp(\lambda/\sigma\kappa)}\frac{\kappa}{\sigma}z^{\frac{1}{\kappa\sigma}-1}\exp\left(-\frac{1}{\kappa^2}z^{\frac{\kappa}{\sigma}}\exp(-\lambda\kappa/\sigma)\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}z^{\frac{1}{\kappa\sigma}-1}\exp\left(\ln(\kappa^{-2/\kappa^2})\right)\exp(-\lambda/\kappa\sigma)\exp\left(\frac{-1}{\kappa^2}z^{\frac{\kappa}{\sigma}}\exp(-\lambda\kappa/\sigma)\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}z^{\frac{1}{\kappa\sigma}-1}\exp\left(\ln(\kappa^{-2/\kappa^2})-\frac{\lambda}{\kappa\sigma}-\frac{-1}{\kappa^2}z^{\frac{\kappa}{\sigma}}\exp(-\lambda\kappa/\sigma)\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}z^{\frac{1}{\kappa\sigma}-1}\exp\left(\frac{1}{\kappa^2}\ln(\frac{1}{\kappa^2}\exp(-\lambda\kappa/\sigma))-\frac{1}{\kappa^2}z^{\frac{\kappa}{\sigma}}\exp(-\lambda\kappa/\sigma)\right).
\end{aligned}$$

Realizando el cambio de variable por $\eta = \exp(-\frac{\kappa\lambda}{\sigma})$ podemos simplificar la función de densidad de la variable aleatoria Z, que la denotaremos por $Z \sim GG(\lambda, \kappa, \sigma)$, en la siguiente expresión:

$$f_Z(z;\lambda,\kappa,\sigma) = \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}z^{\frac{1}{\kappa\sigma}-1}\exp\left(\frac{-\eta}{\kappa^2}z^{\frac{\kappa}{\sigma}}+\frac{1}{\kappa^2}\ln(\frac{\eta}{\kappa^2})\right), \quad z > 0.$$

La media y la varianza de esta variable vienen dadas respectivamente por:

$$E(Z) = \exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}}\frac{\Gamma(1/\kappa^2+\sigma/\kappa)}{\Gamma(1/\kappa^2)}.$$

y

$$V(Z) = (\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}})^2\left\{\frac{\Gamma(1/\kappa^2+2\sigma/\kappa)}{\Gamma(1/\kappa^2)}-\left(\frac{\Gamma(1/\kappa^2+\sigma/\kappa)}{\Gamma(1/\kappa^2)}\right)^2\right\}.$$

Los otros momentos poblacionales de esta variable están dados por:

$$E(Z^r) = (\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}})^r \frac{\Gamma(1/\kappa^2 + r\sigma/\kappa)}{\Gamma(1/\kappa^2)}.$$

Por lo tanto, los parámetros no negativos κ y σ tienen influencia sobre la asimetría y curtosis de esta distribución.

2.3. Distribución gamma generalizada cero-inflacionada

A lo largo de este trabajo consideraremos una variable aleatoria Y con distribución cero-inflacionada mixta, donde los valores positivos presentan una distribución gamma generalizada cero-inflacionada (GGCI) de parámetros $0 \leq \pi \leq 1$, $-\infty < \lambda < \infty$, $\kappa > 0$ y $\sigma > 0$. La función de densidad de esta variable aleatoria viene dada por:

$$f_Y(y; \pi, \lambda, \kappa, \sigma) = \begin{cases} \pi & \text{, si } y{=}0 \\ (1-\pi) f_Z(y; \lambda, \kappa, \sigma) & \text{, si } y > 0 \end{cases} \tag{2.4}$$

donde π es la probabilidad que Y tome el valor cero y $Z \sim GG(\lambda, \kappa, \sigma)$. A la variable aleatoria Y anterior la denotaremos por $Y \sim GGCI(\pi, \lambda, \kappa, \sigma)$. La media y varianza de esta variable son respectivamente:

$$E(Y) = (1-\pi)\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}} \frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)}.$$

y

$$V(Y) = (1-\pi)(\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}})^2 \left\{ \frac{\Gamma(1/\kappa^2 + 2\sigma/\kappa)}{\Gamma(1/\kappa^2)} - \left(\frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)} \right)^2 (1-\pi) \right\}.$$

Note que $Y \sim GGCI(\pi, \lambda, \kappa, \sigma)$ corresponde a la mixtura entre una variable aleatoria con distribución de Bernoulli de parámetro π y la variable aleatoria con distribición $GG(\lambda, \kappa, \sigma)$.

Con el objetivo de estimar la media total, en vez de la condicional, introduciremos un parámetro γ definido como la media de Y; es decir,

$$\gamma = E(Y) = (1-\pi)\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}} \frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)}. \tag{2.5}$$

Utilizaremos ahora el parámetro λ para insertar el nuevo parámetro γ en la función de densidad de GGCI. A partir de la ecuación (2.5), podemos expresar λ en función de los demás parámetros como:

$$\lambda = \ln(\gamma) - \ln(1-\pi) - \frac{2\sigma}{\kappa}\ln(\kappa) + \ln\Gamma\left(\frac{1}{\kappa^2}\right) - \ln\left[\Gamma\left(\frac{1}{\kappa^2} + \frac{\sigma}{\kappa}\right)\right].$$

Esta expresión de λ es reemplazada luego en la función de densidad de la GGCI. La expresión resultante de la densidad en términos de la nueva parametrizacion $0 \leq \pi \leq 1$, $\gamma \geq 0$, $\kappa > 0$ y $\sigma > 0$ estará dada por:

$$f_Y(y;\pi,\gamma,\kappa,\sigma) = \begin{cases} \pi & , \text{si } y = 0 \\ (1-\pi)\frac{\kappa/\sigma}{\Gamma(1/\kappa^2)} y^{\frac{1}{\kappa\sigma}-1} \exp(-\frac{\eta}{\kappa^2} y^{\frac{\kappa}{\sigma}} + \frac{1}{\kappa^2}\ln(\frac{\eta}{\kappa^2})) & , \text{si } y > 0 \end{cases} \tag{2.6}$$

donde

$$\eta = \exp(-\tfrac{\kappa\lambda}{\sigma})$$

$$\text{y} \qquad \lambda = \ln(\gamma) - \ln(1-\pi) - \tfrac{2\sigma}{\kappa}\ln(\kappa) + \ln\left[\Gamma(\tfrac{1}{\kappa^2})\right] - \ln\left[\Gamma\left(\tfrac{1}{\kappa^2} + \tfrac{\sigma}{\kappa}\right)\right]$$

La expresión (2.6) es entonces la función de densidad de probabilidad de la variable aleatoria Y que sigue una distribución gamma generalizada cero-inflacionada bajo la nueva parametrización, la cual la denotaremos por $Y \sim GGCI(\pi,\gamma,\kappa,\sigma)$.

La figura 2.2 muestra la forma de la función de densidad para distintos valores de κ y σ. Cuando $\sigma > 1$, la función inicia en el infinito y cae rapidamente de forma exponencial, independientemente del valor que toma κ. Para el caso que $\kappa < 1$ y $\sigma < 1$, la función se inicia en un punto cercano a cero.

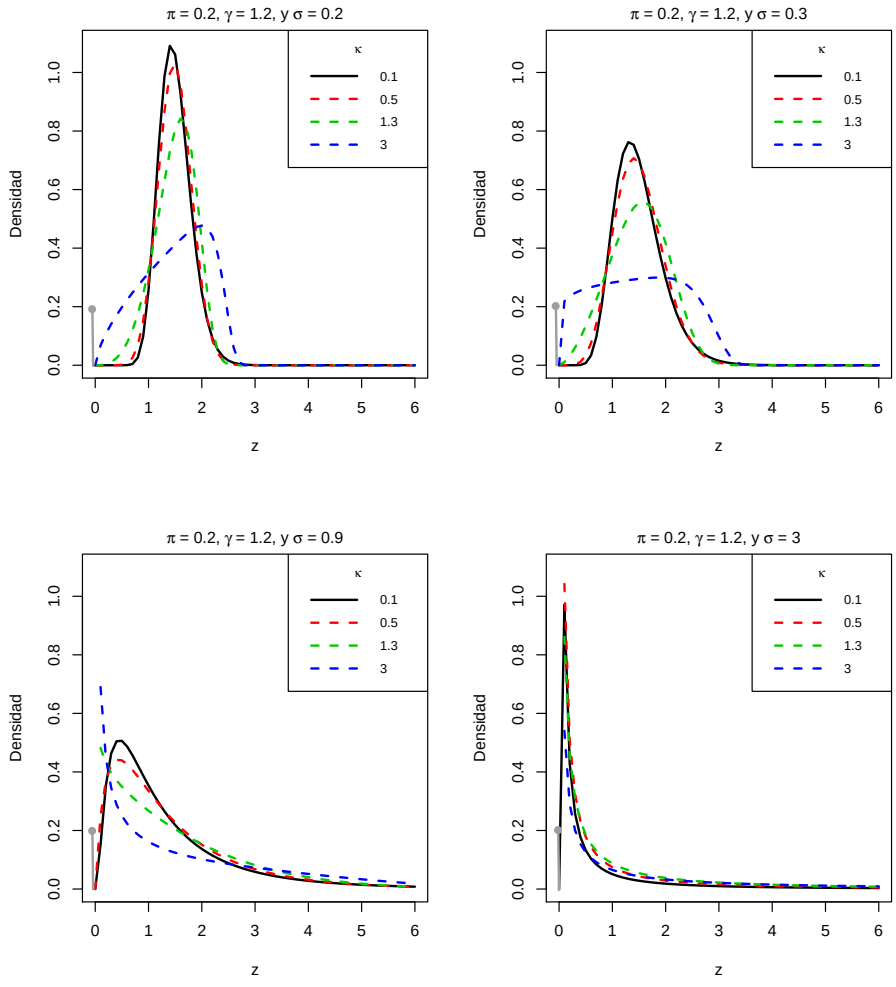

Figura 2.2: Función de densidad de la distribución GGCI (π, γ, κ, σ) para diferentes valores de los parámetros.

2.4. Modelos lineales mixtos

Los modelos lineales mixtos amplían el modelo lineal de manera que estos incluyan efectos aleatorios en su formulación; los cuales son útiles en el modelamiento no solo en datos de corte longitudinal sino también en datos agrupados por conglomerados o con una estrucuta jerárquica.

Según Gumedze y Dunne (2011) los modelos lineales mixtos proporcionan una herramienta poderosa y flexible para el análisis de una amplia variedad de datos que incluyen datos agrupados como datos longitudinales, medidas repetidas, datos bloqueados o multinivel, datos espaciales y datos geoestadísticos. El modelo lineal mixto está dado por:

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

donde $\boldsymbol{Y}$ es un vector de respuestas $n \times 1$, $\boldsymbol{X}$ es una matriz de diseño conocida $n \times p$ para los efectos fijos, $\boldsymbol{\beta}$ es un vector de parámetros de efectos fijos $p \times 1$, $\boldsymbol{Z}$ es una matriz de diseño $n \times q$, $\mathbf{u}$ es un vector de efectos aleatorios $q \times 1$ y $\boldsymbol{\epsilon}$ es un vector de errores aleatorios $n \times 1$, con $E(\mathbf{u}) = \mathbf{0}$ y $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Además, se supone que $\mathbf{u}$ y $\boldsymbol{\epsilon}$ siguen distribuciones normales independientes y multivariadas tales que:

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right).$$

Siendo $\mathbf{G}$ la matriz de varianza covarianza $p \times p$ de los efectos aleatorios y $\mathbf{R}$ la matriz de varianza covarianza de los errores, que usualmente se considera como $\mathbf{R} = \sigma^2 I$, pero sin embargo podría asumir otras estructuras de correlación.

Capítulo 3

Modelos de regresión para respuestas semicontinuas

En el presente capítulo se describe las características de los modelos de regresión que estudiaremos para explicar una variable respuesta semicontinua, los cuales son el modelo de regresión a la media y el de regresión a la media con efectos mixtos. La elección del tipo de modelo dependerá básicamente del tipo de estudio. El modelo de regresión a la media con efectos mixtos es el objetivo de esta tesis, por lo que será desarrollado extensivamente.

3.1. Modelos de regresión a la media para respuestas semicontinuas

Tradicionalmente en la explicación de una variable semicontinua Y se ha venido utilizando el modelo de dos partes propuesto por Duan et al. (1983). La primera parte de este modelo plantea una regresión logística con la finalidad de explicar la presencia o no de valores cero de la respuesta a través de la probabilidad π de que la respuesta tome el valor cero. La segunda parte consiste en una regresión para explicar el nivel medio de la respuesta positiva. Más específicamente en este modelo, la respuesta presenta una distribución semicontinua y separadamente se estiman los efectos de las covariables sobre la respuesta condicionada a valores positivos. Duan et al. (1983) considera por ejemplo una regresión logística para la probabilidad π y un modelo de regresión log-normal para la parte continua. Dado un individuo i este modelo plantea que su probabilidad de que la variable respuesta sea cero y el valor medio de su respuesta positiva satisfacen:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \tilde{\mathbf{x}}_i^{'}\boldsymbol{\alpha},$$

$$\mu_i = E(\ln Y_i | Y_i > 0) = \mathbf{x}_i^{'}\boldsymbol{\gamma}, \tag{3.1}$$

donde $\boldsymbol{\alpha}$ y $\boldsymbol{\gamma}$ son vectores de parámetros de regresión y $\tilde{\mathbf{x}}_i$, $\mathbf{x}_i$ son vectores de covariables para el sujeto i.

Otros autores, sin embargo generalizan (3.1) al incluir funciones de enlace para una regresión sobre $\mu_i = E(Y_i | Y_i > 0)$.

Al ajustar este modelo, la función de verosimilitud resulta ser separable para la componente binaria y continua, y por lo tanto, estas dos partes se ajustan por separado. El componente binario a menudo se modela usando una regresión logística, y el componente

continuo se puede ajustar usando un modelo de regresión estándar, como el log-normal o un modelo de regresión lineal generalizado para variables no negativas.

Smith et al. (2014), por otro lado, desarrollaron el modelo marginalizados de dos partes que en comparación al modelo anterior permite obtener efectos interpretables de las covariables sobre la media global. Para tal propósito se parametriza directamente el efecto de las covariables en terminos de la media global (media marginal), $\gamma_i = E(Y_i)$ sobre los datos en escala original. Este modelo se específica de la siguiente manera:

$$\begin{aligned} \text{logit}(\pi_i) &= \tilde{\mathbf{x}}_i^{'}\boldsymbol{\omega}, \\ E(Y_i) = \gamma_i &= \exp(\mathbf{x}_i^{'}\boldsymbol{\beta}). \end{aligned}$$

Las estimación de parámetros se pueden obtener utilizando rutinas de optimización estándar como Newton-Raphson o por el método de Scoring de Fisher. La media del modelo y los errores estándar también se pueden obtener fácilmente de acuerdo a esta parametrización en un solo proceso de optimización.

Vásquez (2018) presenta un desarrollo del modelo anterior bajo la asunción de que Y sigue una distribución gamma generalizada, al cual llamó modelo de regresión a la media cero-inflacionado (MRCI). En el, como antes, se busca estimar directamente los efectos de las covariables sobre la media total de la respuesta, $\gamma = E(Y) = (1-\pi)\mu$, en vez de la media condicional a valores positivos, $\mu = E(Y|Y > 0)$. Para tal efecto se parametriza el modelo de tal forma que su función de densidad de distribución gamma generalizada (GG) para la parte continua sea expresable en términos de γ.

La formulación del MRCI establece que la función de densidad de Y para un sujeto i, Y_i sea como la dada en la ecuación (2.6), con la diferencia que f_Z se exprese en términos del parámetro γ_i, y no de μ_i. Las ecuaciones de regresión planteadas para π_i y γ_i vienen dadas por:

$$\begin{aligned} g_1(\pi_i) &= \tilde{\mathbf{x}}_i^{'}\boldsymbol{\omega}, \\ g_2(\gamma_i) &= \mathbf{x}_i^{'}\boldsymbol{\beta}, \end{aligned}$$

donde g_1 y g_2 son funciones que enlazan la parte discreta y continua del modelo. Estas funciones son estrictamente monótonas y con segunda derivada continua, como las usualmente consideradas funciones logística y logarítmica. En este último caso el modelo se expresa como:

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \tilde{\mathbf{x}}_i^{'}\boldsymbol{\omega}, \\ \log(\gamma_i) &= \mathbf{x}_i^{'}\boldsymbol{\beta}, \end{aligned}$$

donde $\boldsymbol{\omega} = (\omega_1, ..., \omega_{k_1})^{'}$ y $\boldsymbol{\beta} = (\beta_1, ..., \beta_{k_2})^{'}$ son vectores columna de los parámetros de regresión y $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, ..., \tilde{x}_{ik_1})^{'}$, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ik_2})^{'}$ son vectores columna de covariables para el sujeto i. Vale aclarar que la primera componente de los vectores de covariables podría ser uno para permitir interceptos en el modelo.

3.2. Modelos de regresión a la media con efectos mixtos

Los modelos de regresión anteriormente desarrollados son aplicables a datos de corte tranversal, sin embargo Smith et al. (2017) propuso un modelo longitudinal de dos partes denominado MRMS-LSNCI que directamente busca explicar el efecto de las covariables sobre la media global de Y, extendiendo el modelo marginalizado de dos partes desarrollado por Smith et al. (2014) para datos semicontinuos transversales. Este modelo asume que la variable respuesta Y_{ij} del i-ésimo sujeto para la j-ésima medición sigue una distribución cero-inflacionada similar a la dada en (2.4); pero plantea una distribución Log Skew Normal (LSN) como componente continua, en lugar de la distribución gamma generalizada aquí planteada. El modelo se escribe como:

$$\begin{aligned} g_1(\pi_{ij}) &= \tilde{\mathbf{x}}_{ij}^{'}\boldsymbol{\alpha} + \tilde{\mathbf{z}}_{ij}^{'}\mathbf{a}_i, \\ g_2(\gamma_{ij}) &= \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{d}_i, \end{aligned} \tag{3.2}$$

donde las funciones de enlace g_1 y g_2 tienen las características dadas anteriormente, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)^{'}$ y $\boldsymbol{\beta} = (\beta_1, ..., \beta_l)^{'}$ son vectores de coeficientes de regresión asociados a $\pi_{ij} = P(Y_{ij} = 0)$ y $\gamma_{ij} = E(Y_{ij})$, respectivamente; $\mathbf{a}_i = (a_{i1}, ..., a_{ip})^{'}$ y $\mathbf{d}_i = (d_{i1}, ..., d_{ir})^{'}$ son los efectos aleatorios asociados a π_{ij} y γ_{ij}, respectivamente y $\tilde{\mathbf{x}}_{ij} = (\tilde{x}_{ij1}, ..., \tilde{x}_{ijk})^{'}$, $\tilde{\mathbf{z}}_{ij} = (\tilde{z}_{ij1}, ..., \tilde{z}_{ijp})^{'}$, $\mathbf{x}_{ij} = (x_{ij1}, ..., x_{ijl})^{'}$, $\mathbf{z}_{ij} = (z_{ij1}, ..., z_{ijr})^{'}$ son vectores de covariables. Además asumimos que el vector aleatorio $\mathbf{b}_i^{'} = (\mathbf{a}_i^{'}, \mathbf{d}_i^{'})$, de efectos aleatorios, siguen conjuntamente una distribución normal multivariada.

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{a}_i \\ \mathbf{d}_i \end{pmatrix} \sim \mathbf{N}\left(\mathbf{0}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_{ad} \\ \boldsymbol{\Sigma}_{ad} & \boldsymbol{\Sigma}_d \end{bmatrix}\right), \tag{3.3}$$

siendo $\boldsymbol{\Sigma}_a$, $\boldsymbol{\Sigma}_d$ y $\boldsymbol{\Sigma}_{ad}$ matrices de varianza-covarianza para los efectos aleatorios.

Cabe resaltar que el modelo dado en (3.2) deja de tener dos partes y en adelante será adecuado nombrarlo como regresión de una parte, dado que el proceso de optimización para estimar sus parámetros se realiza en una sola etapa. En adelante denominaremos a (3.2) como un modelo regresión a la media con efectos mixtos para respuestas semicontinuas (MRMS-GGCI).

La formulación general de nuestro modelo MRMS-GGCI la completaremos considerando que la función de densidad de probabilidad de Y_{ij} sea como la dada en la ecuación (2.6); es decir, sigue una distribución gamma generalizada cero-inflacionada, lo cual recordemos estamos denotando por $Y_{ij} \sim GGCI(\pi_{ij}, \gamma_{ij}, \kappa, \sigma)$.

Cualquier función de distribución acumulada de una variable continua puede ser una función de enlace inversa apropiada. Entre ellos tenemos la función de enlace probit que tiene como desventaja aumentar la dificultad en la interpretación de los efectos sobre la variable dependiente. Para facilitar la interpretación, vamos a usar, como funciones de enlace, el inverso de la función de distribución logística acumulada para la probabilidad en cero y la función logarítmica para la media. Las ecuaciones de regresión planteadas en (3.2) tomarán la forma:

$$\begin{aligned} \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \tilde{\mathbf{x}}_{ij}^{'}\boldsymbol{\alpha} + \tilde{\mathbf{z}}_{ij}^{'}\mathbf{a}_i, \\ \log(\gamma_{ij}) &= \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{d}_i, \end{aligned}$$

donde la última ecuación puede también escribirse como $\gamma_{ij} = \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{d}_i)$

Definiendo $\mathbf{b} = (b_1, ..., b_n)^{'}$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, vec(\boldsymbol{\Sigma}), \kappa, \sigma)$ e $Y = (Y_1, ..., Y_n)^{'}$, donde $Y_i = (Y_{i1}, ..., Y_{in_i})^{'}$, para $i = 1, 2, ..., n$, la función de verosimilitud aumentada para nuestro modelo MRMS podrá escribirse como:

$$\begin{aligned} L(\boldsymbol{\theta}, \mathbf{b}|Y) &= p(Y, \mathbf{b}|\boldsymbol{\theta}) = p(Y|\mathbf{b}, \boldsymbol{\theta})p(\mathbf{b}|\boldsymbol{\theta}), \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\pi_i, \gamma_i, \kappa, \sigma) \times \varphi(b_i|\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{3.4}$$

donde $\varphi(.|\mathbf{0}, \boldsymbol{\Sigma})$ denota a la función de densidad de la distribución normal multivariada con vector de media $\mathbf{0}$ y matriz de varianza y covarianza $\boldsymbol{\Sigma}$ dada en (3.3), $\boldsymbol{\pi_i} = (\pi_{i1}, ..., \pi_{in_i})^{'}$ y $\boldsymbol{\gamma_i} = (\gamma_{i1}, ..., \gamma_{in_i})^{'}$ son vectores de parámetros de tamaños n_i y $f_{Y_i}(y_i|\pi_i, \gamma_i, \kappa, \sigma)$ es la distribución de probabilidad conjunta del vector $Y_i = [Y_{i1}, ..., Y_{in_i}]^{'}$, donde asumiendo independencia condicional para los efectos aleatorios, esta se expresa como:

$$f_{Y_i}(y_i|\boldsymbol{\theta}, b_i) = \prod_{j=1}^{n_i} f_{Y_{ij}}(y_{ij}|\pi_{ij}, \gamma_{ij}, \kappa, \sigma),$$

siendo $f_{Y_{ij}}(y_{ij}|\pi_{ij}, \gamma_{ij}, \kappa, \sigma)$ la función de densidad GGCI con parámetros π_{ij}, γ_{ij}, κ, y σ definida en (2.6).

3.3. Inferencia Bayesiana

La estimación de máxima verosimilitud (MV) puede presentar desafíos computacionales cuando se incluyen efectos aleatorios en el modelo. Muchos autores sugieren que la estimación por MV en estos casos puede aún realizarse usándose métodos aproximados como la cuadratura Gaussiana para marginalizar las componentes aleatorias; sin embargo, a menudo estos no logran converger o incluso cuando convergen, los tiempos de ejecución pueden demandar mucho tiempo. Por lo tanto adoptaremos para la estimación de los parámetros de

nuestro modelo un enfoque bayesiano lo cual nos permitirá simplificar la parte computacional y trabajar incluso con modelos de efectos aleatorios multidimensionales correlacionados. A continuación se detalla el procedimiento para la estimación bayesiana.

Tomando en cuenta la función de verosimilitud aumentada descrita en (3.4), la distribución a posteriori aumentada de $\boldsymbol{\theta}$ y $\mathbf{b}$, denotada por $p(\boldsymbol{\theta}, \mathbf{b}|Y)$, viene dada por el teorema de Bayes, por:

$$p(\boldsymbol{\theta}, \mathbf{b}|Y) \propto p(Y|\boldsymbol{\theta}, \mathbf{b}) \times p(\mathbf{b}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \mathbf{b}|Y)p(\boldsymbol{\theta}), \tag{3.5}$$

donde $p(\boldsymbol{\theta})$ es una distribución a priori para $\boldsymbol{\theta}$. En esta tesis consideraremos que $\mathbf{b}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, κ y σ son a priori elementos aleatorios independientes, por lo que la distribución a priori podrá escribirse como:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\mathbf{b})p(\boldsymbol{\Sigma})p(\kappa)p(\sigma). \tag{3.6}$$

Para los efectos fijos proponemos la distribución normal multivariada con $\boldsymbol{\alpha} \sim N_k(\mathbf{0}, \mathbf{A})$ y $\boldsymbol{\beta} \sim N_k(\mathbf{0}, \mathbf{B})$. Para las matrices de covarianza proponemos como distribución a priori a la distribución Inversa de Wishart (IW) tales que $\boldsymbol{\Sigma} \sim IW(\psi, \Psi)$. Para los parámetros de escala y forma, κ y σ se consideran como prioris a la distribución Gamma-Inversa, $\kappa \sim GI(a^*, b^*)$ y la distribución uniforme $\sigma \sim U(c^*, d^*)$. Para estas distribuciones $\mathbf{A}$, $\mathbf{B}$, a^*, b^*, c^* y d^* son hiperparámetros especificados. Estas específicaciones fueron consideradas según lo propuesto por Smith et al. (2017).

Combinando la función de verosimilitud aumentada definida en (3.4) con la distribución a priori definida en (3.6) la distribución a posteriori aumentada definida en (3.5) puede ser escrita de manera más explícita como:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{b}|Y) \propto & \prod_{i=1}^{n} \left[\prod_{j=1}^{n_i} f_{Y_{ij}}(y_{ij}|\pi_{ij}, \gamma_{ij}, \kappa, \sigma) \right] \\ & \times \varphi_p(b_i|\mathbf{0}, \boldsymbol{\Sigma}) \times \varphi_k(\boldsymbol{\alpha}|\mathbf{0}, \mathbf{A}) \times \varphi_L(\boldsymbol{\beta}|\mathbf{0}, \mathbf{B}) \\ & \times h(\boldsymbol{\Sigma}|\psi, \Psi) \times q_1(\kappa|a^*, b^*) \times q_2(\kappa|c^*, d^*), \end{aligned} \tag{3.7}$$

donde $\varphi(\cdot|\mathbf{0}, \boldsymbol{\Sigma})$ denota la función de densidad de probabilidad de una distribución normal multivariada con vector de media $\mathbf{0}$ y matriz de covarianza $\boldsymbol{\Sigma}$, $q_1(\cdot|a^*, b^*)$ denota la distribución gamma inversa, $q_2(\cdot|c^*, d^*)$ denota la distribución uniforme, y $h(\cdot|\psi, \Psi)$ denota la función de densidad de probabilidad de la distribución inversa de Wishart (IW), donde ψ es interpretado como un grado de libertad y Ψ es una matriz de escala $p \times p$.

3.4. Selección de modelos

Existen en la literatura una gran cantidad de criterios de información para evaluar el ajuste de diferentes modelos bajo el enfoque bayesiano. Criterios de información como el de

desvío (DIC) propuesto por Spiegelhalter et al. (2002), Información Ampliamente Aplicable (WAIC), propuesto por Watanabe (2010) y la validación cruzada de Leave-one-out (LOO-CV) propuesto por Vehtari et al. (2017).

El criterio de información WAIC puede verse como una mejora del criterio DIC para modelos bayesianos. El criterio de información DIC debe su popularidad gran parte a su directa implementación en el paquete BUGS, pero se sabe que tiene algunos problemas debido a que no es completamente bayesiano, ya que se basa en una estimación puntual. Por otra parte el WAIC es completamente bayesiano y se aproxima mucho a la validación cruzada bayesiana. A diferencia del DIC, el citerio WAIC es invariante a la parametrización y también funciona para modelos singulares.

Para esta tesis utilizaremos, aparte del WAIC, el método de validación cruzada de Leave-one-out (LOO-CV) en base a un muestreo de importancia suavizado de Pareto (PSIS), el cual es un nuevo procedimiento para regularizar las ponderaciones de importancia. Aunque el WAIC es asintóticamente igual al LOO-CV, se demuestra que PSIS-LOO es más robusto en el caso finito ante la presencia de observaciones influyentes.

Actualmente la validación cruzada de Leave-one-out (LOO-CV) y el criterio de información de Watanabe-Akaike (WAIC) son métodos usados para estimar la precisión de la predicción puntual fuera de la muestra a partir de un modelo bayesiano ajustado, en base a la log-verosimilitud evaluada en las simulaciones posteriores de los valores de los parámetros. El LOO-CV y WAIC tienen varias ventajas sobre estimaciones más simples de el error predictivo, como el DIC, pero se usan menos en la práctica porque su cálculo requiere pasos computacionales adicionales.

A continuación mostramos el desarrollo de criterios de selección en discusión.

Sea $\boldsymbol{\nu} = [\boldsymbol{\theta}, \mathbf{b}]$ el vector de parámetro que contiene todos los parámetros de la distribución a posteriori aumentada (3.7). Definimos el desvío como sigue:

$$\mathcal{D}(\boldsymbol{\nu}) = -2\log(L(\boldsymbol{\nu}|Y)) = -2\log(L(\boldsymbol{\theta}, \mathbf{b}|Y)),$$

donde $L(\boldsymbol{\theta}, \mathbf{b}|Y)$ es la función de verosimilitud aumentada definida en (3.4). El criterio DIC se define en base al desvío anterior como:

$$DIC = \mathcal{D}(\overline{\boldsymbol{\nu}}) + 2 \times p_D,$$

donde $p_D = \overline{\mathcal{D}}(\nu) - \mathcal{D}(\overline{\nu})$ puede ser interpretado como el número efectivo de parámetros. Considerando una simulación de Montecarlo de tamaño M, es decir, $\nu_1, \nu_2, ..., \nu_M$ elegida de la distribución a posteriori aumentada definida en (3.7), los terminos $\overline{\mathcal{D}}(\nu)$ y $\mathcal{D}(\overline{\nu})$ son calculados como:

$$\overline{\mathcal{D}}(\boldsymbol{\nu}) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{D}(\nu_m) \quad \text{y} \quad \mathcal{D}(\overline{\boldsymbol{\nu}}) = \mathcal{D}\left(\frac{1}{M}\sum_{m=1}^{M}\nu_m\right).$$

Por otro lado, el criterio WAIC es calculado de manera similar al DIC, difiriendo sólo en el término de los parámetros efectivos. Watanabe (2010) define el WAIC como:

$$\text{WAIC} = \mathcal{D}(\overline{\nu}) + 2 \times p_{WAIC}$$

donde $p_{WAIC} = \sum_{i=1}^{n} Var(\log(p(Y_i|\boldsymbol{\nu})))$.

En cuanto al criterio Leave-one-out de validación cruzada (LOO-CV), los datos se dividen repetidamente en un conjunto de entrenamiento y prueba, y el modelo se ajusta para los datos de entrenamiento obteniendo una distribución a posteriori $p_{\text{train}}(\boldsymbol{\nu}) = p\left(\boldsymbol{\nu}|y_{\text{train}}\right)$. Con este ajuste evaluado se obtiene una estimación de la densidad predictiva logarítmica de los datos de prueba, $\log p_{\text{train}}\left(y_{\text{prueba}}\right) = \log \int p_{\text{pred}}\left(y_{\text{prueba}}|\boldsymbol{\nu}\right) p_{\text{train}}(\boldsymbol{\nu}) d\boldsymbol{\nu}$.

Suponiendo que la distribución a posteriori $p\left(\boldsymbol{\nu}|y_{\text{train}}\right)$ es resumido por S simulaciones $\boldsymbol{\nu}^s$, calculamos el logaritmo de la densidad predictiva como $\log\left(\frac{1}{S}\sum_{s=1}^{S} p\left(y_{\text{prueba}}|\boldsymbol{\nu}^s\right)\right)$.

Para simplificar, restringiremos nuestra atención aquí a la validación cruzada de omisión, un caso especial con n particiones en el que cada conjunto de prueba representa un solo dato puntual. Realizar el análisis para cada uno de los n puntos de datos (o quizás un subconjunto aleatorio para cómputo eficiente si n es grande) produce n diferentes inferencias $p_{\text{post}}\left(-i\right)$, cada una resumida por S simulaciones a posteriori $\boldsymbol{\nu}^{is}$.

La estimación bayesiana de LOO-CV del ajuste predictivo fuera de la muestra es

$$\text{lppd}_{\text{loo-cv}} = \sum_{i=1}^{n} \log p_{\text{post}(-i)}\left(y_i\right), \text{calculado como} \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\left(y_i|\boldsymbol{\nu}^{is}\right)\right).$$

Para la obtención de los criterios mencionados se hará uso de la librería **loo** del programa **R** el cual fue implementado por Vehtari et al. (2017). Muy aparte de realizar la validación cruzada, el paquete también proporciona errores predictivos y otras técnicas de ponderación de modelos.

Un valor inferior del WAIC y el LOO indica un mejor ajuste, y por tanto el modelo con el valor mas bajo de estos indicadores podría considerarse como el mejor modelo.

3.5. Métodos de cadenas de Markov de Monte Carlo Hamiltonianos

Los modelos bayesianos jerárquicos son pilares de la comunidad de aprendizaje automático y estadístico. Sin embargo, la inferencia a posteriori exacta en tales modelos es raramente tratable, por lo que los investigadores y profesionales generalmente deben recurrir a métodos aproximados de inferencia estadística. Los algoritmos aproximados de inferencia determinísti-

ca pueden ser eficientes, pero introducen sesgos y pueden ser difíciles de aplicar a algunos modelos. En lugar de calcular una aproximación determinista para una distribución objetivo, los métodos de cadenas de Markov Monte Carlo (MCMC) ofrecen esquemas para generar una serie de muestras correlacionadas que convergerán en distribución a la distribución objetivo Neal (1993).

No todos los algoritmos de MCMC se implementan de manera similar. Para modelos complicados con muchos parámetros, métodos simples como el de Metrópolis de camino aleatorio y el muestreador de Gibbs pueden requerir un tiempo inaceptablemente largo para converger a la distribución objetivo. Esto se debe en gran parte a la tendencia de estos métodos a explorar el espacio de parámetros a través de caminatas aleatorias ineficientes Neal (1993). Cuando los parámetros del modelo son continuos en lugar de discretos, el método de Monte Carlo Hamiltoniano (HMC), también conocido como Monte Carlo híbrido, es capaz de suprimir dicho comportamiento de paseo aleatorio por medio de un inteligente esquema de variables auxiliares que transforma el problema del muestreo de una distribución objetivo en un problema de la simulación de la dinámica hamiltoniana Neal et al. (2011). Fue precisamente Duane et al. (1987), quienes sugirieron el uso de sistemas dinámicos Hamiltonianos, dando lugar al algoritmo Híbrido o Hamiltoniano de Monte Carlo (HMC). La dinámica Hamiltoniana se usa para describir cómo se mueven los objetos en un sistema y se define en terminos de ubicación del objeto y su impulso (equivalente a la velocidad de masa del objeto) es un momento específico. Para cada ubicación del objeto hay una energía potencial asociada $U(\mathbf{x})$ y una energía cinética $K(\mathbf{p})$. La energía total del sistema es constante y se denomina $H(\mathbf{x}, \mathbf{p})$, definido como la suma de la energía potencial y la energía cinética:

$$H(\mathbf{x}, \mathbf{p}) = U(\mathbf{x}) + K(\mathbf{p}),$$

y la evolución del espacio de estados a lo largo del tiempo t obedece a las ecuaciones Hamiltonianas:

$$\begin{cases} \dot{\mathbf{x}} &= \frac{\partial H}{\partial \mathbf{p}} = \nabla_{\mathbf{p}} K(\mathbf{p}), \\ \dot{\mathbf{p}} &= -\frac{\partial H}{\partial \mathbf{x}} = -\nabla_{\mathbf{x}} U(\mathbf{x}), \end{cases} \tag{3.8}$$

donde el punto denota a la derivada con respecto al tiempo y ∇ al operador gradiente. En otras palabras, si resolvemos estas ecuaciones sabremos en cada instante cuál es la posición y el momento o velocidad del objeto.

Las ecuaciones de Hamiltonianas en (3.8), que raramente tienen soluciones explícitas, describen el movimiento de un objeto con respecto al tiempo, el cual es una variable continua. Para resolverlas o simularlas en una computadora, las ecuaciones de Hamiltonianas deben aproximarse numéricamente por métodos de discretización del tiempo. Esto se hace dividiendo el intervalo de tiempo en pequeños de intervalos o pasos de longitud ε. La forma más conocida de aproximar la solución del sistema (3.8) es el método de Leapfrog.

El método de las cadenas de Markov de Monte Carlo Hamiltonianos (HMC) se sintetiza

en un algoritmo de cadenas de Markov de Monte Carlo (MCMC) que adopta la dinámica de un sistema físico en lugar de una distribución de probabilidades para proponer futuros estados en la cadena. Esto permite a la cadena de Markov explorar la distribución objetivo de forma mucho más eficiente, lo que resulta en una convergencia más rápida. En otras palabras evita el comportamiento de la caminata aleatoria y la sensibilidad a los parámetros correlacionados que son comunes en los distintos métodos de MCMC. Estas características le permiten converger rápidamente a comparación de los métodos más simples como el Metropolis o el muestreador de Gibbs. Sin embargo, el rendimiento de HMC es muy sensible a dos parámetros especificados por el usuario: el tamaño de paso ϵ y el número deseado de pasos L.

- Un ϵ muy pequeño ciertamente incrementará el número de cálculos y uno muy grande hará que el esquema Leapfrog sea muy inexacto produciéndose por tanto bajas tasas de aceptación.
- Si L es pequeño las sucesivas muestras serán muy cercanas unas de otras resultando en el indeseable comportamiento de un camino aleatorio que buscábamos justamente evitar. Si L es muy grande el algoritmo HMC generará trayectorias que pudiesen retornar y retrasar sus pasos pudiendo incluso obtenerse una cadena que no converge. Neal et al. (2011)

Hoffman y Gelman (2014) desarrollaron el algoritmo No-U-Turn Sampler (NUTS), el cual es un algoritmo MCMC que se asemeja mucho al HMC, pero elimina la necesidad de elegir el parámetro de número de pasos L. NUTS usa un algoritmo recursivo para construir un conjunto de probables puntos candidatos que abarca una amplia franja de la distribución objetivo, deteniéndose automáticamente cuando el algoritmo de Leapfrog comienza a retroceder, lo cual es una condición suficiente para que la cadena converja.

La principal ventaja de NUTS frente a otros algoritmos HMC es que NUTS es el algoritmo con más probabilidad de producir muestras aproximadamente independientes, en el sentido que generan simulaciones con baja autocorrelación.

3.6. Implementación computacional

Para la implementación del modelo de regresión a la media con efectos mixtos para respuestas semicontinuas (MRMS) usaremos el paquete *rstan* del programa **R**. El proyecto Stan desarrolla un lenguaje de programación probabilística que implementa la inferencia estadística bayesiana a través de Cadenas de Markov de Monte Carlo Hamiltonianos. Este paquete implementa el algoritmo No-U-Turn (NUTS) en el método de Monte Carlo Hamiltoniano (HMC). El paquete *rstan* le permite a uno ajustar convenientemente los modelos y acceder a las salidas que incluyen las inferencias a posteriori, las evaluaciones de la densidad a posteriori y sus gradientes.

Capítulo 4

Estudio de Simulación

En esta sección se realizará un estudio de simulación para el modelo regresión a la media con efectos mixtos para respuestas semicontinuas (MRMS-GGCI) para diferentes tamaños de muestra. Lo que se busca es comprobar si el modelo propuesto es capaz de recuperar los parámetros teóricos. A continuación se detalla los principales objetivos del estudio de simulación.

4.1. Objetivos

- Simular un conjunto de datos semicontinuos de tipo longitudinal a partir de la definición del modelo MRMS-GGCI.
- Evaluar la precisión en la recuperación de los parámetros del modelo MRMS-GGCI mediante el muestreador No-U-Turn (NUTS) usando los criterios como el sesgo, sesgo relativo, RMSE (Raíz del Error Cuadrático Medio) y la cobertura al 95 % de confianza.

4.2. Algoritmo para simular los datos

Para simular un conjunto de datos de tipo longitudinal para el modelo MRMS-GGCI, se procedió de la siguiente manera:

- Consideramos diferentes escenarios para el tamaño de muestra (n=100, 300, 600 y 1000) y 5 mediciones por cada sujeto. Para fines prácticos se consideró una sola covariable x_{ij}, donde $x_{ij} = N(0.5, 1) + j \times U(0, 0.1)$. Con esta definición se busca incorporar una tendencia en los datos y garantizar una estructura de dependencia entre un mismo individuo.
- Para los efectos aleatorios, consideramos $a_i \sim N(0, \sigma_a^2)$ y $d_i \sim N(0, \sigma_d^2)$. Se muestrearon los efectos aleatorios, a_i y d_i de una distribución normal con media cero y varianzas $\sigma_a^2 = \sigma_d^2 = 1.0$.
- Para generar las respuestas semicontinuas, se incorpora covariables a los parámetros π y γ de nuestro modelo MRMS. Establecemos los coeficientes de efectos fijos como $\boldsymbol{\beta} = (\beta_0, \beta_1)' = (-0.5, 1.1)'$ y $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)' = (-1, 0.25)'$.
- Para el caso de la distribución gamma generalizada cero-inflacionada, se generó tantas observaciones (respuestas) según el número de individuos y el número de mediciones por individuo con la siguiente especificación: $Y_{ij} \sim GGCI(\pi_{ij}, \gamma_{ij}, \kappa, \sigma)$, donde $g_1(\pi_{ij}) =$

$\tilde{\mathbf{x}}_{ij}^{'}\boldsymbol{\alpha} + \tilde{\mathbf{z}}_{ij}^{'}\mathbf{a}_i$ y $g_2(\gamma_{ij}) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{d}_i$, tales que g_1 y g_2 son funciones de enlace de tipo logística inversa (logit) y logarítmica, respectivamente. Los parámetros κ y σ se fijarón en 0.10 y 0.30, respectivamente.

Por fines prácticos las ecuaciones de regresión planteadas anteriormente tomaron la forma:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_0 + \alpha_1 x_{ij} + a_i,$$
$$\log(\gamma_{ij}) = \beta_0 + \beta_1 x_{ij} + d_i,$$

Según la figura 4.1 se observa que la variable Y_{ij} presenta una distribución continua no negativa. Recordar que un caso particular de esta distribución es que cuando $\kappa = \sigma$ la distribución resultante es la distribución gamma estándar, por tal motivo se puede usar esta distribución para la parte continua como un modelo alternativo para fines de comparación.

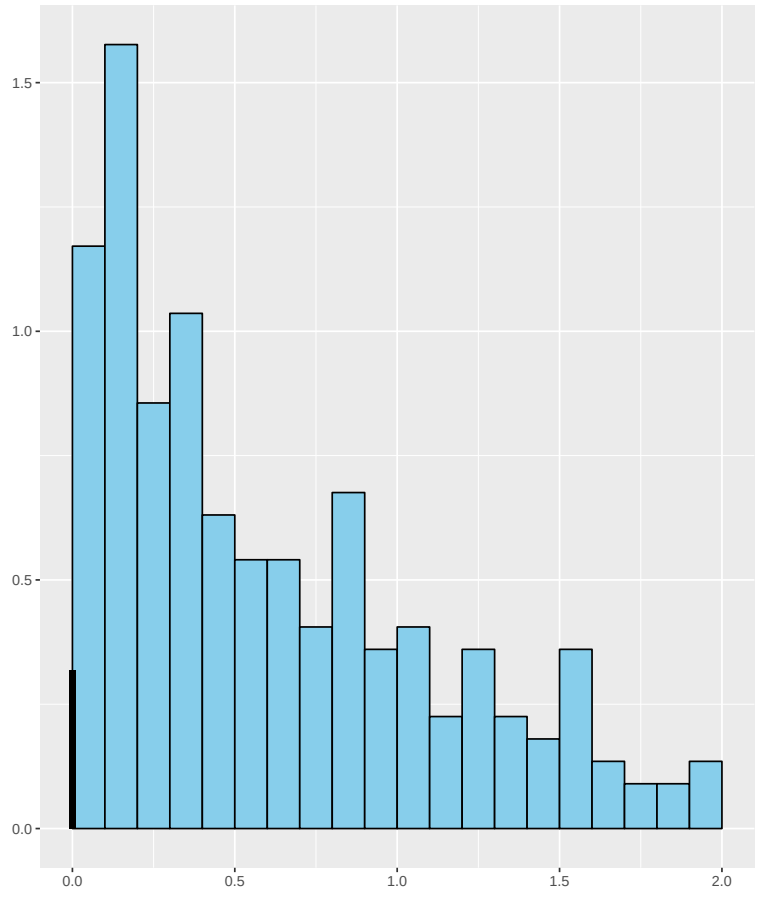

Figura 4.1: Simulación de un conjunto de datos con distribución $GGCI(\pi_{ij}, \gamma_{ij}, \kappa = 0.10, \sigma = 0.30)$.

4.3. Método para estimar los parámetros

Para la estimación de los parámetros usaremos los métodos MCMC Hamiltonianos, los cuales se encuentran implementados en el paquete *rstan* del **R**. Este paquete implementa un algoritmo adaptativo del método de Monte Carlo Hamiltonianos (también conocido como HMC) utilizando el muestreador No-U-Turn (NUTS).

4.4. Criterios para evaluar la precisión de la simulación

Para medir qué tan bien los métodos de estimación logran recuperar los parámetros del modelo, utilizaremos los siguientes indicadores: el sesgo, el sesgo relativo, la raíz del error cuadrático medio (RMSE) y la cobertura al 95 % del intervalo de credibilidad. A continuación se presentan estos indicadores:

$$\text{Sesgo}\left(\hat{\theta}_j\right) = \frac{\sum_{i=1}^{m} \hat{\theta}_j^{(i)}}{m} - \theta_j$$

$$\text{Sesgo\_Relativo}\left(\hat{\theta}_j\right) = \frac{\text{Sesgo}\left(\hat{\theta}_j\right)}{\theta_j} \times 100\,\%$$

$$\text{RMSE}\left(\hat{\theta}_j\right) = \sqrt{\frac{\sum_{i=1}^{m}\left(\hat{\theta}_j^{(i)} - \theta_j\right)^2}{m}}$$

$$\text{Cobertura}\left(\hat{\theta}_j\right) = \frac{\sum_{i=1}^{m} \mathbf{I}\left[\theta_j \in \text{IC}\left(\hat{\theta}_j^{(i)}\right) \text{de}\, 95\,\%\right]}{m} \times 100\,\%$$

donde $\hat{\theta}_j^{(i)}$ es la estimación de los parámetros θ_j en la i-ésima simulación y m representa el número de simulaciones.

4.5. Resultados

El cuadro 4.1 muestra los resultados para estudio de simulación con las especificaciones dadas anteriormente. El cuadro mencionado muestra el sesgo, sesgo relativo, la raíz del error cuadrático medio y la cobertura del estimador. Para los resultados mostrados se consideró 300 réplicas, en cada réplica se estimó el modelo considerando 10 mil iteraciones y se eliminaron las primeras 2 mil iteraciones como periódo de burn-in.

Con relación al sesgo, los estimadores de los parámetros de regresión presentan sesgos relativamente altos para el escenario cuando n =100, sin embargo conforme aumenta el tamaño de muestra, estos sesgos comienzan a disminuir. De manera similar se puede observar en el sesgo relativo.

Respecto al RMSE se observa que los estimadores presentan valores aceptables, en general son relativamente bajos, tanto para la parte discreta como para la parte continua. Por otro

lado, es evidente que conforme aumentemos el tamaño de la muestra el RMSE de la mayoria de los estimadores tienden a disminuir.

La cobertura de los parámetros de regresión en todos los escenarios presentan un valor cercano al 95 %. Esta mejora a medida que el tamaño de muestra se incrementa.

Cuadro 4.1: Estudio de simulación del modelo MRMS-GGCI considerando diferentes tamaños de muestra.

Tamaño de muestra	**Parámetro**	**Valor verdadero**	**Sesgo**	**Sesgo relativo**	**RMSE**	**Cobertura al 95 % (%)**
$n = 100$	α_0	-1.00	-0.0144	1.4400	0.1660	96.6
	α_1	0.25	0.0030	1.2000	0.0260	92.7
	β_0	-0.50	-0.0158	3.1600	0.1072	94.0
	β_1	1.10	-0.0002	-0.0182	0.0081	94.5
	σ_a	1.00	0.0060	0.6000	0.1301	96.7
	σ_d	1.00	0.0003	0.0300	0.0751	94.6
	κ	0.10	0.0102	10.2322	0.1576	96.0
	σ	0.30	0.0278	9.2773	0.1703	92.6
$n = 300$	α_0	-1.00	-0.0091	0.9120	0.1124	96.1
	α_1	0.25	0.0018	0.7200	0.0260	92.4
	β_0	-0.50	-0.0121	2.4200	0.0946	94.6
	β_1	1.10	-0.0001	-0.0127	0.0042	94.7
	σ_a	1.00	0.0036	0.3600	0.1091	96.2
	σ_d	1.00	0.0002	0.0160	0.0389	94.8
	κ	0.10	0.0098	9.8000	0.1210	95.8
	σ	0.30	0.0260	8.6700	0.1692	94.0
$n = 600$	α_0	-1.00	-0.0052	0.5210	0.1019	95.7
	α_1	0.25	0.0011	0.4360	0.0180	94.6
	β_0	-0.50	-0.0103	2.0600	0.0543	94.8
	β_1	1.10	-0.0001	-0.0100	0.0028	94.5
	σ_a	1.00	0.0021	0.2100	0.0958	96.0
	σ_d	1.00	0.0001	0.0109	0.0267	94.4
	κ	0.10	0.0071	7.1000	0.1009	95.6
	σ	0.30	0.0193	6.4167	0.1270	94.2
$n = 1000$	α_0	-1.00	-0.0032	0.3192	0.0868	95.0
	α_1	0.25	0.0007	0.2618	0.0167	94.9
	β_0	-0.50	-0.0052	1.0420	0.0454	95.0
	β_1	1.10	-0.0001	-0.0045	0.0012	94.8
	σ_a	1.00	0.0013	0.1300	0.0717	95.1
	σ_d	1.00	0.0001	0.0063	0.0147	94.9
	κ	0.10	0.0030	3.0147	0.0865	95.2
	σ	0.30	0.0081	2.7071	0.1155	94.6

Capítulo 5

Aplicación

En esta sección se desarrollará la aplicación del modelo regresión a la media con efectos mixtos para respuestas semicontinuas (MRMS-GGCI). Nuestro interés radicará en modelar la variable: Gasto de los hogares destinados a la educación, para ello se usará la base de datos de la Encuesta Nacional de Hogares (ENAHO) con datos de tipo panel para el periódo 2013-2017.

El motivo por el cuál se eligió esta base de datos, se debe basicamente, a que como la ENAHO lo referencia: "La ENAHO constituye la fuente más importante y oportuna que dispone el país para la obtención de información estadística de carácter social, demográfica y económica de los hogares. Además permite medir el nivel de vida de la población, el análisis de políticas en el área social y la evaluación del impacto en las condiciones de vida de la población".

5.1. Justificación de la aplicación

Se considero oportuno analizar el contexto educativo debido a la importancia que implica este sector en la economía de nuestro país, pues la educación es uno de los derechos fundamentales de la niñez y adolescencia que genera beneficios como acceso a oportunidades laborales y mejoras salariales futuros. Ambos aspectos son considerados como un incentivo para que las familias decidan invertir en la educación de sus hijos. Para este propósito, las familias deben distribuir recursos que no solamente cubran mensualidades y matrículas, sino que también incluyan otros costos indirectos, como uniformes escolares, útiles escolares, transporte y otros. En la educación pública, si bien no existen pagos por matrículas, incluso los hogares más pobres, deben recurrir a costos indirectos.

5.2. Fuente de datos

La base de datos para esta aplicación proviene de la ENAHO sobre Condiciones de Vida y Pobreza. La ENAHO, que es realizada por el Instituto Nacional de Estadística e Informática (INEI), manifiesta tener como finalidad: "Proporcionar información estadística, demográfica, social y económica de los hogares, que permita medir la pobreza y caracterizar las condiciones de vida de la población. Medir estos resultados contribuirá al análisis y diseño de políticas en el área social y la evaluación del impacto en las condiciones de vida de la población".

5.3. Tipo de muestra

La muestra en la ENAHO es del tipo probabilística, de áreas, estratificada, multietápica e independiente en cada departamento de estudio.

5.4. Descripción de los datos

La ENAHO cuenta con los módulos de educación, salud, empleo, demografía y población. Como nuestro objetivo es modelar el gasto de los hogares destinados a la educación usaremos el módulo de educación y el módulo Summaria (resumen de indicadores). Los datos para la aplicación corresponden al panel de hogares de los periódos 2013-2017.

Definimos a la variable de interés como la suma de los gastos directos e indirectos destinados a la educación, realizados por algún miembro del hogar. Para los gastos directos se considera a la matrícula, mensualidades y algunos aportes a la asociación de padres y madres. Los gastos indirectos se deben a las asesorías particulares, uniformes escolares, libros o útiles escolares. Cabe tener en cuenta que los valores cero registrados de la respuesta significa que los hogares no están invertiendo en educación, a pesar de que algún miembro del hogar esté en una edad de hacerlo o, en caso contrario, se trate de un hogar que no tiene la necesidad de invertir en educación, como por ejemplo un hogar con presencia de solo personas mayores.

En esta aplicación buscaremos identificar las variables o factores que determinan los niveles de gasto en educación de los hogares en el Perú. Para ello, las posibles covariables a considerar en este modelo serán:

- **Pobreza monetaria:** Indicador de bienestar que utiliza al gasto en que incurre el hogar por compras, autoconsumo, autosuministro, pagos en especies, transferencias de otros hogares y donaciones públicas. Las categorías de esta variable son: pobre extremo= 1, pobre no extremo= 2 y no pobre= 3.
- **Nivel educativo del jefe de hogar:** Máximo nivel educativo alcanzado por el jefe de hogar, donde las categorias estan codificadas como: sin estudios= 1, nivel primaria= 2, nivel secundaria = 3 y nivel universitario= 4.
- **Sexo del jefe de hogar**: Dicotomizó como = 1, si es hombre, y = 0 si es mujer.
- **Edad del jefe de hogar**: Edad en años del jefe de hogar al momento de la encuesta.
- **Área geográfica**: Ámbito geográfico donde se ubica el hogar, siendo: 0 rural= 0 y 1 urbano.
- **Región natural**: Región donde se ubica el hogar, siendo: = 1 costa, 2 sierra y 3 selva.
- **Miembros del hogar**: Indica el total de integrantes que conforman el hogar.

5.5. Análisis descriptivo

Analizando los cuadros y las siguientes figuras, se propone la inclusión de algunas características como covariables en el modelo de regresión. Entonces una suposición es que el gasto de los hogares destinados a la educación se asocia fuertemente con el máximo nivel educativo alcanzado por el jefe de hogar, sexo del jefe de hogar, el nivel de pobreza monetaria, región natural y el ámbito geográfico (urbano y/o rural) donde se ubica el hogar.

Cuadro 5.1: Número de hogares y el gasto promedio (miles de soles) destinados a la educación, según el año y el nivel educativo del jefe de hogar

Año/nivel	**Número de hogares**			**Gasto promedio de los hogares**
	No gastó	**Si gastó**	**Total**	
2013	54	673	727	7031.46
Sin estudios	7	31	38	136.716
Primaria	19	272	291	2177.195
Secundaria	22	236	258	2459.958
Universitaria	6	134	140	2257.591
2014	41	686	727	7703.839
Sin estudios	3	35	38	179.185
Primaria	15	279	294	2082.519
Secundaria	16	249	265	2553.955
Universitaria	7	123	130	2888.18
2015	37	690	727	8180.192
Sin estudios	2	31	33	201.142
Primaria	13	277	290	1911.291
Secundaria	19	251	270	3328.079
Universitaria	3	131	134	2739.68
2016	35	692	727	9465.361
Sin estudios	2	34	36	133.97
Primaria	14	279	293	2492.644
Secundaria	16	250	266	3344.814
Universitaria	3	129	132	3493.933
2017	43	684	727	8855.908
Sin estudios	3	33	36	216.413
Primaria	18	269	287	2184.07
Secundaria	16	252	268	3878.781
Universitaria	6	130	136	2576.644

Cuadro 5.2: Número de hogares y el gasto promedio (miles de soles) destinados a la educación, según el año y el sexo del jefe de hogar

Año/sexo	**Número de hogares**			**Gasto promedio de los hogares**
	No gastó	**Si gastó**	**Total**	
2013	54	673	727	7031.46
hombre	43	561	604	5864.90
mujer	11	112	123	1166.56
2014	41	686	727	7703.84
hombre	35	574	609	6280.93
mujer	6	112	118	1422.91
2015	37	690	727	8180.19
hombre	32	580	612	7160.32
mujer	5	110	115	1019.87
2016	35	692	727	9465.36
hombre	30	583	613	8033.56
mujer	5	109	114	1431.80
2017	43	684	727	8855.91
hombre	37	575	612	7629.43
mujer	6	109	115	1226.48

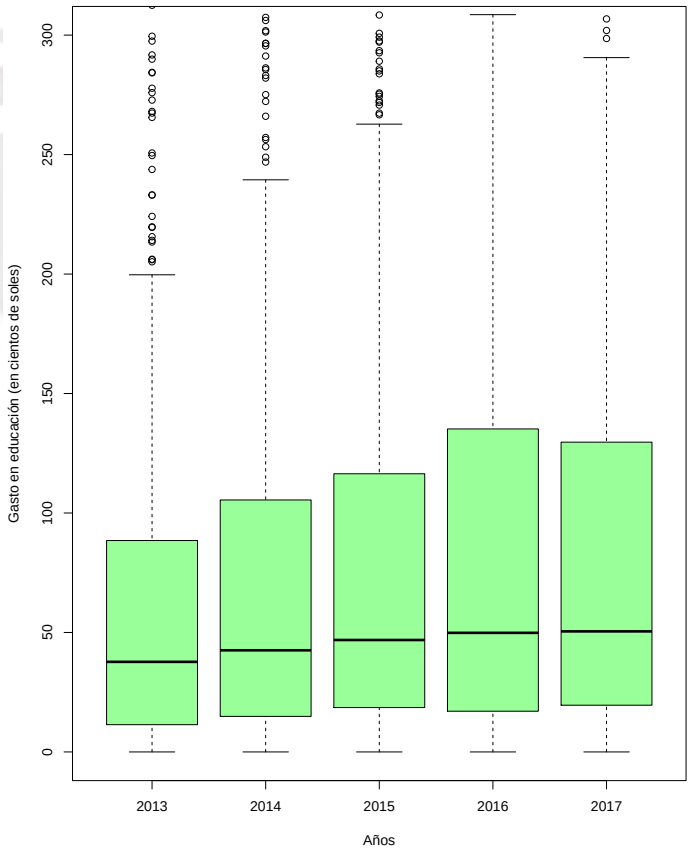

Figura 5.1: Evolución del gasto en educación para los periódos del 2013 al 2017.

Según la figura 5.1 el gasto de los hogares destinados a la educación (en cientos de soles) se mantiene casi constante con una pequeña tendencia al alza, lo cual era de esperarse, pues se trata de un panel de hogares. Por otro lado, según la figura 5.2 se puede verificar que los hogares más ricos son los que más invierten en educación, mientras que los hogares en pobreza extrema son los que menos invierten en educación.

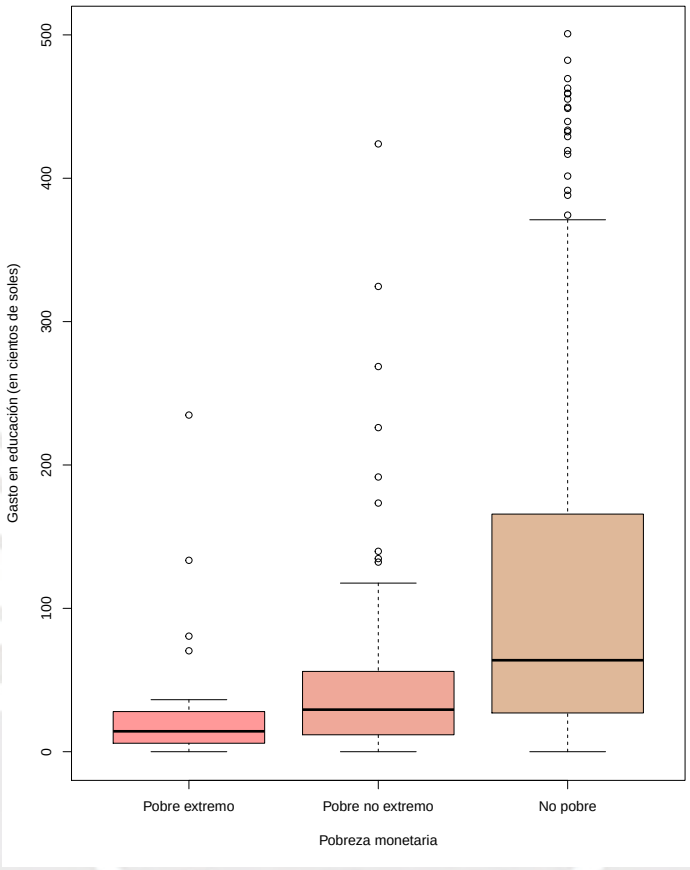

Figura 5.2: Gastos en educación, según el nivel de pobreza del hogar.

Respecto al nivel educativo del jefe de hogar se pudo observar que el nivel educativo del jefe de hogar está altamente asociado al gasto destinado a educación, es decir, los hogares donde el jefe de hogar cuenta con un nivel educativo superior son los que más invierten en educación, mientras que los que menos invierten son los que no tienen estudios.

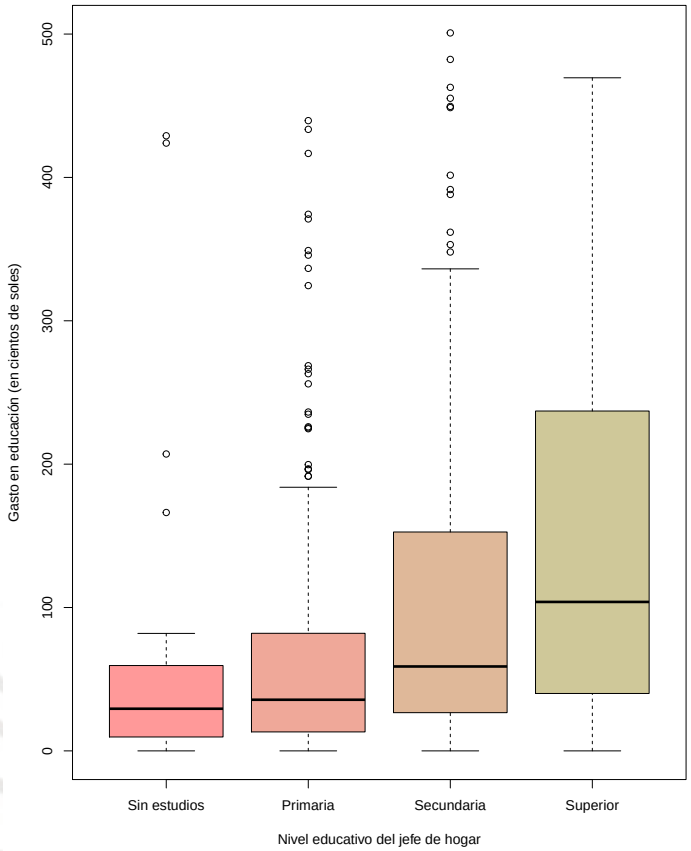

Figura 5.3: Gastos en educación, según nivel educativo alcanzado por el jefe de hogar.

Analizando el gasto en relación a los hogares según región natural, se puede concluir que en la región costa existe una mayor inversión de los hogares en educación, mientras que en la sierra y la selva se observa un menor inversión. Este comportamiento se debe basicamente a la brecha que existe en las condiciones de vida.

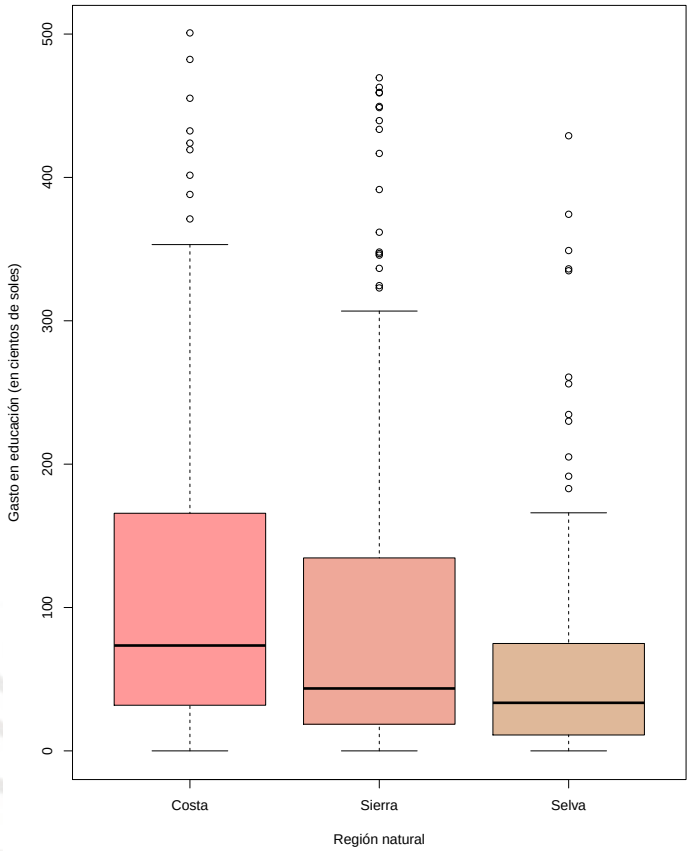

Figura 5.4: Gastos en educación, según región natural.

Un análisis por tipo de ámbito geográfico nos muestra que los hogares de las zonas urbanas son los que más invierten en educación, comparado con el ámbito rural. Por otra parte si analizamos el gasto por sexo del jefe de hogar, no existe diferencias marcadas en la inversión en educación.

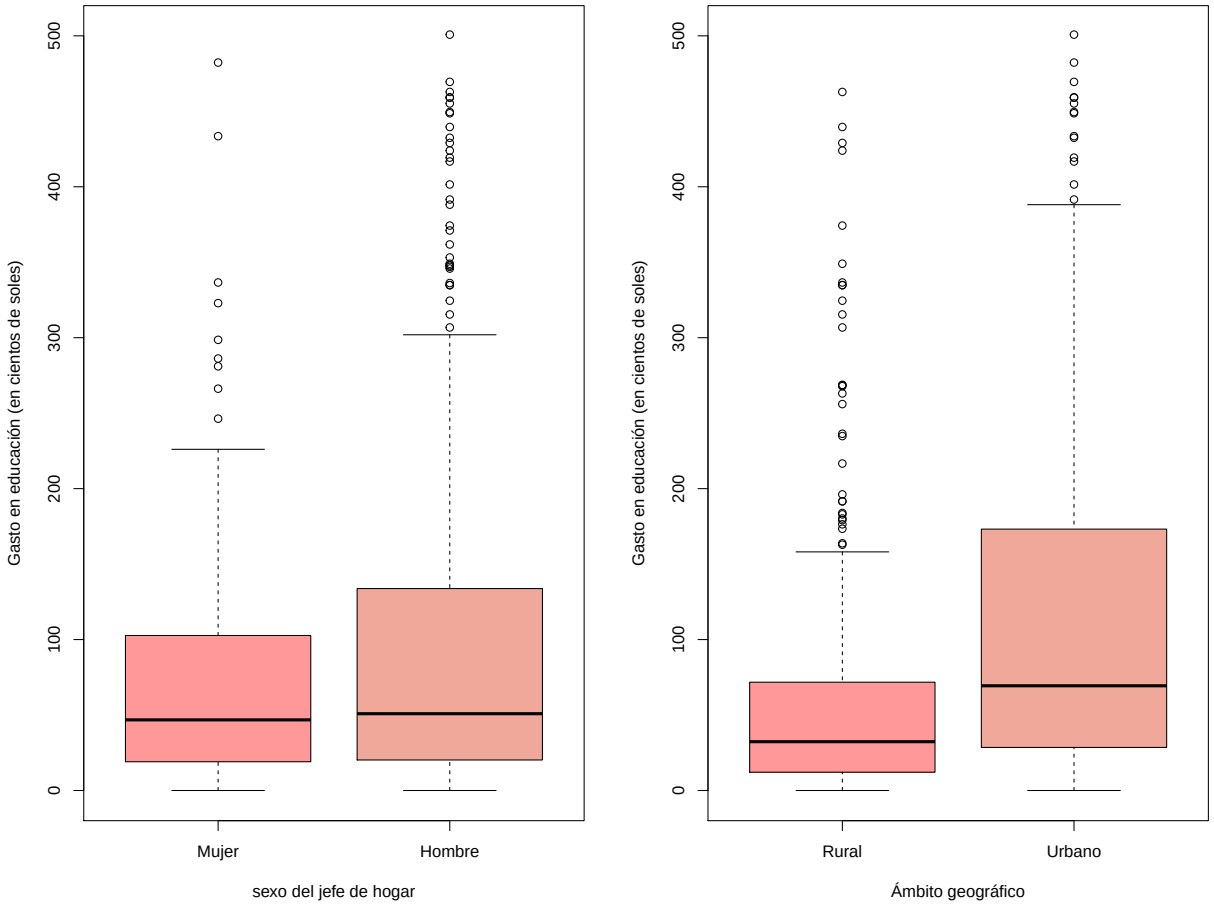

Figura 5.5: Gastos en educación, según el ámbito geográfico y sexo del jefe de hogar.

5.6. Especificación del modelo

En el modelo MRMS-GGCI asumiremos que la variable respuesta "Gasto de los hogares destinados a la educación" para el hogar i tiene distribución $Y_{ij} \sim GGCI(\pi_{ij}, \gamma_{ij}, \kappa, \sigma)$, donde los parámetros $\pi_i = P(Y_i = 0)$ y $\gamma_i = E(Y_i)$ estarán relacionados con un conjunto de covariables mediante:

$$Y_{ij} \sim GGCI(\pi_{ij}, \gamma_{ij}, \kappa, \sigma)$$

$$\text{logit}(\pi_{ij}) = \alpha_0 + \alpha_1 x_{1ij} + \alpha_2 x_{2ij} + \alpha_3 x_{3ij} + \alpha_4 x_{4ij} + \alpha_5 x_{5ij} + \alpha_6 x_{6ij} + \alpha_7 x_{7ij} + \alpha_8 x_{8ij} + a_i$$

$$y$$

$$\log(\gamma_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} + \beta_7 x_{7ij} + \beta_8 x_{8ij} + d_i,$$

donde la variable x_{1ij} es el "sexo del jefe hogar", x_{2ij} es la"edad del jefe de hogar", x_{3ij} es "nivel educativo primaria", x_{4ij} es "nivel educativo secundaria", x_{5ij} es "nivel educativo superior", x_{6ij} es el "número de miembros del hogar", x_{7ij} es el "hogar en pobreza no extrema", y por último x_{8ij} es el "hogar no pobre". Se debe tener en cuenta que en la espe-

cificación del modelo anterior los parámetros κ y σ son considerados constantes para todas las observaciones.

5.7. Resultados

En el cuadro 5.3 se aprecian las estimaciones de los parámetros de regresión del modelo MRMS-GGCI, los errores estándar, los cuantiles de los parámetros que definen los intervalos de credibilidad y la exponencial de los coeficientes estimados para fines de interpretación.

En el modelo MRMS-GGCI, la exponenciación de los coeficiente de regresión β_j pueden interpretarse como efectos multiplicativos. Esto es, dado un incremento en una unidad de la correspondiente covariable, ella mide en que proporción se incrementa o reduce el valor esperado de la variable respuesta, $E(Y_{ij}) = \gamma_{ij}$. Por ejemplo, el efecto multiplicativo de x_{2ij} está dado por:

$$\frac{E\left(Y_{ij}|x_{2ij} = p+1\right)}{E\left(Y_{ij}|x_{2ij} = p\right)} = \exp\left(\beta_2\right),$$

donde p es cualquier valor dado. Recordando que la variable x_{2ij} es la edad en años del jefe de hogar, tendremos que un incremento de un año en la edad del jefe de hogar, se esperará que incremente la media del gasto en educación en un $(e^{\beta_2} - 1)100\,\%$.

Usando la ENAHO sobre el periódo 2013-2017, se procedió a estimar los parámetros del modelo. Para esto se consideraron, en el HMC, 10 mil iteraciones y se eliminaron las primeras 2 mil iteraciones como periódo de burn-in. La convergencia del proceso se evaluó observando el comportamiento de las cadenas para cada parámetro, los cuales junto a las autocorrelaciones mostrarón evidencias de convergencia. A continuación se muestra el detalle de la estimación.

Para el modelo MRMS-GGCI se obtuvieron los siguientes resultados para la parte discreta, si el jefe de hogar es hombre el ratio de odds de la probabilidad de no gastar en educación se incrementa en un 34 %. El nivel de pobreza del hogar, el número de miembros del hogar y el nivel educativo del jefe de hogar tienen, por otro lado, un efecto negativo sobre la probabilidad de que no se gaste en educación, es decir, aparentemente estas covariables son determinantes en el gasto directo.

Analizando los resultados para la parte continua, si el jefe de hogar es hombre conllevaría a una disminución en un 16 % del valor esperado del Gasto en educación, mientras que a medida que se incremente en una unidad el número de miembros en el hogar conllevaría al incremento en un 24 % del valor esperado del Gasto en educación. Por otra parte el impacto del nivel educativo del jefe de hogar es más significativo a medida que posea mayor formación educativa.

Finalmente, acorde a los criterios WAIC y LOO, calculados en el cuadro 5.4 para los distintos modelos, verificamos que el modelo de regresión a la media con efectos mixtos para respuestas semicontinuas con distribución gamma generalizada cero-inflacionda en su parte continua (MRMS-GGCI) tiene un mejor ajuste respecto al modelo de regresión a la media con efectos mixtos para respuestas semicontinuas con distribución gamma cero-inflacionada en su parte continua (MRMS-GCI) y el modelo de regresión a la media con efectos mixtos para respuestas semicontinuas con distribución Log Skew Normal en su parte continua (MRMS-LSNCI).

Cuadro 5.3: Estimación de los coeficientes de regresión del modelo MRMS-GGCI

Parámetro	**Covariable**	$\hat{\beta}$	**Error estándar**	**IC (95 %)**	**Rhat**	**Exp($\hat{\beta}$)**
π_{ij}	Intercepto	-0.309	0.137	(-2.829, 2.754)	1.003	0.734
	Sexo del jefe de hogar	0.296	0.008	(-0.094, 0.657)	0.999	1.344
	Edad del jefe de hogar	-0.001	0.000	(-0.014, 0.01)	0.999	0.999
	Jefe del hogar con educación primaria	-0.535	0.011	(-1.099, 0.067)	0.999	0.586
	Jefe del hogar con educación secundaria	-0.221	0.011	(-0.849, 0.431)	0.999	0.802
	Jefe del hogar con educación superior	-0.843	0.013	(-1.532, -0.049)	1.000	0.430
	Número de miembros del hogar	-0.332	0.002	(-0.44, -0.233)	1.004	0.717
	Pobre no extremo	-0.647	0.010	(-1.114, -0.133)	0.999	0.524
	No pobre	-0.997	0.009	(-1.459, -0.486)	1.000	0.369
γ_{ij}	Intercepto	1.437	0.112	(-3.719, 3.795)	0.999	4.208
	Sexo del jefe de hogar	-0.175	0.002	(-0.291, -0.066)	1.000	0.839
	Edad del jefe de hogar	0.002	0.000	(-0.003, 0.006)	0.999	1.002
	Jefe del hogar con educación primaria	0.216	0.004	(0.002, 0.421)	1.003	1.241
	Jefe del hogar con educación secundaria	0.636	0.004	(0.416, 0.859)	1.003	1.889
	Jefe del hogar con educación superior	1.180	0.004	(0.944, 1.408)	1.003	3.254
	Número de miembros del hogar	0.214	0.000	(0.189, 0.24)	0.999	1.239
	Pobre no extremo	0.870	0.004	(0.672, 1.057)	1.001	2.387
	No pobre	1.636	0.003	(1.46, 1.817)	1.000	5.135
σ_a		6.335	0.287	(3.376, 8.299)	1.023	
σ_d		6.251	0.298	(3.312, 8.559)	1.065	
κ		0.277	0.001	(0.214, 0.344)	1.000	
σ		1.229	0.001	(1.202, 1.259)	1.000	

Cuadro 5.4: Criterios de comparación de los modelos alternativos

Criterios	**MRMS-GGCI**	**MRMS-GCI**	**MRMS-LSN**
WAIC	-16.7	-12.5	-14.1
LOO	-1.6	1.2	1.0

Capítulo 6

Conclusiones

Se presentó en este trabajo, la teoría, implementación y aplicación de un nuevo modelo de regresión a la media con efectos mixtos para respuestas semicontinuas que siguen una distribución gamma generalizada cero-inflacionada (MRMS-GGCI). Este modelo se sugirió como alternativa al modelo de regresión a la media con efectos mixtos para respuestas semicontinuas con distribución Log Skew Normal cero-inflacionada (MRMS-LSNCI). Ambos modelos permiten no solo interpretar los efectos de un conjunto de covariables sobre la media marginal de la respuesta, en vez de la media condicionada a valores positivos, sino también estimar los efectos que estas covariables podrían ejercer sobre la probabilidad de que la respuesta tome el valor cero o no. Mediante un estudio de simulación, hemos verificado que las estimaciones obtenidas en el modelo MRMS-GGCI por *Stan*, a través del método de las cadenas de Markov de Monte Carlo Hamiltoniano y el uso del muestreador No-U-Turn (NUTS), tuvieron un adecuado desempeño en términos del sesgo y la raíz del error cuadrático medio. Finalmente, realizamos una aplicación del MRMS-GGCI para investigar que factores están relacionados con los gastos en educación de los hogares en el Perú. Para ello consideramos la Encuesta Nacional de Hogares (ENAHO), con datos de tipo panel entre los años 2013 y 2017. El modelo MRMS-GGCI mostró, en esta aplicación, un mejor ajuste a los datos con respecto al modelo alternativo MRMS-LSNCI.

6.1. Sugerencias para investigaciones futuras

Como trabajos futuros al desarrollo de este nuevo modelo se podría sugerir los siguientes:

- Estudiar el MRMS-GGCI bajo una especificación que incluya una ecuación de regresión para el parámetro de dispersión de la parte continua del modelo.
- Estudiar el MRMS-GGCI con una especificación que incluya funciones de enlace distintas a la logarítmica o la logística.
- Aplicar el MRMS-GGCI a datos de semicontinuos provenientes de otras encuestas con formato de panel como la ENDES y los niños de milenio ("Young lives").
- Implementar la estimación del modelo por MCMC usando otros métodos y programas, como JAGS, INLA, entre otros.

Apéndice A

Código en R y Stan: Simulación y aplicación

A.1. Código en R para la simulación de los datos usando la distribución Gamma Generalizada Cero-Inflacionada-GGCI

```
#Simulacion de la data usando la distribución GGCI
rm(list = ls(all=TRUE))
#install.packages("gamlss.dist")
#install.packages("boot")
#install.packages("ggplot2")
#install.packages("rstan", repos = "https://cloud.r-project.org/")
library(gamlss.dist)
library(boot)
library(ggplot2)
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)

set.seed(2018)
R   <- 300       #Número de réplicas
n   <- 100       #Número de individuos
m   <- 5         #Número de mediciones por individuo

#Definiendo los parámetros a recuperar
k  <- 0.10 #Kappa
s  <- 0.30 #Sigma

alpha0 <- -1
alpha1 <- 0.25
beta0  <- -0.5
beta1  <- 1.1
sigmaa <- 1
sigmad <- 1
```

```
#Definiendo la covariable
#-------------------------------------------------------
X <- matrix(0,n,m)
for (i in 1:n) {
  for (j in 1:m) {
    X[i,j] <- rnorm(1,mean=0.5,sd=1) + j*runif(1,min=0,max=0.1)
  }
}

X <- as.vector(t(X))

#Lista de datos
#-------------------------------------------------------
BD=list()
for(i in 1:R) {

BD[[i]]=cbind.data.frame(ID =rep(seq_len(n), each = m),
                         tiempo = c(rep(1:5,n)),X)

                #Efectos aleatorios de la parte cero
                a0 = replicate(n,rep(rnorm(1,mean=0, sd=sigmaa),m))
                BD[[i]]$a=as.vector(a0)

                #Efectos aleatorios de la parte continua
                d0 = replicate(n,rep(rnorm(1,mean=0, sd=sigmad),m))
                BD[[i]]$d=as.vector(d0)

                #Predictor lineal para la parte continua
                eta_gamma <- beta0+beta1*X+BD[[i]]$d
                BD[[i]]$gammaij <- exp(eta_gamma)

                #Predictor lineal para la parte cero
                eta_pi <- alpha0+alpha1*X+BD[[i]]$a
                BD[[i]]$pij <- inv.logit(eta_pi)

                #Simulando la variabel Yij ~ GGCI(pi,gamma,kappa,sigma)
                BD[[i]]$Y <-rep(0,n*m)
                BD[[i]]$u <-runif(n*m,0,1)
                BD[[i]]$h <-(BD[[i]]$u-BD[[i]]$pij)/(1-BD[[i]]$pij)
                index <- which(BD[[i]]$u > BD[[i]]$pij,arr.ind=FALSE)
                lambda <- log(BD[[i]]$gammaij[index])-log(1-BD[[i]]$pij[index])-
                ((2*s)/k)*log(k)+log(gamma(1/k^2))-log(gamma((1/k^2)+(s/k)))
```

```
                #Transformando los parámetros Y~ gGG(x|mu,sigma,nu)
                #mu=vector of location parameter values
                #sigma=vector of scale parameter values
                #nu=vector of shape parameter values
                sigma0 <- k*s
                nu0 <- k/s
                mu0 <- exp(lambda)
                BD[[i]]$Y[index] <-qGG(p=BD[[i]]$h[index], mu = mu0, sigma =sigma0,
                                       nu=nu0,lower.tail = TRUE, log.p = FALSE)
        }
```

A.2. Código en Stan para la recuperación de parámetros usando la distribución Gamma Generalizada Cero-Inflacionada-GGCI

```
GammaGCI <-

'
// Definimos la función Gamma Generalizada-GG

functions {
real GammaG_lpdf(real x, real lambda, real kappa, real sigma)
{

real etha;
real GG;
etha = exp((-kappa*lambda)/sigma);
GG   = log(kappa/sigma)-log(tgamma(1/(kappa^2)))+(1/(kappa*sigma)-1)*log(x)-
       (etha/(kappa^2))*x^(kappa/sigma)+(1/(kappa^2))*log(etha/(kappa^2));
return(GG);

  }

// Definimos la función Gamma Generalizada Cero Inflacionada-GGCI

  real GGCI_lpdf(real x, real pi, real gamma0, real kappa, real sigma)
  {
    real prob;
    real lprob;
    real lambda;
    lambda = log(gamma0)-log(1-pi)-((2*sigma)/kappa)*log(kappa)+
         log(tgamma(1/(kappa^2)))-log(tgamma((1/(kappa^2))+(sigma/kappa)));
```

```
    if (x == 0)
      prob = pi;

    else
    {
      prob = (1-pi);
      prob = prob*exp(GammaG_lpdf(x|lambda,kappa,sigma));
    }
    lprob = log(prob);
    return(lprob);
  }
     }

data
{
  int<lower=0> N;        // número de casos
  int<lower=1> M;        // número de sujetos
  real<lower=0> Y[N];    // variables respuesta
  int<lower=1> ID[N];    // identificador
  int K1;                // número de columnas de la matriz de efectos fijos
  matrix[N,K1] X;        // matriz del modelo para los efectos fijos
    }

parameters
  {
  vector[K1] beta;          // Efectos fijos (gamma0))
  real di[M];               // Efectos aleatorios (gamma0)
  real<lower=0> sigmad;     // varianza de los efectos aleatorios(gamma0)
  vector[K1] alpha;         // Efectos fijos (pi)
  real ai[M];               // Efectos aleatorios (pi)
  real<lower=0> sigmaa;     // varianza de los efectos aleatorios(pi)
  real<lower=0.001> kappa; // parámetro kappa
  real<lower=0.001> sigma; // parámetro sigma
}

transformed parameters
{
  vector[N] linpred_gamma0;
  vector[N] gamma0_est;
  vector[N] linpred_pi;
  vector[N] pi_est;
```

```
  linpred_gamma0 = X*beta; // Predictor lineal solo con efectos fijos(gamma0)
  linpred_pi = X*alpha;    // Predictor lineal solo con efectos fijos(pi)

for(i in 1:N)
{

  // Agregando efectos aleatorios al predictor lineal (pi)
  linpred_pi[i] = linpred_pi[i] + ai[ID[i]];

  // Agregando efectos aleatorios al predictor lineal (gamma0)
  linpred_gamma0[i] = linpred_gamma0[i] + di[ID[i]];

  pi_est[i] = inv_logit(linpred_pi[i]);
  gamma0_est[i] = exp(linpred_gamma0[i]);
   }
}

model
{
  // Definición de a prioris

  sigmaa ~ inv_gamma(5,50);
  sigmad ~ inv_gamma(5,50);

for(j in 1:M)
{
  di[j] ~ normal(0,sigmad); // efectos aleatorios(gamma0)
  ai[j] ~ normal(0,sigmaa); // efectos aleatorios(pi)
}
  for(i in 1:K1)
  {
  alpha[i] ~ normal(0,10000); // efectos fijos (pi)
  beta[i] ~ normal(0,10000); //  efectos fijos (gamma0)

  }

  kappa ~ gamma(0.1,0.01); // parámetro kappa
  sigma ~ gamma(0.1,0.01); // parámetro sigma

 // Distribución de la variable dependiente

for(j in 1:N)
```

```
  {
  Y[j] ~ GGCI(pi_est[j],gamma0_est[j], kappa, sigma);
  }
}
  ,
```

A.3. Aplicación del modelo GGCI

```
#Cargando la data y las librerias necesarias
#-----------------------------------------------------------------
rm(list = ls(all=TRUE))
library(haven)
library(boot)
library(ggplot2)
library(rstan)
library(loo)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)

setwd("D:/TESIS_FINAL/APLICACIÓN")
datos <- read_dta("BD_Tesis.dta")

datos$sexo_jh<- factor(datos$sexo_jh,levels = c(0,1),labels = c("Mujer", "Hombre"))
datos$area <- factor(datos$area,levels = c(0,1),labels = c("Rural", "Urbano"))
datos$reg_natural <- factor(datos$reg_natural,levels = c(1,2,3),
labels = c("Costa", "Sierra","Selva"))
datos$nivel_jh<-factor(datos$nivel_jh,levels = c(1,2,3,4),
labels = c("Sin estudios", "Primaria","Secundaria","Superior"))
datos$estado_civil<-factor(datos$estado_civil,levels = c(1,2,3,4,5,6),
labels = c("conviviente", "Casado","Viudo","Divorciado","Separado","Soltero"))
datos$pobreza <-factor(datos$pobreza,levels = c(1,2,3),
labels = c("Pobre extremo", "Pobre no extremo","No pobre"))
datos$ingr_neto<-datos$ingr_neto/100
datos$gasto_educ<-datos$gasto_educ/100

datos$id<-as.numeric(datos$id)

#Definiendo la función GGCI en Stan
#-----------------------------------------------------
GammaGCI <-
 ,
// Definimos la función Gamma Generalizada-GG
```

```
functions {
real GammaG_lpdf(real x, real lambda, real kappa, real sigma)

{
  real etha;
  real GG;
  etha = exp((-kappa*lambda)/sigma);
  GG   = log(kappa/sigma)-log(tgamma(1/(kappa^2)))+(1/(kappa*sigma)-1)
  *log(x)-(etha/(kappa^2))*x^(kappa/sigma)+(1/(kappa^2))*log(etha/(kappa^2));
  return(GG);
}

// Definimos la función Gamma Generalizada Cero Inflacionada-GGCI

real GGCI_lpdf(real x, real pi, real gamma0, real kappa, real sigma)
{
  real prob;
  real lprob;
  real lambda;
  lambda = log(gamma0)-log(1-pi)-((2*sigma)/kappa)*log(kappa)+
  log(tgamma(1/(kappa^2)))-log(tgamma((1/(kappa^2))+(sigma/kappa)));

  if (x == 0)
  prob = pi;

  else
  {
  prob = (1-pi);
  prob = prob*exp(GammaG_lpdf(x|lambda,kappa,sigma));
  }

  lprob = log(prob);
  return(lprob);
}
  }

data
{
  int<lower=0> N;      // número de casos
  int<lower=1> M;      // número de sujetos
  real<lower=0> Y[N];  // variables respuesta
```

```
  int<lower=1> ID[N];  // identificador
  int K1;              // número de columnas de la matriz del modelo de efectos fijos
  matrix[N,K1] X;      // matriz del modelo para los efectos fijos
 }

parameters
{
  vector[K1] beta;         // Efectos fijos (gamma0))
  real di[M];              // Efectos aleatorios (gamma0)
  real<lower=0> sigmad;    // varianza de los efectos aleatorios(gamma0)
  vector[K1] alpha;        // Efectos fijos (pi)
  real ai[M];              // Efectos aleatorios (pi)
  real<lower=0> sigmaa;    // varianza de los efectos aleatorios(pi)
  real<lower=0.001> kappa; // parámetro kappa
  real<lower=0.001> sigma; // parámetro sigma
}

transformed parameters
{
  vector[N] linpred_gamma0;
  vector[N] gamma0_est;
  vector[N] linpred_pi;
  vector[N] pi_est;
  linpred_gamma0 = X*beta; // Cálculo del predictor lineal solo con efectos fijos
  linpred_pi = X*alpha;    // Cálculo del predictor lineal solo con efectos fijos

  for(i in 1:N)
  {

  // Agregando efectos aleatorios al predictor lineal (pi)
  linpred_pi[i] = linpred_pi[i] + ai[ID[i]];

  // Agregando efectos aleatorios al predictor lineal (gamma0)
  linpred_gamma0[i] = linpred_gamma0[i] + di[ID[i]];

  pi_est[i] = inv_logit(linpred_pi[i]);
  gamma0_est[i] = exp(linpred_gamma0[i]);

  }
    }

model
```

```
{
  // Definición de a prioris

  sigmaa ~ inv_gamma(5,50);
  sigmad ~ inv_gamma(5,50);

  for(j in 1:M)
  {

  di[j] ~ normal(0,sigmad); // efectos aleatorios(gamma0)
  ai[j] ~ normal(0,sigmaa); // efectos aleatorios(pi)

  }

  for(i in 1:K1)
  {
  alpha[i] ~ normal(0,10000); // efectos fijos (pi)
  beta[i] ~ normal(0,10000); //  efectos fijos (gamma0)
    }
  kappa ~ gamma(0.01,0.01); // parámetro kappa
  sigma ~ gamma(0.01,0.01); // parámetro sigma

  // Distribución de la variable dependiente

  for(j in 1:N)
  {
  Y[j] ~ GGCI(pi_est[j],gamma0_est[j], kappa, sigma);
  }
}
  ,
#Estimación del modelo MRMS
#-------------------------------------------------------

#Declarando la data
#---------------------
BD <- list(N = length(datos$id),
           M= length(datos$id)/5,
           X= model.matrix(~ datos$sexo_jh+datos$edad_jh+factor(datos$nivel_jh)+
           datos$totmieho+factor(datos$pobreza)),
           Y= datos$gasto_educ,
           ID= datos$id,
           K1=9)
```

```
#Ajuste
#---------------------
fit <- stan(model_code = GammaGCI , data = BD , iter = 10000 , warmup = 2000,
init="random" , chains = 1,thin = 10)
print(fit , digits=3 , pars=c("alpha","beta","sigmaa","sigmad","kappa","sigma"),
probs=c(0.025,0.5,0.975))
traceplot( fit , c("alpha","beta","sigmaa","sigmad","kappa","sigma"))
traceplot( fit , c("sigmaa","sigmad","kappa","sigma"))

#Densidad a posteriori
#---------------------
posterior <-extract(fit)
par(mfrow=c(3,3))
plot(density(posterior$alpha[,1]), main="Intercepto",col=2)
plot(density(posterior$alpha[,2]), main="Sexo",col=2)
plot(density(posterior$alpha[,3]), main="Edad",col=2)
plot(density(posterior$alpha[,4]), main="Nivel primaria",col=2)
plot(density(posterior$alpha[,5]), main="Nivel secundaria",col=2)
plot(density(posterior$alpha[,6]), main="Nivel superior",col=2)
plot(density(posterior$alpha[,7]), main="Miembros del hogar",col=2)
plot(density(posterior$alpha[,8]), main="Pobre no extremo",col=2)
plot(density(posterior$alpha[,9]), main="No pobre",col=2)

par(mfrow=c(3,3))
plot(density(posterior$beta[,1]), main="Intercepto",col=3)
plot(density(posterior$beta[,2]), main="Sexo",col=3)
plot(density(posterior$beta[,3]), main="Edad",col=3)
plot(density(posterior$beta[,4]), main="Nivel primaria",col=3)
plot(density(posterior$beta[,5]), main="Nivel secundaria",col=3)
plot(density(posterior$beta[,6]), main="Nivel superior",col=3)
plot(density(posterior$beta[,7]), main="Miembros del hogar",col=3)
plot(density(posterior$beta[,8]), main="Pobre no extremo",col=3)
plot(density(posterior$beta[,9]), main="No pobre",col=3)

par(mfrow=c(2,2))
plot(density(posterior$sigmaa), main="Sigma_a",col=6,type="l",xlab="",ylab="")
plot(density(posterior$sigmad), main="Sigma_d",col=6,type="l",xlab="",ylab="")
plot(density(posterior$kappa), main="Kappa",col=6,type="l",xlab="",ylab="")
plot(density(posterior$sigma), main="Sigma",col=6,type="l",xlab="",ylab="")

#Gráfico de cadenas
```

```
#---------------------
par(mfrow=c(3,3))
plot(posterior$alpha[,1], main="Intercepto",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,2], main="Sexo",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,3], main="Edad",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,4], main="Nivel primaria",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,5], main="Nivel secundaria",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,6], main="Nivel superior",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,7], main="Miembros del hogar",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,8], main="Pobre no extremo",col=4,type="l",xlab="",ylab="")
plot(posterior$alpha[,9], main="No pobre",col=4,type="l",xlab="",ylab="")

par(mfrow=c(3,3))
plot(posterior$beta[,1], main="Intercepto",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,2], main="Sexo",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,3], main="Edad",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,4], main="Nivel primaria",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,5], main="Nivel secundaria",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,6], main="Nivel superior",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,7], main="Miembros del hogar",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,8], main="Pobre no extremo",col=2,type="l",xlab="",ylab="")
plot(posterior$beta[,9], main="No pobre",col=2,type="l",xlab="",ylab="")

par(mfrow=c(2,2))
plot(posterior$sigmaa, main="Sigma_a",col=3,type="l",xlab="",ylab="")
plot(posterior$sigmad, main="Sigma_d",col=3,type="l",xlab="",ylab="")
plot(posterior$kappa, main="Kappa",col=3,type="l",xlab="",ylab="")
plot(posterior$sigma, main="Sigma",col=3,type="l",xlab="",ylab="")

#Gráfico de autocorrelaciones
#---------------------
par(mfrow=c(3,3))
acf(posterior$alpha[,1], main="Intercepto")
acf(posterior$alpha[,2], main="Sexo")
acf(posterior$alpha[,3], main="Edad")
acf(posterior$alpha[,4], main="Nivel primaria")
acf(posterior$alpha[,5], main="Nivel secundaria")
acf(posterior$alpha[,6], main="Nivel superior")
acf(posterior$alpha[,7], main="Miembros del hogar")
acf(posterior$alpha[,8], main="Pobre no extremo")
acf(posterior$alpha[,9], main="No pobre")
```

```
par(mfrow=c(3,3))
acf(posterior$beta[,1], main="Intercepto")
acf(posterior$beta[,2], main="Sexo")
acf(posterior$beta[,3], main="Edad")
acf(posterior$beta[,4], main="Nivel primaria")
acf(posterior$beta[,5], main="Nivel secundaria")
acf(posterior$beta[,6], main="Nivel superior")
acf(posterior$beta[,7], main="Miembros del hogar")
acf(posterior$beta[,8], main="Pobre no extremo")
acf(posterior$beta[,9], main="No pobre")

par(mfrow=c(2,2))
acf(posterior$sigmaa, main="Sigma_a")
acf(posterior$sigmad, main="Sigma_d")
acf(posterior$kappa, main="Kappa")
acf(posterior$sigma, main="Sigma")

#WAIC y LOO
#---------------------
log_lik1 <- extract_log_lik(fit,parameter_name=c("alpha","beta","sigmaa",
"sigmad","kappa","sigma"), merge_chains = TRUE)

waic1 <- waic(log_lik1)
loo1 <- loo(log_lik1)

#Guardando el fit
#---------------------
saveRDS(fit, file = "fit.rds")
fit <- readRDS("fit.rds")
```

Bibliografía

Bayes, C. L. y Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables, *Journal of Applied Statistics* **43**(10): 1814–1830.

Duan, N., Manning, W. G., Morris, C. N. y Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care, *Journal of business & economic statistics* **1**(2): 115–126.

Duane, S., Kennedy, A. D., Pendleton, B. J. y Roweth, D. (1987). Hybrid monte carlo, *Physics letters B* **195**(2): 216–222.

Gumedze, F. y Dunne, T. (2011). Parameter estimation and inference in the linear mixed model, *Linear Algebra and its Applications* **435**(8): 1920–1944.

Hoffman, M. D. y Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo., *Journal of Machine Learning Research* **15**(1): 1593–1623.

Manning, W. G., Basu, A. y Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data, *Journal of health economics* **24**(3): 465–488.

Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics, *Handbook of Markov Chain Monte Carlo* **2**(11): 2.

Olsen, M. K. y Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association* **96**(454): 730–745.

Smith, V. A., Neelon, B., Preisser, J. S. y Maciejewski, M. L. (2017). A marginalized two-part model for longitudinal semicontinuous data, *Statistical methods in medical research* **26**(4): 1949–1968.

Smith, V. A., Preisser, J. S., Neelon, B. y Maciejewski, M. L. (2014). A marginalized two-part model for semicontinuous data, *Statistics in medicine* **33**(28): 4891–4903.

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.

Stacy, E. W. (1962). A generalization of the gamma distribution, *The Annals of mathematical statistics* pp. 1187–1192.

Stacy, E. W. y Mihram, G. A. (1965). Parameter estimation for a generalized gamma distribution, *Technometrics* **7**(3): 349–358.

Vehtari, A., Gelman, A. y Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic, *Statistics and computing* **27**(5): 1413–1432.

Vásquez (2018). *Modelos de regresión gamma generalizada cero-inflacionada para la media con aplicación a gastos en educación de hogares en situación de pobreza*, Master's thesis, Pontificia Universidad Católica del Perú.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**(Dec): 3571–3594.