PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Modelo de regresión Dirichlet bayesiano: Aplicación para estimar la prevalencia del nivel de anemia infantil en centros poblados del Perú

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Francisco Mauricio Andrade Chávez

Asesora: Dra. Zaida Jesús Quiroz Cornejo

Miembros del jurado:
Dr. Cristian Luis Bayes Rodríguez
Dr. Luis Enrique Benites Sánchez
Dra. Zaida Jesús Quiroz Cornejo

Lima, Agosto 2020

Agradecimentos

Al terminar esta tesis quiero agradecer a Dios por acompañarme siempre.

A la profesora Zaida Quiroz por el liderazgo mostrado durante todo el proceso de elaboración de tesis y por haber mantenido la objetividad en especial en esta época de Covid-19.

Y a mis amigas y amigos de la maestría por su calidez en este viaje.

Resumen

La anemia es una afección causada por un bajo nivel de hemoglobina en la sangre causada principalmente por un déficit en el consumo de hierro. En el Perú, es un problema de salud pública y nutrición principalmente en niñas y niños menores de cinco años, por ello el Instituto Nacional de Estadística (INEI) realiza una prueba para determinar anemia en niñas y niños a través de la Encuesta Demográfica y de Salud Familiar (ENDES). En esta encuesta se clasifica los niveles de anemia como severa si es menor a 7,0 g/dl, moderada si está entre 7,0 y 9,9 g/dl o leve si varía entre 10,0 y 11,9 g/dl. En este contexto, en esta tesis se propone aplicar el modelo de regresión de Dirichlet para estimar la prevalencia de los niveles de anemia infantil a nivel de centros poblados en el año 2017. Se propone estimar los parámetros usando inferencia bayesiana, a través del método Halmitoniano de Monte Carlo (HMC) usando *Rstan*. El modelo propuesto también permite identificar posibles factores determinantes de la prevalencia de la anemia infantil y tiene el propósito de mejorar las políticas públicas dirigidas a la reducción de la anemia en el país.

Palabras-clave: anemia, inferencia bayesiana, método halmitoniano de Monte Carlo, regresión de Dirichlet.

Abstract

Anemia is a condition caused by a low level of hemoglobin in the blood, it is caused mainly by an iron deficiency consumption. In Perú, it is a problem of public health and nutrition in childs under five years old, for this reason the Instituto Nacional de Estadística (INEI) conducts a test to determine anemia in childs through the survey "Encuesta Demográfica y de Salud Familiar" (ENDES). In this survey, anemia levels are classified as severe (< 7.0 g/dl), moderate (7.0-9.9 g/dl) or mild (10.0-11.9 g/dl). In this context, in this thesis, it is proposed to apply the Dirichlet regression model to estimate the prevalence of levels of childhood anemia at the level of population centers in 2017. It is proposed to estimate the parameters using bayesian inference, through the Halmitonian Monte Carlo (HMC) method using *Rstan*. The model let us identify possible determining factors of the prevalence of childhood anemia. These results are important to improve public policies aimed at reducing anemia in the country.

Keywords: anemia, Bayesian inference, Halmitonian Monte Carlo method, Dirichlet regression.

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

La anemia es el problema de salud pública y nutrición más extendido en el mundo, en especial en los países, regiones y grupos poblacionales con mayor nivel de pobreza. Se estima que más de 1600 millones de personas registran algún grado de anemia que puede ser leve, moderado o severo. La principal causa de anemia es el déficit en el consumo de hierro, elemento principal para la formación de hemoglobina (Hb). La hemoglobina es una proteína rica en hierro, da a la sangre su color rojo y transporta oxígeno desde los pulmones hacia el resto del cuerpo. La concentración de Hb es el indicador más fácil de medir y más confiable para detectar la anemia. La anemia y el déficit de hierro en el organismo de un niño en crecimiento, afectan gravemente el proceso de maduración cerebral teniendo efectos negativos sobre la capacidad de aprendizaje y en el sistema inmune. En el caso de anemia en la madre durante la etapa gestacional, implica mayor riesgo de prematuridad, bajo peso al nacer, deterioro de la salud y mayor riesgo de mortalidad materna e infantil, (World Health Organization, 2008).

El 2017 en el Perú, 43.6 % de los niños entre 6 y 36 meses de edad tuvieron algún grado de anemia. De la misma forma el 28 % de las gestantes presentaron anemia. La proporción de anemia varía según la situación económica familiar, está presente en el 53 % de los niños del quintil de menores ingresos y en el 28 % de los niños del quintil de mayores ingresos. De cada 100 casos de anemia, 64 corresponden a anemia leve. Las mayores prevalencias se registran en regiones de la Sierra Sur, Central y en la Amazonía. En el período 2016 y 2017, trece de las veinticinco regiones del país incrementaron los niveles de anemia; este es el caso de Puno donde la anemia afecta al 75 % de los niños. Entre las causas indirectas de la anemia por un bajo consumo de hierro se encuentran: la falta de acceso a alimentos de calidad, malos hábitos de alimentación y nutrición, condiciones insalubres de vivienda y entorno comunitario, carencia de agua segura y alcantarillado, bajo nivel educativo, etc., (Colegio Médico del Perú, 2018).

La Organización Mundial de la Salud, grupos de expertos mundiales y el Colegio Médico del Perú recomiendan priorizar políticas públicas y sanitarias para reducir la prevalencia de anemia. Para esto e s necesario identificar las causas determinantes de la anemia en niñas y

niños menores de 5 años (de 0 a 59 meses de edad) a nivel de centros poblados con el propósito de mejorar las políticas públicas dirigidas a la reducción de la anemia en el país.

En este contexto, en esta tesis se propone estudiar las proporciones de los niveles de anemia a nivel de centros poblados del 2017 en el Perú. Estas proporciones son datos composicionales obtenidos de la Encuesta Demográfica y de Salud Familiar (2017) que requieren ser modelados de acuerdo a sus características.

Los datos composicionales son realizaciones de vectores aleatorios de sumas constantes, es decir, cualquier vector, cuyas componentes representan partes de un todo, sujeto a la restricción de que la respuesta observada está compuesta por un conjunto o vector de valores positivos que suman uno o en el caso general, una constante; por lo tanto, el espacio muestral es un simplejo, (Camargo et al., 2012).

En esta tesis se utilizará la distribución Dirichlet, que es la versión multivariada de la distribución beta para modelar datos composicionales. Los modelos de regresión Dirichlet se pueden usar para analizar un conjunto de variables que se encuentran en un intervalo acotado que suman una constante, que muestran sesgo y heterocedasticidad, sin tener que transformar los datos, (Maier, 2014).

1.2. Revisión de la literatura

A continuación se presentan algunos de los documentos que se utilizan en el desarrollo de la tesis:

Cribari-Neto (2004) propone un modelo de regresión en el que la respuesta tiene distribución beta reparametrizada con la media y la dispersión cuyo criterio de reparametrización es utilizado en la distribución Dirichlet. El modelo propuesto es útil para situaciones donde la variable de interés es continua y restringida en el intervalo $(0, 1)$ y está relacionada con otras variables a través de una estructura de regresión.

Hijazi y Jernigan (2007) introduce los modelos y métodos de regresión Dirichlet con un algoritmo eficiente para elegir los valores iniciales en la optimización numérica de la función de máxima verosimilitud y propone dos medidas de variabilidad explicada en los datos composicionales para evaluar la adecuación del ajuste de log ratio y modelos Dirichlet.

Maier (2014) proporciona una breve base teórica de las reparametrizaciones de la distribución Dirichlet y describe la implementación de los métodos de regresión Dirichlet y aplica una transformación multivariada a los parámetros de la distribución Dirichlet que llega a una formulación alternativa que tiene la ventaja de modelar el valor esperado de una observación por separado de su precisión. La metodología para la optimización del modelo está plasmada en el paquete DirichletReg en R de su autoría.

van der Merwe (2018) explora algunos enfoques existentes del modelo de regresión bayesiana de datos composicionales y luego introduce un nuevo enfoque de simulación para ajustar

dichos modelos, basado en el marco bayesiano.

Sennhenn-Reulen (2018) ofrece una breve introducción a los modelos de regresión Dirichlet y en base a una formulación desarrollada recientemente, ilustra la implementación en el marco de inferencia bayesiana.

1.3. Objetivos

El objetivo general de la tesis es aplicar el modelo de regresión Dirichlet bayesiano para estimar los niveles de prevalencia de anemia en niños en el Perú a nivel de centros poblados del año 2017. Esto incluye el estudio de las propiedades de la distribución Dirichlet y su modelo de regresión, así como su estimación desde el punto de vista bayesiano, la simulación, implementación computacional y aplicación a un conjunto de datos reales en el ámbito de medición. De manera específica:

- Estudiar la distribución Dirichlet y sus propiedades.
- Estudiar y aplicar la regresión Dirichlet para modelar los niveles de prevalencia de anemia en el Perú: no anemia, leve y moderado-severo.
- Estimar los parámetros del modelo con métodos de inferencia bayesiana aplicando un software de uso libre.
- Evaluar el desempeño del modelo propuesto a través estudios de simulación.
- Identificar en base al modelo obtenido las causas de la anemia en niños y niñas menores de 5 años (6 a 59 meses de edad) a nivel de centros poblados en el Perú en el año 2017.

1.4. Organización del Trabajo

En el Capítulo 2 se describe la distribución Dirichlet y se señala la importancia de la re-parametrización propuesta. El capítulo 3 describe la estructura del modelo de regresión de regresión Dirichlet, presenta su función de probabilidad y se definen las distribuciones a priori de los parámetros que permite calcular la distribución a posteriori. En el capítulo 4 se presentan los resultados obtenidos de un estudio de simulación para evaluar la bondad de ajuste del modelo de regresión Dirichlet bayesiano. El Capítulo 5 incluye los resultados de la aplicación del modelo de regresión Dirichlet bayesiano a datos de niveles de anemia de niños menores de 5 años en el Perú. Los comentarios finales se presentan en el Capítulo 6.

Capítulo 2

Marco teórico

Este capítulo presenta el sustento teórico para desarrollar el modelo de regresión Dirichlet bayesiano. Un modelo de regresión trata del estudio de la dependencia de una variable dependiente o de respuesta $\boldsymbol{Y}$ respecto de una o más variables explicativas $\boldsymbol{X}$ (Gujarati y Porter, 2010). En nuestro caso la variable de respuesta consiste en datos composicionales que siguen una distribución Dirichlet. A continuación se estudian la definición de datos composicionales, las dos reparametrizaciones de la distribución Dirichlet a las que se conocen como común y alternativa, además de la inferencia bayesiana que se requiere para el desarrollo de la tesis, los métodos computacionales para extraer los parámetros del modelo de regresión y los criterios para su evaluar y seleccionar el mejor modelo. Hijazi (2003) y Hijazi y Jernigan (2007) exploraron los modelos de regresión Dirichlet; Hijazi (2003) utilizó el enfoque común en el que los parámetros $\boldsymbol{\alpha}$ de la distribución Dirichlet se modelan directamente mediante covariables con un enlace logarítmico. La reparametrización alternativa es una extensión del enfoque implementado para modelar proporciones con la regresión beta (Cribari-Neto, 2004) y (Cribari y Zeileis, 2010).

2.1. Datos composicionales

Los datos composicionales consisten en vectores de proporciones que surgen cuando se clasifican i-ésimos sujetos en categorías disjuntas y se registran las frecuencias relativas resultantes, o si se divide una medida completa en contribuciones porcentuales de sus diversas partes. Bajo la restricción de sumar 1 (Hijazi y Jernigan, 2007).

Aitchison (1986) manifiesta que en una composición en términos más generales, se usan los números enteros $1, 2, \cdots, C$.

Cada i-ésima fila de la matriz de datos corresponde a un sujeto, a una única unidad experimental o de observación. Cada columna c corresponde a una parte específica de cada composición, ó sea $y_{i,1}, y_{i,2}, \cdots, y_{i.C}$.

Los componentes de cualquier composición de la parte $\boldsymbol{c} = (y_{i,1} \cdots, y_{i,C})$ deben satisfacer la restricción de que cada componente no sea negativo, es decir:

$$y_{i,1} \geq 0, \cdots y_{i,C} \geq 0.$$

Según lo mencionado en los párrafos anteriores, a continuación se presenta la estructura de un conjunto de datos composicionales:

$$\boldsymbol{y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,C} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,C} \\ \vdots & \vdots & \cdots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,C} \end{bmatrix},$$

donde $y_{i,1} + \cdots + y_{i,C} = 1$.

2.2. Reparametrización común de la distribución Dirichlet

En el caso de datos composicionales el conjunto de resultados del experimento aleatorio, es decir el espacio muestral, es una parte restringida del espacio real denominada simplejo y las operaciones en sus elementos juegan un papel fundamental en los problemas de los datos de composición, ya sea en su totalidad o como una parte importante del espacio muestral, (Aitchison, 1986). La distribución Dirichlet es un candidato natural para analizar este tipo de datos ya que su soporte es el simplejo (Tsagris y Stewart, 2017).

Para definir la función de densidad de probabilidad (f.d.p.) de la distribución Dirichlet se define un simplejo cerrado de n dimensiones en $\mathbb{R}^C$ y un simplejo abierto de $(C-1)$ dimensiones en $\mathbb{R}^{(C-1)}$ dado respectivamente por:

$$\begin{aligned} \mathbb{T}_C &= \left\{ (y_1, ..., y_C)^\top : y_c > 0, 1 \leq c \leq C, \sum_{c=1}^{C} y_c = 1 \right\}, \\ \mathbb{V}_{C-1} &= \left\{ (y_1, ..., y_{C-1})^\top : y_c > 0, 1 \leq c \leq C-1, \sum_{c=1}^{C-1} y_c < 1 \right\}. \end{aligned}$$

Luego, un vector aleatorio $\mathbf{Y} = (Y_1, ..., Y_C)^\top \in \mathbb{T}_c$ tiene distribución Dirichlet si la f.d.p. de Y es dada por:

$$f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{c=1}^{C} \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} y_c^{\alpha_c - 1}, \quad \mathbf{y} \in \mathbb{T}_c, \tag{2.1}$$

donde $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_C)^\top$ es un vector de parámetros positivos ya que $\alpha_c > 0$ para todo

$c = 1, \cdots, C$ y $\Gamma(\cdot)$ representa la función gamma: $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}du$. Podemos denotar la distribución por: $\mathbf{Y} \sim \text{Dir}(\boldsymbol{\alpha})$ en $\mathbb{T}_c$ o $\mathbf{Y}_{-c} \sim \text{Dir}(\alpha_1, ..., \alpha_{C-1,\alpha_C})$ en $\mathbb{V}_{C-1}$, (Wang et al., 2011). Cuando $c = 2$ la distribución $\text{Dir}(\alpha_1, \alpha_2)$ se reduce a una distribución $\text{Beta}(\alpha_1, \alpha_2)$. En esta tesis trabajaremos con $c = 3$; como se muestra en la Figura 2.1.

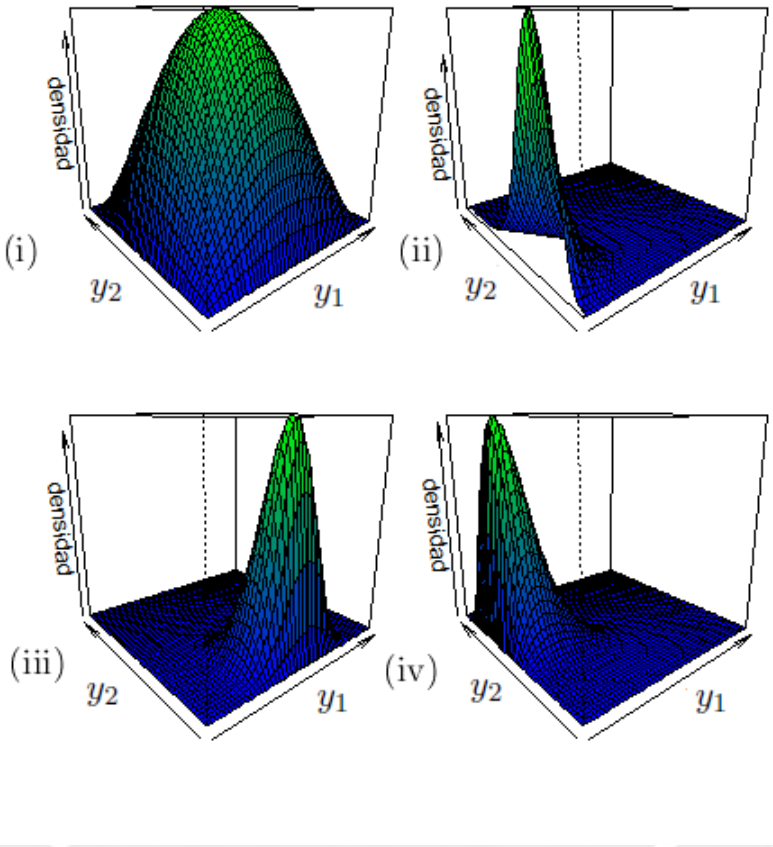


Figura 2.1: Función de probabilidad de $\mathbf{Y} = (Y_1, Y_2)^\top \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ con: (i) $\alpha_1 = \alpha_2 = \alpha_3 = 2$, (ii) $\alpha_1 = 1, \alpha_2 = 5, \alpha_3 = 10$ (iii) $\alpha_1 = 10, \alpha_2 = 3, \alpha_3 = 8$ y (iv) $\alpha_1 = 2, \alpha_2 = 10, \alpha_3 = 4$.

Sea el vector aleatorio $\mathbf{Y} = (Y_1, ..., Y_C)^\top \sim \text{Dir}(\boldsymbol{\alpha})$ con f.d.p. dada en (2.1), cuyas componentes cumplen las siguientes propiedades:

$$\begin{aligned} E(Y_c) &= \frac{\alpha_c}{\alpha_+}, \quad c = 1, ..., C. \\ Var(Y_c) &= \frac{\alpha_c(\alpha_+ - \alpha_c)}{\alpha_+^2(1+\alpha_+)}, \quad c = 1, ..., C. \\ Cov(Y_c, Y_j) &= \frac{-\alpha_c\alpha_j}{\alpha_+^2(1+\alpha_+)}, \quad c \neq j \quad c, j = 1, ..., C. \end{aligned}$$

donde $y_C = 1 - \sum_{c=1}^{C-1} y_c$ y $\alpha_+ \hat{=} \sum_{c=1}^{C} \alpha_c$.

2.3. Reparametrización alternativa de la distribución Dirichlet

Se define un conjunto de parámetros $\mu_c = E(Y_c)$ para los diferentes valores esperados de la variable Y_c y θ para modelar la precisión. Se llega a esta parametrización asumiendo que $\mu_c = \alpha_c/\alpha_+$ y $\theta = \alpha_+$, (Maier, 2014). Esta parametrización se denota por $\mathbf{Y} \sim D(\boldsymbol{\mu}, \theta)$ por lo que se puede introducir las siguientes propiedades:

$$\begin{aligned} E(Y_c) &= \mu_c, \\ Var(Y_c) &= \frac{\mu_c(1-\mu_c)}{(\theta+1)}, \\ Cov(Y_c, Y_j) &= \frac{-\mu_c\mu_j}{(\theta+1)}. \end{aligned}$$

Por lo tanto la distribución Dirichlet alternativa se representa de la siguiente forma:

$$f(\mathbf{y}|\boldsymbol{\mu}, \theta) = \frac{\Gamma\left(\sum_{c=1}^{C} \mu_c\theta\right)}{\prod_{c=1}^{C} \Gamma\left(\mu_c\theta\right)} \prod_{c=1}^{C} y_c^{(\mu_c\theta-1)}, \quad \mathbf{y} \in \mathbb{T}_n,$$

donde $\sum_{c=1}^{C} \mu_c = 1$, $\mu_c \in (0,1)$ y $\theta > 0$. Esta distribución se utilizará en el modelo de regresión Dirichlet.

2.4. Inferencia bayesiana

Thomas Bayes y Pierre-Simon Laplace, en el siglo XVIII, fueron los primeros pensadores en considerar el azar y la aleatoriedad de una manera cuantitativa y científica a través de una relación que se conoce como teorema de Bayes, definido para un parámetro o vector de parámetros $\boldsymbol{\beta}$ y una variable o vector aleatorio $\mathbf{Y}$ como sigue:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\boldsymbol{y})},$$

donde $p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{\beta}) d\boldsymbol{\beta} = \int p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}$, que generalmente es difícil de calcular. Dado que $p(\boldsymbol{y})$ no depende de $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}|\mathbf{y})$ puede escribirse de forma más compacta,

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}). \tag{2.2}$$

La proporcionalidad $\propto$ indica que el factor $1/p(\mathbf{y})$ es una constante normalizadora y puede ignorarse, (Gelman et al., 2003).

Una de las aplicaciones del teorema de Bayes es la inferencia bayesiana que tiene por meta principal calcular la distribución a posteriori $p(\boldsymbol{\beta}|\boldsymbol{y})$ de un vector de parámetros desconocidos $\boldsymbol{\beta} = (\beta_1, ..., \beta_Q)^\top$ después de tomar en cuenta datos nuevos representados por la verosimilitud de los datos $p(\boldsymbol{y}|\boldsymbol{\beta})$, es decir actualizando la probabilidad a priori $p(\boldsymbol{\beta})$ en base al conocimiento actual antes de un experimento. La distribución a posteriori contiene toda la información actual sobre el parámetro $\boldsymbol{\beta}$ que permite obtener: intervalos de credibilidad, estimaciones puntuales de los parámetros, realizar inferencia de predicción para datos futuros y evaluaciones probabilísticas para su hipótesis, (Gelman et al., 2003).

2.4.1. Selección de distribuciones a priori

La selección de distribuciones a priori $p(\boldsymbol{\beta})$ es un aspecto crucial en inferencia bayesiana. Si se tiene un conocimiento basado en datos históricos o una conjetura de expertos de cuales deberían ser los parámetros del modelo, esa información se puede incluir como una distribución a priori subjetiva, pero si el modelo se vuelve más complicado resulta difícil identificar que distribuciones a priori utilizar para cada parámetro desconocido. Si bien el subjetivismo se ha convertido en la base filosófica dominante de la inferencia bayesiana, cuando las distribuciones a priori no tienen una base poblacional, pueden ser difíciles de construir (Gelman et al., 2003). En la práctica, la mayoría de análisis bayesianos se llevan a cabo con distribuciones a priori vagas, fiduciarias, difusas o no informativas, (Kass y Wasserman, 1994). Se dicen no informativas ya que su función es dejar que los datos hablen por sí mismos al desempeñar un papel mínimo en la distribución a posteriori, de modo que las inferencias no se vean afectadas por la información externa a los datos, (Gelman, 2006). Se puede trabajar con una distribución normal donde a medida que aumenta la cantidad de observaciones en el conjunto de datos, la verosimilitud diluye el impacto de las distribuciones a priori.

Gelman (2006) considera cualquier distribución a priori no informativa o débilmente informativa como provisional hasta que el modelo se haya ajustado, se debe mirar la distribución a posteriori y evaluar si tiene sentido. Si la distribución a posteriori no tiene sentido, implica que se dispone de conocimiento previo adicional que no se ha incluido en el modelo, y que contradice los supuestos de la distribución a priori que se ha utilizado. Entonces es apropiado modificar la distribución a priori para ser más consistente con el conocimiento externo.

2.4.2. Distribución conjunta a posteriori

Según Gelman et al. (2003) la distribución conjunta a posteriori $p(\boldsymbol{\beta}|\boldsymbol{y})$ que contiene toda la información actual sobre los parámetros desconocidos $\boldsymbol{\beta}$ debe ser propia, es decir:

$$\int_{\boldsymbol{\beta}} p(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta} = 1,$$

Para hacer inferencia bayesiana se requiere evaluar analíticamente las integrales requeridas,

por lo cual deben darse situaciones conjugadas en las que la distribución a priori $p(\boldsymbol{\beta})$ y distribución a posteriori $p(\boldsymbol{\beta}|\boldsymbol{y})$ pertenezcan a la misma familia de distribuciones. Las situaciones conjugadas, que permiten extraer directamente muestras aleatorias los parámetros de la distribución a posteriori, no están disponibles más allá del modelo lineal normal (Wakefield, 2013).

Las distribuciones marginales a posteriori $p(\boldsymbol{\beta}|\boldsymbol{y})$ contienen información completa de los parámetros de interés; se obtienen al integrar $p(\boldsymbol{\beta}|\boldsymbol{y})$ sobre los parámetros desconocidos que no son de interés inmediato; $p(\boldsymbol{\beta}|\boldsymbol{y})$ es típicamente multivariada. La distribución marginal univariada para β_i es

$$p(\beta_i|\boldsymbol{y}) \propto \int p(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta}_{-i}.$$

donde $\boldsymbol{\beta}_{-i}$ es el vector $\boldsymbol{\beta}$ excluyendo β_i, es decir, $\boldsymbol{\beta}_{-i} = [\beta_1, \cdots, \beta_{-i}, \beta_{i+1}, \cdots, \beta_p]^\top$ (Wakefield, 2013).

2.4.3. Distribución predictiva a posteriori

Otra de las metas de la inferencia bayesiana es la predicción para lo cual se utiliza la distribución a posteriori predictiva $p(y_{n+1}|\boldsymbol{y})$ (Gelman et al., 2003). Para un conjunto de datos $\boldsymbol{y}$, y una nueva observación y_{n+1} deseamos encontrar la distribución condicional de Y_{n+1} dado $\boldsymbol{Y} = y$. Si Y_{n+1} e $\boldsymbol{Y}$ son condicionalmente independientes dado el parámetro continuo $\boldsymbol{\beta}$; la distribución predictiva a posteriori estaría dada por:

$$\begin{aligned} p(y_{n+1}|\boldsymbol{y}) &= \int p(y_{n+1}, \boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta} \\ &= \int p(y_{n+1}|\boldsymbol{\beta}, \boldsymbol{y})p(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta} \\ &= \int p(y_{n+1}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta}. \end{aligned}$$

La segunda y tercera línea muestran la distribución a posteriori predictiva como un promedio de las predicciones condicionales sobre la distribución a posteriori de $\boldsymbol{\beta}$.

2.4.4. Intervalos de credibilidad

Un intervalo de credibilidad de un parámetro $\boldsymbol{\beta}$ de tamaño $1-\alpha$ es un intervalo $[a, b]$ tal que

$$\begin{aligned} P(a \leq \boldsymbol{\beta} \leq b|\boldsymbol{y}) &= 1-\alpha \\ \int_a^b p(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta} &= 1-\alpha. \end{aligned}$$

La perspectiva frecuentista considera a $\boldsymbol{\beta}$ como fijo y hacen uso del concepto de probabilidad antes de observar los datos. Para cualquier valor fijo de $\boldsymbol{\beta}$, un intervalo de confianza frecuentista contendrá el verdadero parámetro $\boldsymbol{\beta}$ con cierta probabilidad, por ejemplo, 0.95. Por otro lado la perspectiva bayesiana considera a $\boldsymbol{\beta}$ como una variable aleatoria y utilizan el concepto de probabilidad después de observar los datos. Para algún conjunto particular de datos $\mathbf{Y} = \boldsymbol{y}$, la variable aleatoria $\boldsymbol{\beta}$ se encuentra en un intervalo de credibilidad con cierta probabilidad.

2.5. Inferencia bayesiana usando MCMC

Los métodos Monte Carlo vía Cadenas de Markov (MCMC) permiten estimar computacionalmente la distribución a posteriori en caso de ser una expresión compleja y no tener forma analítica conocida (Gelman et al., 2003). Se extraen muestras de $\boldsymbol{\beta}$ provenientes de la distribución aproximada a la distribución a posteriori $p(\boldsymbol{\beta}|\boldsymbol{y})$.

A continuación se analizan brevemente las simulaciones de Monte Carlo, las cadenas de Markov y los métodos MCMC:

- Las simulaciones de Monte Carlo se utilizan para estimar el área de distribuciones con formas complicadas, al generar repetidamente muestras aleatorias de un parámetro y proporcionar una aproximación de dicho parámetro ya que calcularlo directamente es computacionalmente costoso. El proceso de simulación requiere que se sigan los siguientes pasos:

  1. Generar M conjuntos de datos bajo la condición de interés.
  2. Calcular el valor numérico de los estimadores:

  $$(T_1, T_2, \cdots, T_M)$$

  3. Si M es suficientemente grande, las estimaciones de $(T_1, T_2, \cdots, T_M)$ deberían ser una adecuada aproximación de los parámetros.

- Una cadena de Markov es un proceso estocástico que describe una secuencia de posibles eventos en los que la probabilidad de cada evento depende solo del estado alcanzado en el evento anterior, de acuerdo con un conjunto fijo de probabilidades. Una característica importante de las cadenas de Markov es que no tienen memoria; todo lo que se necesita para predecir el próximo evento está disponible en el estado actual, y no se obtiene información nueva al conocer el historial de eventos.

Los métodos MCMC generan un conjunto de cadenas similares a las cadenas de Markov, a partir de un conjunto de puntos elegidos arbitrariamente y separados entre sí. Estas cadenas son procesos estocásticos de *random walkers* que realizan paseos aleatorios de acuerdo con

un algoritmo que busca valores que contribuyan a la integral que se desea aproximar numéricamente, asignándoles mayor probabilidad, para luego pasar a la siguiente ubicación hasta encontrar la distribución estacionaria. Las cadenas de Markov tienen como distribución estacionaria a la distribución a posteriori. El estimador MCMC tiene una varianza baja debido a que se extraen muestras de regiones importantes del espacio de parámetros (Gelman et al., 2003).

2.5.1. Hamiltoniano Monte Carlo (HMC)

HMC es un método MCMC que evita el comportamiento de paseos aleatorios y sensibilidad a los parámetros correlacionados, que afectan a muchos métodos MCMC al tomar una serie de pasos informados por la gradiente de primer orden.

En general, la energía total de un sistema cerrado y conservador es dado por la función halmitoniana $H(\theta, \phi) = U(\theta) + C(\phi)$ donde $U(\theta)$ es la energía potencial y $C(\phi)$ es la energía cinética del sistema. En estadística el Monte Carlo halmitoniano (HMC) toma los parámetros θ como si colectivamente denotasen la posición de la partícula en algún espacio con momento (cantidad de movimiento) ϕ. Así se asume que la energía potencial del sistema halmitoniano es $U(\theta) = -\log(p(\theta \mid Y))$, donde $p(\theta \mid Y)$ es la fdp de la distribución a posteriori, mientras que la energía cinética es $C(\phi) = p(\phi) = \phi^T M^{-1} \phi$, que es en esencia el núcleo de una distribución normal con covarianza M. Se asume que la fdp aposteriori $p(\theta, \phi \mid Y) \propto \exp(-H(\theta, \phi)) = \exp[\log(p(\theta \mid Y)) - \phi^T M^{-1} \phi)]$. Luego el algoritmo consiste en simular θ y ϕ a partir de $p(\theta, \phi \mid Y)$ a pesar de que el interés está solo en estimar θ. Así el vector ϕ es una variable auxiliar introducida en el algoritmo para que la cadena de Markov se mueva más rápido en todo el espacio paramétrico.

El algoritmo HMC combina MCMC y métodos de simulación determinística. El algoritmo consiste en repetir los pasos siguientes L veces:

1. En la iteración t, actualiza ϕ a partir de $\phi \sim N(0, M)$.
2. Actualizar $\theta^{(t-1)}$ y $\phi^{(t-1)}$ simulatáneamente a través de L pasos de "leapfrog", tal que cada paso es escalado po ϵ, un valor pequeño que sirve para el tuning del algoritmo.

   a) Use el gradiente $\frac{d \log(p(\theta, \phi \mid Y))}{d\theta}$ para actualizar ϕ,

   $$\phi^\star \leftarrow \phi^{(t-1)} + \frac{1}{2}\epsilon \frac{d \log(p(\theta^{(t-1)}, \phi^{(t-1)} \mid Y))}{d\theta}.$$

   b) Use el vector momento $\phi^\star$ para actualizar la posición del vector θ:

   $$\theta^\star \leftarrow \theta^{(t-1)} + \epsilon M^{-1} \phi^\star.$$

   c) Actualizar ϕ:

   $$\phi^\star \leftarrow \phi^\star + \frac{1}{2}\epsilon \frac{d \log(p(\theta^\star, \phi^\star \mid Y))}{d\theta}.$$

3. Calcule

$$r = \frac{p(\theta^{\star} \mid Y)p(\phi^{\star})}{p(\theta^{(t-1)} \mid Y)p(\phi^{(t-1)})},$$

donde $\theta^{\star}$ y $\phi^{\star}$ son los valores iniciales, mientras que $\theta^{\star}$ y $\phi^{\star}$ son los valores candidatos.

4. Actualice $\theta^{(t)} = \theta^{\star}$ con probabilidad $min(r, 1)$ o no actualice $\theta^{(t)} = \theta^{(t-1)}$ caso contrario.

5. Repetir los pasos 1 a 4 hasta llegar a la convergencia de las cadenas.

Se debe tener en cuenta que ϵ no debe ser demasiado grande o sino el integrador realizará muchos pasos pequeños lo que incrementaría el costo computacional. Se debe evitar que L sea demasiado pequeño, el algoritmo exhibe un comportamiento de oscilación aleatorio no deseado (Hoffman y Gelman, 2011).

2.5.2. No-U-Turn Sampler (NUTS)

En esta tesis se utilizará el algoritmo No-U-Turn Sampler (NUTS), una variante de HMC (Hoffman y Gelman, 2011). NUTS simula el movimiento de la partícula ficticia que representa los valores de los parámetros que no se detiene hasta que hace un giro en U (Stan Development Team, 2019).

NUTS es más eficiente explorando la distribución a posteriori, se elimina la necesidad de establecer una serie de pasos L; ya que utiliza un algoritmo recursivo para crear un conjunto de valores candidatos probables que abarca una amplia franja de la distribución objetivo; selecciona automáticamente un número apropiado de pasos en cada iteración para permitir que los valores candidatos crucen la distribución a posteriori fácilmente.

Según Stan Development Team (2019), NUTS genera una propuesta comenzando en una posición inicial determinada por los parámetros seleccionados en la última iteración. Luego genera un vector aleatorio normal estándar independiente. Expande el sistema inicial para adelante y para atrás al mismo tiempo en la forma un árbol binario balanceado. En cada iteración del algoritmo NUTS, la profundidad del árbol aumenta en uno, duplicando el número de saltos y el tiempo de cálculo. El algoritmo termina y vuelve sobre sus pasos en una de dos formas; si se cumple el criterio NUTS (es decir, un cambio de sentido en el espacio euclidiano en un subárbol) para un nuevo subárbol o el árbol completado, ó la profundidad del árbol completo alcanza la profundidad máxima permitida. NUTS no requiere la intervención del usuario ni ejecuciones de *tunning*.

2.5.3. Diagnósticos de convergencia y autocorrelación

El uso de los métodos MCMC puede presentar ciertas dificultades que requieren que se tenga en cuenta lo siguiente:

1. Es necesario determinar cuándo una cadena de Markov inicializada aleatoriamente ha convergido a su distribución de equilibrio.

2. Los valores de una cadena de Markov pueden estar correlacionados. Mientras que las muestras aleatorias del integrando utilizado en una integración convencional de Monte Carlo son estadísticamente independientes, las utilizadas en los métodos MCMC están autocorrelacionadas y, por lo tanto, el teorema del límite central sobre el error de estimación no se puede aplicar.

3. Antes de hacer inferencia, por lo general se eliminan un número inicial de iteraciones, este procedimiento se conoce como "burn-in", la razón es que los resúmenes inferenciales no deben estar influenciados por puntos iniciales que podrían estar lejos de la masa principal de la distribución a posteriori (Gelman et al., 2003).

2.6. Evaluación y selección de modelos

Se presentan varios métodos de evaluación y selección de modelos con el objetivo de seleccionar el mejor, (Amaral et al., 2019).

2.6.1. Ajuste del modelo

Se busca cuantificar las discrepancias entre el modelo y el conjunto de datos disponibles.

Verosimilitud pseudo-marginal (LPML).- Estas medidas pueden construirse comparando las características de la distribución predictiva bajo el modelo condicional a los datos observados con otros datos también observados. Esto es posible en el contexto de la validación cruzada, si la muestra original es grande, es posible dividirla en dos partes; una parte como muestra de entrenamiento $\boldsymbol{Y} = (Y_1, ..., Y_k)^\top$ para generar una distribución predictiva a posteriori $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$, y otra parte como muestra de evaluación $\tilde{\boldsymbol{Y}} = (Y_{k+1}, ..., Y_n)^\top$, independiente de $\boldsymbol{Y}$; para determinar la validez del modelo mediante la inspección de la distribución predictiva a posteriori correspondiente $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$.

Otra alternativa, es usar el método de Jackknife que consiste en repetir la validación cruzada n veces, siempre dejando por fuera una observación de la muestra de entrenamiento. Esa observación juega el papel de validación de la muestra de evaluación.

Sea $\boldsymbol{Y}_{(-i)} = (Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_k)^\top$, el vector de todas las variables aleatorias excepto Y_i. Se puede obtener las distribuciones predictivas condicionales $p(\tilde{y_i}|y_{(-i)})$,

$$p(\tilde{y_i}|y_{(-i)}) = \int p(\tilde{y_i}|\beta, y_{(-i)})p(\beta|y_{-i})d\beta,$$

El valor de $p(\tilde{y_i}|y_{(-i)})$ evaluado en y_i se conoce como conditional predictive ordinate (CPO). Los valores son evidencia de la probabilidad de cada observación dado el resto de las observaciones y, por lo tanto, los valores bajos de CPO corresponden a observaciones mal ajustadas. En este sentido cuanto mayor es la suma de los valores logarítmicos de CPO, también conocida como pseudo-verosimilitud marginal (LPML), mejor es el modelo. LPLM se define como:

$$\text{LPML} = \sum_{i=1}^{n} \ln \ CPO_i = \ln \ \prod_{i=1}^{n} p(y_i|y_{(-i)}).$$

2.6.2. Capacidad predictiva

Los criterios de información se utilizan para la selección de modelos entre un conjunto finito de modelos. Al ajustar modelos, es posible aumentar la verosimilitud agregando parámetros lo que puede resultar en un sobreajuste. BIC, DIC y WAIC intentan resolver este problema introduciendo un término de penalización por el número de parámetros del modelo. Cuanto menor sea el valor de estos criterios, mejor será el modelo y reflejará el uso predictivo del mismo.

Criterio de información bayesiana (BIC): este criterio de selección de modelos que utiliza la aproximación de una muestra de tamaño n grande de la distribución marginal de los datos $p(\boldsymbol{y}) = E_{p(\boldsymbol{\beta})}[p(\boldsymbol{y}|\boldsymbol{\beta})]$ se define como:

$$\text{BIC}^{\text{CL}} = -2 \ln p(\boldsymbol{y}|\hat{\boldsymbol{\beta}}) + p_B \ \ln(n),$$

donde $\ln p(\boldsymbol{y}|\hat{\boldsymbol{\beta}}) \simeq \ln p(\boldsymbol{y}) + (p_B/2)\ln(n)$, $\hat{\boldsymbol{\beta}}$ es la estimación de $\boldsymbol{\beta}$ y p_B es el número de parámetros del modelo.

Para evitar la maximización basada en la simulación, se utiliza una modificación de este criterio utilizando la media a posteriori de la probabilidad de registro en lugar de $\ln p(\boldsymbol{y}|\hat{\boldsymbol{\beta}})$. Esta versión modificada del BIC se define como

$$\text{BIC} = -2E_{\beta|\boldsymbol{y}}[\ln p(\boldsymbol{y}|\beta)] + p_B \ \ln(n).$$

Criterio de información de desviación (DIC): se utiliza para comparar la bondad de ajuste de los modelos analizados debido a que la distribución a priori y el tipo de estructura del modelo tienden a afectar el grado de sobreajuste. El DIC se define como:

$$\text{DIC} = D(\overline{\boldsymbol{\beta}}) + 2p_D = \overline{D(\boldsymbol{\beta})} + p_D = 2\overline{D(\beta)} - D(\overline{\boldsymbol{\beta}}),$$

donde la aproximación $\overline{\boldsymbol{\beta}}$ es una estimación bayesiana que sustituye la estimación de máxima verosimilitud de $\hat{\boldsymbol{\beta}}$; $\overline{\boldsymbol{\beta}}$ suele ser la media a posteriori de $\boldsymbol{\beta}$. $D(\boldsymbol{\beta})$ es la devianza que se define como el doble logaritmo del estadístico de verosimilitud $D(\boldsymbol{\beta}) = -2loglikelihood$. $p_D = \overline{D(\boldsymbol{\beta})} - D(\overline{\boldsymbol{\beta}})$ es una medida de la complejidad del modelo, la medida correspondiente de precisión predictiva que define el número efectivo de parámetros. Cuanto mayor sea el número efectivo de parámetros, más fácil será para el modelo ajustar los datos, por lo tanto, la desviación debe ser penalizada. $\overline{D(\boldsymbol{\beta})} = E_{\boldsymbol{\beta}|\boldsymbol{y}}[D(\boldsymbol{\beta})]$, la esperanza a posteriori de $D(\boldsymbol{\beta})$.

Criterio de información ampliamente aplicable (WAIC).- en este criterio la medida de precisión predictiva dentro de la muestra asociada con el WAIC no implica aproximaciones de complemento como los criterios discutidos anteriormente. Se evalúa con una corrección por exceso de ajuste p_w conocida tambié como término de penalización o como dimensionalidad del modelo efectivo:

$$\tilde{A}_{\text{WAIC}} = \sum_{i=1}^{n} \ln E_{\boldsymbol{\beta}|\boldsymbol{y}}[p(\boldsymbol{y}_i|\boldsymbol{\beta})] - p_w.$$

p_w que implica una expresión en términos de los datos individuales. Una de las propuestas para determinar p_w es similar a la utilizada en DIC y se expresa como:

$$p_{w_1} = -2\sum_{i=1}^{n} E_{\boldsymbol{\beta}|\boldsymbol{y}}[\ln[p(\boldsymbol{y}_i|\boldsymbol{\beta})]] - \ln E_{\boldsymbol{\beta}|\boldsymbol{y}}[ln[p(\boldsymbol{y}_i|\boldsymbol{\beta})]],$$

la otra propuesta para p_w basada en la varianza a posteriori de $ln[f(\boldsymbol{y}_i|\boldsymbol{\beta})]$ se define como:

$$p_{w_2} = \sum_{i=1}^{n} Var_{\boldsymbol{\beta}|\boldsymbol{y}}[\ln[p(\boldsymbol{y}_i|\boldsymbol{\beta})]]$$

El resultado de realizar una transformación a $\tilde{A}_{\text{WAIC}}$ facililta que el criterio esté la misma escala que los anteriores y se da por:

$$\text{WAIC} = -2\sum_{i=1}^{n} \ln E_{\boldsymbol{\beta}|\boldsymbol{y}}[p(\boldsymbol{y}_i|\boldsymbol{\beta})] + 2p_w.$$

Capítulo 3

Modelo de regresión Dirichlet

En este capítulo se presenta la definición del modelo de regresión Dirichlet y sus parámetros, donde se establece la relación que existe entre la variable de respuesta o dependiente $\mathbf{Y}$ y las variables predictoras o independientes $\mathbf{X}$. Más adelante se explican las prioris escogidas para la distribución a posteriori.

3.1. Parametrización común del modelo de regresión Dirichlet

Con fines prácticos se asume $\mathbf{Y} = (Y_{1\bullet}, Y_{2\bullet}, \cdots, Y_{n\bullet})^\top$ una variable de respuesta multivariada tal que:

$$\mathbf{Y} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \cdots & Y_{1,C} \\ Y_{2,1} & Y_{2,2} & \cdots & Y_{2,C} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{n,1} & Y_{n,2} & \cdots & Y_{n,C} \end{bmatrix} = \begin{bmatrix} Y_{1\bullet} \\ Y_{2\bullet} \\ \vdots \\ Y_{n\bullet} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_C \end{bmatrix},$$

entonces $Y_{i\bullet} = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,C}) \in \mathbb{T}_C, i = 1, 2, \cdots, n$, donde n indica el número de sujetos del estudio, C representa el número de categorías que asume la respuesta $Y_{i\bullet}$ e $\boldsymbol{Y}_c$ es el vector de la variable de respuesta con la categoría c de los n sujetos de estudio.

Sea la matriz de covariables asociadas a la variable de respuesta multivariada $\mathbf{Y}$ dada por $\mathbf{X} = (x_{1\bullet}, x_{2\bullet}, \cdots, x_{n\bullet})^\top$ donde $x_{i\bullet} \in \mathbb{R}^Q$, tal que

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,Q} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,Q} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,Q} \end{bmatrix} = \begin{bmatrix} x_{1\bullet} \\ x_{2\bullet} \\ \vdots \\ x_{n\bullet} \end{bmatrix}.$$

Asumiendo que en $Y_{i\bullet} \sim Dir(\alpha_1(x_{i\bullet}), \cdots, \alpha_C(x_{i\bullet}))$ para $i = 1, \cdots, n$, se modela cada parámetro $\alpha_c(x_{i\bullet})$, $c = 1, \cdots, C$, como una función lineal de las variables explicativas, donde

$$\begin{aligned} \alpha_c(x_{i\bullet}) &= x_{i,1}\beta_{1,c} + x_{i,2}\beta_{2,c} +, \cdots, +x_{i,Q}\beta_{Q,c} \\ &= x_{i\bullet}\beta_{\bullet c} \quad \text{para } c = 1, \cdots, C. \end{aligned}$$

Así la matriz de coeficientes de regresión está definida por:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,C} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,C} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{Q,1} & \beta_{Q,2} & \cdots & \beta_{Q,C} \end{bmatrix} = \begin{bmatrix} \beta_{\bullet 1} & \beta_{\bullet 2} & \cdots & \beta_{\bullet C} \end{bmatrix}.$$

Por lo tanto, los parámetros a estimar, son los coeficientes de regresión sujetos a la restricción $\alpha_c(x_i) > 0 \quad \forall i = 1, \cdots, n, c = 1, \cdots, C$.

La desventaja de este enfoque es que no hay una interpretación clara de los coeficientes de regresión, en función de la media de la observación $Y_{i,c}$. Para mayor información vea Camargo et al. (2012).

3.2. Parametrización alternativa del modelo de regresión Dirichlet

Se asume que $\boldsymbol{\mu}_i = E(Y_{i\bullet})$, se tiene,

$$Y_{i\bullet} \sim Dir(\boldsymbol{\mu}_i, \theta) \quad \text{para } i = 1, 2, \cdots, n,$$

donde

$$\boldsymbol{\mu}_i = (\mu_{i,1}, \mu_{i,2}, \cdots, \mu_{i,C}), \tag{3.1}$$

$$\mu_{ic} = \alpha_c(x_{i\bullet})/\alpha_{i+}, \quad \text{donde} \quad \alpha_{i+} = \sum_{c=1}^{C} \alpha_c(x_{i\bullet}) = \theta, \tag{3.2}$$

y θ es el parámetro de precisión.

Tanto $\boldsymbol{\mu}_i$ como el parámetro de precisión pueden asociarse a coeficientes de regresión a través de funciones de enlace $g_\mu(\cdot)$ y $g_\theta(\cdot)$,

en particular, se asume que

$$\begin{aligned} g_\mu(\mu_{i,c}) &= x_{i,1}\beta_{1,c} + x_{i,2}\beta_{2,c}+, \cdots, +x_{i,Q}\beta_{Q,c}, \\ g_\mu(\mu_{i,c}) &= \mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet c}, \quad c = 1, 2, \cdots, C \\ g_\theta(\theta_i) &= \gamma. \end{aligned}$$

La elección natural para modelar la media de $g_\mu(\cdot)$ debe considerar valores entre $(0, 1)$ es el enlace logit similar al de una regresión multinomial, $\mathbf{x}_i$ es la variable explicativa, $\boldsymbol{\beta}$ y γ son los coeficientes de regresión. Luego:

$$\text{logit}(\mu_{i,c}) = \mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet c} \tag{3.3}$$

$$\mu_{i,c} = \frac{\exp(\mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet c})}{\sum_{d=1}^{C}\exp(\mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet d})}, \quad c = 2, \cdots, C \tag{3.4}$$

$$\mu_{i,1} = \frac{\exp(\mathbf{x}_{i\bullet}0)}{\sum_{d=1}^{C}\exp(\mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet d})} = \frac{1}{\sum_{d=1}^{C}\exp(\mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet d})}, \tag{3.5}$$

donde debido a que tenemos c categorías para las cuales las medias siempre deben sumar 1, se emplea una estrategia logit multinomial, donde los coeficientes de regresión de una categoría $c = 1$ son tomadas como referencia, $\beta_{\bullet,1} = \mathbf{0}$.

Para g_{θ_i} se define una función de enlace logarítmica en la que $\ln(\theta) = \gamma$.

Finalmente la función de verosimilitud es definida por:

$$L(\boldsymbol{\beta}, \gamma; \mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{y}_{i\bullet}|(\boldsymbol{\beta}, \gamma)).$$

3.3. Estimación de parámetros de la regresión Dirichlet por máxima verosimilitud

Hijazi y Jernigan (2007) propusieron estimar los parámetros $\boldsymbol{\beta}$ del modelo por medio del método de máxima verosimilitud obtienendo así la gradiente de la función log verosimilitud, es decir ($\nabla \log L$). Sin embargo $\nabla \log L = 0$ es analíticamente intratable, (Camargo et al., 2012).

Maier (2014) optimiza en tres etapas la estimación de este tipo de modelos que pueden contener un gran número de variables dependientes: (1) valores iniciales y refinamiento de los mismos, (2) optimización por medio del algoritmo de Broyden-Fletcher-Goldfarb-Shanno (BFGS) y (3) optimización por el algoritmo de Newton-Raphson.

Los valores iniciales deben elegirse cuidadosamente para que el algoritmo de optimización de

la parametrización alternativa converja; están compuestos de coeficientes de regresión para las respuestas transformadas de la media y la precisión.

Para la optimización BFGS se utiliza la gradiente analítica en lugar de la numérica; así se determinan los puntos donde la derivada es igual a cero. Se utilizan los valores iniciales y se aplica el algoritmo BFGS con un criterio de convergencia de 10^{-5}. Los errores estándar de los valores estimados obtenidos de la matriz hesiana no son confiables por lo que debe aplicarse el siguiente paso.

Con las estimaciones obtenidas anteriormente, se realiza la estimación final con el algoritmo de Newton-Raphson con las derivadas de primer y segundo orden obtenidas de forma analítica para acelerar la convergencia (el criterio es 10^{-10}) y para asegurar la estabilidad numérica, los errores estándar se obtienen de la hessiana ($H(\theta)$).

3.4. Distribución conjunta a posteriori para la distribución Dirichlet con parametrización alternativa

Bajo el enfoque bayesiano Sennhenn-Reulen (2018) introducen distribuciones a priori objetivas $\boldsymbol{\beta}$ con distribución normal. Debido a que el modelo de regresión propuesto en esta tesis es complicado, no se puede incluir una distribución a priori subjetiva ya que no se tiene un conocimiento basado en datos históricos o una conjetura de expertos la cual debería ser el parámetro. Por lo tanto, según lo explicado en el Capítulo 2, cuando se calcula la distribución a posteriori, la mayoría o toda la inferencia surgirá de la función de verosimilitud. El modelo completo incluyendo la asignación de las distribuciones a priori es dado por:

$$\begin{aligned} Y_{i\bullet} &= (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,C}), \quad i = 1, ..., n \\ Y_{i\bullet}|\boldsymbol{\beta}, \gamma &\overset{ind}{\sim} Dir(\boldsymbol{\mu}_i, \exp(\gamma)). \\ \beta_{q,c} &\sim N(a, b), c = 1, ..., C, \quad q = 1, ..., Q \\ \gamma &\sim N(c, d), \end{aligned}$$

donde $\boldsymbol{\mu}_i = (\mu_{i,1}, \mu_{i,2}, \cdots, \mu_{i,C})$ y según (3.3), $\mu_{i,c} = \text{logit}^{-1}(\mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet c})$.

Como se asume que $Y_{i\bullet}$ son independientes, entonces:

$$f(Y_{i\bullet}|\boldsymbol{\beta}, \gamma) = \frac{\Gamma(\sum_{c=1}^{C} \exp(\gamma)\mu_{ic})}{\prod_{c=1}^{C} \Gamma(\exp(\gamma)\mu_{i,c})} \prod_{c=1}^{C} y_c^{(\exp(\gamma)\mu_{i,c}-1)}.$$

Luego, la función de verosimilitud está dada por:

$$L(\boldsymbol{\beta}, \gamma; \mathbf{Y}) = \prod_{i=1}^{n} \left\{ \frac{\Gamma(\sum_{c=1}^{C} \exp(\gamma)\mu_{ic})}{\prod_{c=1}^{C} \Gamma(\exp(\gamma)\mu_{i,c})} \prod_{c=1}^{C} y_c^{(\exp(\gamma)\mu_{i,c}-1)} \right\}.$$

Luego, la función de densidad conjunta a posteriori es dada por:

$$\begin{aligned} p(\boldsymbol{\beta}, \gamma \mid \mathbf{Y}) &\propto L(\boldsymbol{\beta}, \gamma; \mathbf{Y}) p(\boldsymbol{\beta}, \gamma), \\ &\propto L(\boldsymbol{\beta}, \gamma; \mathbf{Y}) p(\boldsymbol{\beta}) p(\gamma), \\ &\propto L(\boldsymbol{\beta}, \gamma; \mathbf{Y}) \left[\prod_{i=1}^{n} \prod_{c=1}^{C} p(\beta_{i,c}) \right] p(\gamma), \\ &\propto \prod_{i=1}^{n} \left\{ \frac{\Gamma(\sum_{c=1}^{C} \exp(\gamma)\mu_{ic})}{\prod_{c=1}^{C} \Gamma(\exp(\gamma)\mu_{i,c})} \prod_{c=1}^{C} y_c^{(\exp(\gamma)\mu_{i,c}-1)} \right\} \times \\ &\left[\prod_{i=1}^{n} \prod_{c=1}^{C} \frac{1}{\sqrt{2\pi b}} \exp\left\{ \frac{-1}{2b} (\beta_{i,c} - a)^2 \right\} \right] \times \\ &\frac{1}{\sqrt{2\pi d}} \exp\left\{ \frac{-1}{2d^2} (\gamma - c)^2 \right\}. \end{aligned} \tag{3.6}$$

Sea $\theta = \{\beta, \gamma\}$, el Monte Carlo halmitoniano (HMC) toma los parámetros θ como si colectivamente denotasen la posición de la partícula en algún espacio con momento (cantidad de movimiento) ϕ. Así se asume que la energía potencial del sistema halmitoniano es $U(\theta) = -\log(p(\theta \mid Y))$, donde $p(\theta \mid Y) = p(\beta, \gamma \mid Y)$ es la fdp de la distribución a posteriori en la ecuación 3.6. Se asume que la fdp aposteriori $p(\theta, \phi \mid Y) \propto \exp(-H(\theta, \phi)) = \exp[\log(p(\theta \mid Y)) - \phi^T M^{-1} \phi)]$. Luego el algoritmo consiste en simular θ y ϕ a partir de $p(\theta, \phi \mid Y)$. El algoritmo HMC consiste en repetir los pasos siguientes L veces:

1. En la iteración t, actualiza ϕ a partir de $\phi^{(t-1)} \sim N(0, I)$.
2. Actualizar $\theta^{(t-1)}$ y $\phi^{(t-1)}$ simultáneamente a través de L pasos de "leapfrog", tal que cada paso es escalado por ϵ, un valor pequeño que sirve para el tuning del algoritmo.

   a) Actualizar ϕ,

   $$\phi^\star \leftarrow \phi^{(t-1)} + \frac{1}{2} \epsilon \nabla_{\theta^{(t-1)}} \log(p(\theta^{(t-1)}, \phi^{(t-1)} \mid Y)).$$

   b) Use el vector momento $\phi^\star$ para actualizar la posición del vector θ:

   $$\theta^\star \leftarrow \theta^{(t-1)} + \epsilon M^{-1} \phi^\star.$$

   c) Actualizar ϕ:

   $$\phi^\star \leftarrow \phi^\star + \frac{1}{2} \epsilon \nabla_{\theta^\star} \log(p(\theta^\star, \phi^\star \mid Y)).$$

3. Calcule

$$r = \frac{p(\theta^\star \mid Y)p(\phi^\star)}{p(\theta^{(t-1)} \mid Y)p(\phi^{(t-1)})},$$

donde $\theta^\star$ y $\phi^\star$ son los valores iniciales, mientras que $\theta^\star$ y $\phi^\star$ son los valores candidatos y ∇_θ es el gradiente con respecto a θ.

4. Actualice $\theta^{(t)} = \theta^\star$ con probabilidad $min(r, 1)$ o no actualice $\theta^{(t)} = \theta^{(t-1)}$ caso contrario.

5. Repetir los pasos 1 a 4 hasta llegar a la convergencia de las cadenas.

Un método más eficiente para generar muestras de las distribuciones a posteriori de $\boldsymbol{\beta}$ y γ es el algoritmo NUTS.

Capítulo 4

Estudio de simulación

En el presente capítulo se muestran los resultados del estudio de simulación de recuperación de parámetros con distintos escenarios considerando el modelo de regresión Dirichlet bajo un enfoque bayesiano. Para medir la efectividad de la recuperación también se obtienen los parámetros por inferencia clásica. Para la inferencia clásica se usa la libreria *DirichletReg* en R y para la inferencia bayesiana, se utiliza el software libre *Stan* con su interface en R. Se mide la calidad de las predicciones obtenidas con las estimaciones de los parámetros. Finalmente se realizan réplicas de las simulaciones de uno de los escenarios para verificar la consistencia de los parámetros conforme el tamaño de muestra se incrementa.

4.1. Generación de los datos

En el estudio se simula una matriz $\mathbf{Y}$ de datos composicionales; para lo cual se establece la cantidad de sujetos $n = 720$ y las categorías $c = 3$ para cada Y_i. Se utiliza un conjunto de covariables $\mathbf{X}$, y los coeficientes de la regresión $\boldsymbol{\beta}$ y el parámetro de precisión γ que se recuperán más adelante con la simulación. Se asume una covariable dicotómica que puede ser 0 ó 1; entonces:

$$\mathbf{X} = \begin{bmatrix} x_{1\bullet} \\ x_{2\bullet} \\ \vdots \\ x_{720\bullet} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}.$$

Para $\boldsymbol{\beta}$ se define la siguiente matriz de coeficientes de regresión:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{\bullet 1} & \beta_{\bullet 2} & \beta_{\bullet 3} \end{bmatrix} = \begin{bmatrix} \beta_{1,1} & \beta_{2,1} & 0 \\ \beta_{1,2} & \beta_{2,2} & 0 \end{bmatrix}.$$

Con la matriz de covariables $\mathbf{X}$ y coeficientes de regresión $\boldsymbol{\beta}$ se procede a calcular el vector $\boldsymbol{\mu}_i = (\mu_{i,1}, \mu_{i,2}, \mu_{i,3})$ dado en (3.1), según (3.4) y (3.5). Enseguida con la precisión $\gamma = 4{,}22$ se puede calcular θ.

A continuación se simulan los datos de acuerdo al modelo de regresión con parametrización alternativa $Y_{i\bullet} \sim Dir(\boldsymbol{\mu}_i, \theta)$ para $i = 1, 2, \cdots, n,$.

Para simular los conjuntos de datos se utilizó el paquete `Compositional` de R el cual requiere que se especifique el tamaño de la muestra y $\alpha_c(x_{i\bullet})$, para $c = 1, \ldots, C,$; parámetros asociados con los valores de los parámetros $\boldsymbol{\beta}$ y γ. La matriz resultante $\mathbf{Y}$ con los datos simulados para el caso de $c = 3$ tiene la siguiente estructura:

$$\mathbf{Y} = \begin{bmatrix} Y_{1\bullet} \\ Y_{2\bullet} \\ \vdots \\ Y_{720\bullet} \end{bmatrix} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & Y_{1,3} \\ Y_{2,1} & Y_{2,2} & Y_{2,3} \\ \vdots & \vdots & \vdots \\ Y_{720,1} & Y_{720,2} & Y_{720,3} \end{bmatrix}.$$

4.2. Consideraciones para la estimación de parámetros

Se consideran 3 escenarios con distintos valores de los parámetros a ser recuperados estimados con las simulaciones, como se muestra en el cuadro 4.1:

Cuadro 4.1: Escenarios para la simulación de datos.

	Escenarios		
Parámetro	**No. 1.**	**No. 2.**	**No. 3.**
$\beta_{1,1}$	0.46	1.11	1.43
$\beta_{2,1}$	-0.47	-0.07	-0.10
$\beta_{1,2}$	-0.12	0.54	-1.79
$\beta_{2,2}$	-0.56	-0.12	0.11
γ	4.22	4.00	6.33

Las distribuciones a priori seleccionadas son débilmente informativas con varianzas grandes; según lo explicado en el capítulo 3, expresan información vaga o general sobre los parámetros de interés y permiten que la información se extraiga de la función de verosimilitud, es decir que tienen un impacto mínimo en la distribución a posteriori:

$$\begin{aligned} \beta_{q,c} &\sim N(0, 1000^2), c = 1, 2, 3, \quad q = 1, 2, \cdots, Q \\ \gamma &\sim N(0, 1000^2). \end{aligned}$$

La estimación de los parámetros de la regresión Dirichlet se realiza mediante inferencia clásica

e inferencia bayesiana.

Para inferencia clásica se utiliza el método de máxima verosimilitud implementado en la función del paquete `Compositional` del programa R.

En inferencia bayesiana se utiliza el código implementado por (Sennhenn-Reulen, 2018) para el modelo de Dirichlet en RStan y códigos propios elaborados para esta tesis en R. Se considera como estimadores de los parámetros a las medias a posteriori, de las muestras extraídas con NUTS, que se describe en los Capítulos 2 y 3. NUTS determina adaptativamente el número de pasos L durante el muestreo, (Stan Development Team, 2019). Se simulan cuatro cadenas con 2000 iteraciones para cada parámetro, las primeras 1000 iteraciones de cada cadena fueron descartadas como *burn-in*, dejando un total de 4000 iteraciones.

4.3. Estimación de parámetros

En los cuadros 4.2, 4.3 y 4.4 se muestran los resultados de las estimaciones de los parámetros $\boldsymbol{\beta}$ y γ bajo los tres escenarios, por máxima verosimilitud (EMV) y los intervalos de confianza, y las estimaciones de la media a posteriori (MAP) obtenidas a través de inferencia bayesiana y los intervalos de credibilidad. En general se obtuvieron los siguientes resultados:

- Los valores de las estimaciones de los parámetros por máxima verosimilitud e inferencia bayesiana se aproximan a los valores originales de los parámetros.
- Los intervalos de credibilidad en los 3 escenarios se consideran adecuados, dado que contienen el verdadero valor de los parámetros.
- $\hat{R} = 1$ permite confirmar la convergencia de las cadenas, se analiza en la siguiente sección de diagnósticos de convergencia y autocorrelación, se define esta estadística utilizada para evaluar la convergencia.

Cabe mencionar que el tiempo total aproximado de cada simulación fue de 0.20 horas en un computador con un procesador intel (R) core (TM) i5-4210u CPU @ 1.70GHz 2.40 GHz con una memoria RAM de 6.00GB con Windows 8.1.

Cuadro 4.2: Resumen de resultados para el escenario 1: estimación por máxima verosimilitud (EMV), intervalo de confianza (95 %), media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

Parámetro	**Original**	**Máxima Verosimilitud**			**Inferencia bayesiana**			
		2.5 %	**EMV**	**97.5 %**	**2.5 %**	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	0.46	0.45	0.48	0.51	0.45	0.48	0.51	1.00
$\beta_{2,1}$	-0.47	-0.54	-0.50	-0.46	-0.54	-0.50	-0.46	1.00
$\beta_{1,2}$	-0.12	-0.15	-0.12	-0.09	-0.15	-0.12	-0.09	1.00
$\beta_{2,2}$	-0.56	-0.61	-0.56	-0.52	-0.61	-0.56	-0.52	1.00
γ	4.22	4.20	4.28	4.35	4.20	4.27	4.35	1.00

A continuación la Figura 4.1 muestra las funciones de densidad a posteriori de los parámetros en el escenario 1, las representaciones gráficas se construyen con las muestras a posteriori de los parámetros $\boldsymbol{\beta}$ y γ. También se observan las estimaciones de las medias a posteriori de los parámetros y las estimaciones de los intervalos de credibilidad. Se puede observar en el gráfico que los límites inferiores y superiores del intervalo de credibilidad contienen al verdadero valor de los parámetros. Además, que las distribuciones a posteriori de los parámetros $\boldsymbol{\beta}$ y γ se asemejan a una distribución normal.

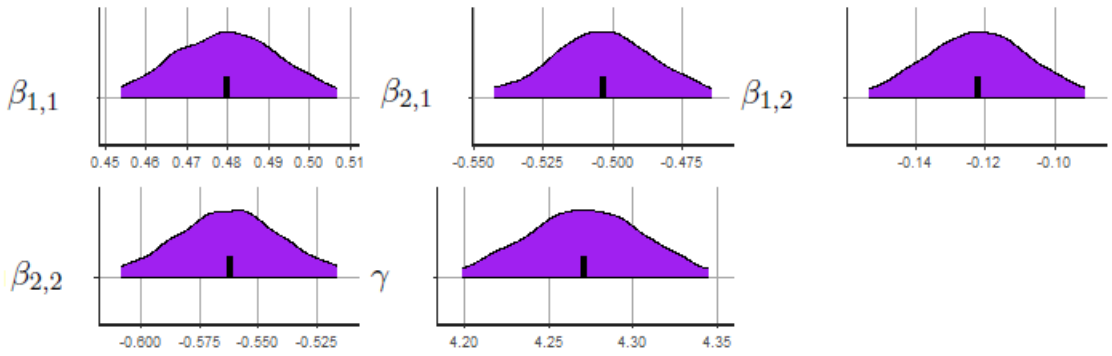


Figura 4.1: Funciones de densidad a posteriori de los parámetros en el escenario 1, medias a posteriori e intervalos de credibilidad.

Los cuadros 4.3 y 4.4 muestran los resultados de los escenarios 2 y 3. Las conclusiones son similares a las expuestas para el escenario 1.

Cuadro 4.3: Resumen de resultados para el escenario 2: estimación por máxima verosimilitud (EMV), intervalo de confianza (95 %), media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

Parámetro	**Original**	**Máxima Verosimilitud**			**Inferencia bayesiana**			
		2.5 %	**EMV**	**97.5 %**	**2.5 %**	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	1.11	1.05	1.09	1.13	1.05	1.09	1.12	1.00
$\beta_{2,1}$	-0.07	-0.12	-0.07	-0.01	-0.12	-0.07	-0.01	1.00
$\beta_{1,2}$	0.54	0.48	0.52	0.56	0.48	0.52	0.56	1.00
$\beta_{2,2}$	-0.12	-0.15	-0.09	-0.03	-0.15	-0.09	-0.03	1.00
γ	4.00	3.92	3.99	4.07	3.92	3.99	4.06	1.00

Cuadro 4.4: Resumen de resultados para el escenario 3: estimación por máxima verosimilitud (EMV), intervalo de confianza (95 %), media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

Parámetro	**Original**	**Máxima Verosimilitud**			**Inferencia bayesiana**			
		2.5 %	**EMV**	**97.5 %**	**2.5 %**	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	1.43	1.42	1.43	1.44	1.42	1.43	1.44	1.00
$\beta_{2,1}$	-0.10	-0.11	-0.10	-0.08	-0.11	-0.10	-0.08	1.00
$\beta_{1,2}$	-1.79	-1.84	-1.81	-1.79	-1.84	-1.81	-1.79	1.00
$\beta_{2,2}$	0.11	0.11	0.15	0.18	0.11	0.15	0.18	1.00
γ	6.33	6.31	6.38	6.46	6.30	6.38	6.45	1.00

4.4. Diagnósticos de convergencia y de autocorrelación

Según lo analizado en el Capítulo 2, es necesario realizar un diagnóstico de convergencia para verificar si las cadenas de Markov alcanzaron su estacionariedad. A continuación la figura 4.2 presenta los diagnósticos de convergencia, donde los gráficos de traza muestran una adecuada convergencia para el escenario 1. Se observa que el centro de las cadenas de cada muestra a posteriori de cada parámetro parece estar alrededor de los valores originales de los parámetros, con fluctuaciones que se ajustan al intervalo de credibilidad. Las cadenas se mezclan bien ya que realizan pasos largos y atraviesan las distribuciones rápidamente lo que indica que podrían haber alcanzado estacionariedad.

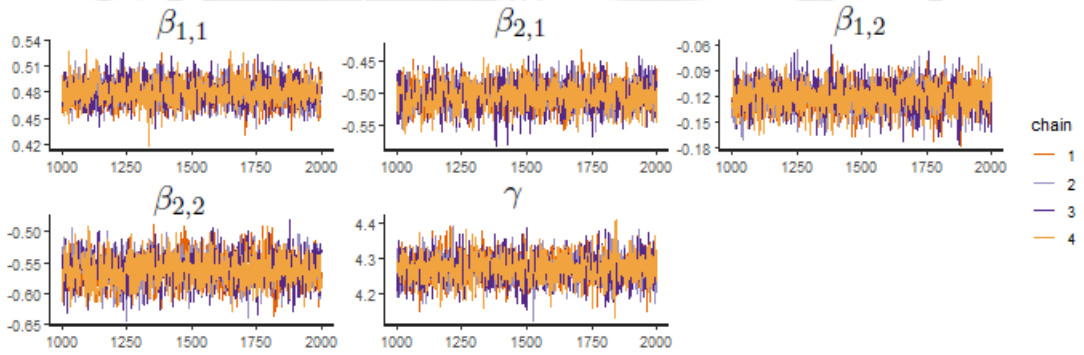


Figura 4.2: Convergencia de las cadenas para el escenario 1.

El siguiente diagnóstico es el de autocorrelación, el cual mide dependencia entre las muestras de las cadenas. La Figura 4.3 muestra una baja autocorrelación en todos los parámetros, lo que confirma que la mezcla de cadenas es eficiente. Por lo tanto, no es necesario usar saltos *thining*, es decir tomar la k-ésima observación en lugar de todos los valores obtenidos en la cadena, a fin de reducir la autocorrelación.

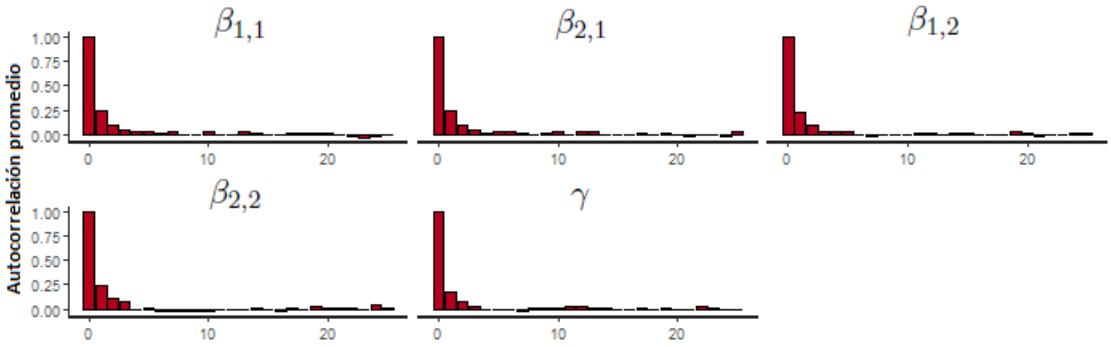


Figura 4.3: Gráficos de autocorrelación para la muestra a posteriori en el escenario 1.

La última verificación realizada es el test estadístico MCMC de Gelman y Rubin (Gelman et al., 2003), que permite confirmar la convergencia de las cadenas. Se conoce también como factor de reducción de escala potencial $\hat{R}$ que hace una comparación de las variaciones dentro de la cadena y entre las cadenas. Una gran desviación entre estas dos varianzas indica la no convergencia. Se obtuvo $\hat{R}$=1 para todos los parámetros, en los 3 escenarios, lo que indica que todas las cadenas simuladas se mezclaron adecuadamente.

Finalmente, para evaluar la calidad de las predicciones del modelo de regresión de Dirichlet, se calculó $\hat{\mathbf{Y}}$ con las estimaciones de los parámetros $\hat{\boldsymbol{\beta}}$ y $\hat{\gamma}$ bajo los 3 escenarios. Luego se utilizó el error cuadrático medio de predicción (MSE), que se define como la diferencia cuadrática promedio entre los valores observados $\mathbf{Y}$ y las predicciones $\hat{\mathbf{Y}}$, (Rawlins et al., 1986), es decir,

$$\text{MSE} = \frac{1}{n}(\mathbf{Y}_{i,c} - \hat{\mathbf{Y}}_{i,c})^2, \quad n = 1, ..., 720, c = 1, 2, 3.$$

En el cuadro 4.5 se observa que los MSE obtenidos por máxima verosimilitud e inferencia bayesiana son pequeños y similares. Sin embargo podemos resaltar que en el escenario 3 el MSE más bajo corresponde al de inferencia bayesiana.

Cuadro 4.5: Errores medios cuadráticos de los 3 escenarios de simulación.

Escenario	**Máxima verosimilitud**	**Inferencia bayesiana**
1	0.002890	0.002891
2	0.003665	0.003665
3	0.046549	0.043061

4.5. Réplicas

Se realizan cien réplicas de la estimación de los parámetros β y γ bajo el escenario 1, con inferencia bayesiana para diferentes tamaños de muestra $n = (50, 100, 250, 500, 720)$ con el fin de examinar la consistencia de los estimadores. Para medir la eficiencia de los resultados obtenidos se utilizó el sesgo, MSE y cobertura del intervalo de credibilidad para cada parámetro. Los resultados de las réplicas realizadas se presentan en el cuadro 4.6 donde se demuestra la capacidad del modelo para recuperar los parámetros β y γ con distintos tamaños de n con valores de sesgo y MSE pequeños. En el caso de la cobertura del intervalo de credibilidad, todos fueron igual a uno, por lo cual se omite este resultado.

Cuadro 4.6: Error medio cuadráticos y sesgo de los parámetros β y γ en réplicas de simulación del escenario 1

n	**Parámetro**	**Valor real**	**Sesgo**	**MSE**
50	$\beta_{1,1}$	0.46	-0.020485	0.000421
	$\beta_{2,1}$	-0.47	0.005579	0.000035
	$\beta_{1,2}$	-0.12	0.052244	0.002731
	$\beta_{2,2}$	-0-56	-0.006344	0.000045
	γ	4.22	-0.199132	0.039663
100	$\beta_{1,1}$	0.46	-0.011922	0.000143
	$\beta_{2,1}$	-0.47	-0.020934	0.000440
	$\beta_{1,2}$	-0.12	0.012668	0.000161
	$\beta_{2,2}$	-0-56	0.054871	0.003013
	γ	4.22	-0.042797	0.001835
250	$\beta_{1,1}$	0.46	0.000845	0.000018
	$\beta_{2,1}$	-0.47	-0.026911	0.000758
	$\beta_{1,2}$	-0.12	-0.009183	0.000102
	$\beta_{2,2}$	-0-56	0.025016	0.000663
	γ	4.22	0.006290	0.000097
500	$\beta_{1,1}$	0.46	-0.006347	0.000040
	$\beta_{2,1}$	-0.47	0.002425	0.000006
	$\beta_{1,2}$	-0.12	-0.016447	0.000271
	$\beta_{2,2}$	-0-56	0.012336	0.000152
	γ	4.22	0.020206	0.000409
720	$\beta_{1,1}$	0.46	0.020416	0.000417
	$\beta_{2,1}$	-0.47	-0.033959	0.001153
	$\beta_{1,2}$	-0.12	-0.002039	0.000004
	$\beta_{2,2}$	-0-56	-0.002328	0.000006
	γ	4.22	0.049397	0.002440

Capítulo 5

Aplicación

En este capítulo se aplica el modelo de regresión Dirichlet bayesiano a datos composicionales obtenidos de la Encuesta Demográfica y de Salud Familiar (2017), para estimar la prevalencia de niveles de anemia en niños de los centros poblados del Perú. En la aplicación se utilizan la información recabada y análisis realizados en el desarrollo de la tesis; llevando a cabo las siguientes actividades: descripción de la problemática de la anemia en el Perú y de los datos disponibles, presentación del modelo de regresión Dirichlet y relevancia, modelamiento y evaluación del modelo.

5.1. Importancia de modelar la prevalencia de anemia en el Perú

La anemia es una afección en la que la sangre no transporta suficiente oxígeno hacia los órganos del cuerpo; se produce por la falta de glóbulos rojos o la presencia de glóbulos rojos disfuncionales. En Perú, el Instituto Nacional de Estadística (INEI) realiza una prueba para determinar anemia en niñas y niños a través de la Encuesta Demográfica y de Salud Familiar (ENDES). En esta encuesta se clasifica los niveles de anemia como severa ($<$ 7.0 g/dl), moderada (7.0-9.9 g/dl) o leve (10.0-11.9 g/dl). En general, en el 2017, 43.6 % de los niños entre 6 y 36 meses de edad tuvieron algún grado de anemia (Colegio Médico del Perú, 2018). Los efectos de esta enfermedad se resumen a continuación.

Alcazar (2012) y Zavaleta y Astete (2017) explican que la anemia provoca deficiencias en el desarrollo cognitivo de los niños, en especial en sus habilidades psicomotrices, cognitivas y de socialización. Los niños con deficiencia de hierro tienen menos capacidad de atención, son más tímidos y dubitativos, menos perseverantes, menos alegres y desarrollan menos sus habilidades motrices. Entre los síntomas de la anemia (especialmente de la anemia por deficiencia de hierro) se encuentran la fatiga y el letargo, por lo que es de esperarse que estos problemas tengan un efecto en el desempeño laboral de los adultos. Esto afectaría en particular a aquellas personas cuyas labores implican trabajo físico sostenido (trabajadores agrícolas, obreros, etc.), generando una reducción de su productividad a lo largo de su jornada. Dado que el pago por estas labores está directamente ligado a la cantidad producida, por tanto, la fatiga causada por la anemia tiene un efecto negativo en la cantidad que un individuo produce y, por ende una disminución en sus ingresos.

Ciertos estudios encuentran que a pesar de curarse la anemia luego del primer año el efecto negativo en el desarrollo cognitivo no se revierte, pues los niños que padecieron de anemia siguen teniendo menores puntajes en las pruebas de desarrollo cognitivo con relación a los niños que no la padecieron. Esto conlleva el costo directo que asume el estado peruano por la posible mayor tasa de repitencia o deserción escolar de los niños a causa de la anemia, (Alcazar, 2012).

La anemia genera una carga pesada en el desarrollo de los individuos desde temprana edad, ejerce un efecto negativo a largo plazo no solo en la vida de cada persona que la padece, sino también sobre la sociedad el gobierno en términos de gasto de educación, salud, productividad y retorno económico.

Se puede afirmar que es necesario contar con una herramienta que permita entender las causas más importantes de la anemia en niñas y niños menores de 5 años (de 0 a 59 meses de edad) a nivel de centros poblados con el propósito de mejorar las políticas públicas dirigidas a la reducción de la anemia en el país. Según (Moghadas y Star, 2010) el modelamiento matemático puede ser una herramienta para la toma de decisiones para mitigar adecuadamente las causas de una enfermedad con los recursos limitados con los que cuenta el gobierno. Puede utilizarse para ilustrar diferentes escenarios de los potenciales resultados de los niveles de prevalencia de anemia dependiendo de variaciones en las causas de la anemia. Otro de sus usos sería programar el día a día de las operaciones de salud pública. La importancia de esta tesis se basa en que la literatura es pobre respecto a estudios que realicen modelamiento de la prevalencia de los niveles de anemia (severa, moderada o leve) en todo un país con fines de explicación de las causas o predicción, sino únicamente estudios campo que recolectan muestras de sangre con el fin de medir la prevalencia de anemia o conocer sus causas en poblaciones pequeñas.

5.2. Entendimiento de los datos y análisis descriptivo

Para alcanzar el objetivo del estudio en este capítulo, es decir estimar los niveles de prevalencia de anemia en niños y niñas menores de 5 años a nivel de centros poblados se utilizan los datos provenientes de la Encuesta Demográfica y de Salud del 2017 realizada por el Instituto Nacional de Estadística e Informática (INEI) de Perú que contiene información de 35, 910 viviendas de todos los departamentos del Perú. Estas bases de datos se examinan y consolidan en un conjunto de datos para facilitar su análisis.

De forma resumida el proceso de modelamiento requiere que se sigan los pasos siguientes:

- Entendimiento de la importancia de modelar la prevalencia de anemia en centros poblados del Perú.
- Explorar el conjunto de datos a analizar.
- Unificar bases de datos con las variables recopiladas en los cuestionarios de hogar, mujer y de salud.

- Agrupar datos de información de los niveles de anemia (severa- moderada o leve (10.0-11.9 g/dl)) y no anemia para cada centro poblado como datos composicionales con tres categorías. Así se obtiene la matriz de la variable dependiente $\boldsymbol{Y}$. Cabe mencionar que en la encuesta realizada por el INEI existen 134 centros poblados donde se consideran datos de un solo niño por centro poblado, 174 centros poblados a 2 niños, 152 centros poblados a 3 niños, etc. En estos casos las proporciones podrían no ajustarse a la realidad del centro poblado i ya que las muestras tomadas son pequeñas.
- Verificar si el conjunto de datos no tiene valores duplicados y eliminar filas con NA.
- Identificar variables explicativas $\boldsymbol{X}$ que podrían estar asociadas a una mayor prevalencia de anemia en la población y realizar un análisis descriptivo para verificar asociación.
- Estimar los parámetros $\boldsymbol{\beta}$ y γ con su intervalos de credibilidad mediante métodos MCMC; en nuestro caso a través del algoritmo No-U-Turn Sampler (NUTS). Realizar el diagnóstico de convergencia de los parámetros y de autocorrelación.
- Determinar el modelo que mejor ajusta a los datos con las técnicas de MSE y WAIC.
- Con el mejor modelo obtenido interpretar los parámetros $\boldsymbol{\beta}$ con sus intervalos de credibilidad.
- Los códigos de programación en R se incluyen como anexo.

En la Figura 5.1 se presenta un diagrama ternario con la representación bidimensional de las proporciones de los niveles de anemia (severa - moderada o leve) y no anemia por centros poblados en el Perú. Cada vértice del triángulo representa una categoría con una concentración del 100 %. Cada punto en el interior del triángulo representa un centro poblado con las 3 categorías de anemia. La suma de estas proporciones es igual a 1. Se observa que casi todos los centros poblados del Perú tienen algún nivel de anemia. La mayoría de centros poblados tiene entre el 15 % y 50 % de anemia en sus niños. Existen centros poblados en los que la totalidad de los niños tienen algun grado de anemia. De acuerdo a criterios emitidos por una nutricionista que confirma lo observado en el diagrama ternario; la anemia en un nivel leve casi no presenta sintomatología por lo que gran cantidad de personas tiene anemia leve pero no está al tanto.

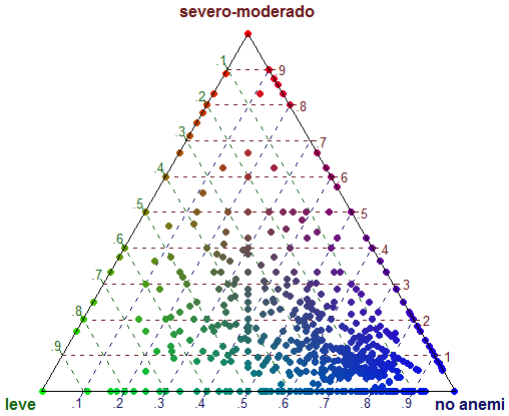


Figura 5.1: Diagrama ternario de prevalencia de anemia en niños del Perú por centros poblados.

A continuación se presentan diagramas de dispersión de cuatro potenciales covariables que podrían explicar la prevalencia de los niveles de anemia en niños y niñas menores de cinco años en centro poblados del Perú.

En la Figura 5.2 se observa que la prevalencia de anemia severa y moderada en los centros poblados podría tener mayor impacto en los niños con menor edad ya que la proporción de centros poblados con niños con niveles de anemia severa/moderada se reduce con el aumento de la edad promedio en meses.

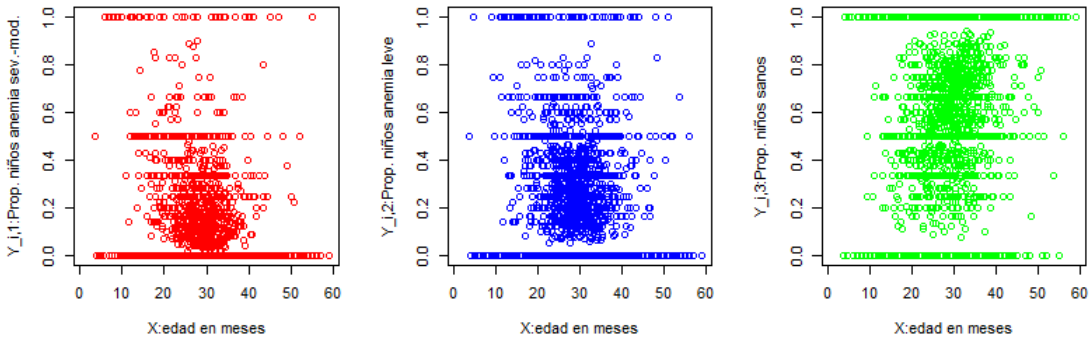


Figura 5.2: Diagramas de dispersión de la edad promedio (meses) de los niños en centros poblados vs.la proporción de niños con niveles de anemia severa o moderada (circulos rojos), anemia leve (circulos azules) y sin anemia (circulos verdes).

La Figura 5.3 indica que a mayor proporción de madres con educación superior en los centros poblados, menor proporción de con niños con anemia severa/moderada y leve ó bien en dichos centros poblados; a mayor proporción de madres con educación superior en los centros poblados, mayor proporción de niños sin anemia en los centros poblados.

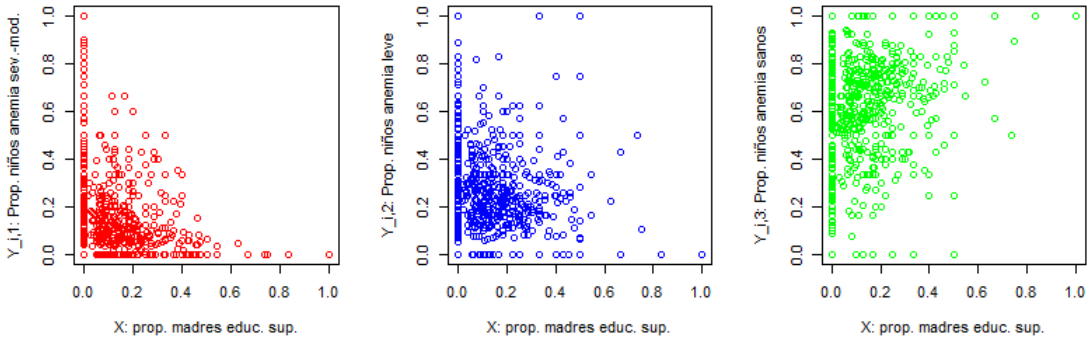


Figura 5.3: Diagramas de dispersión de la proporción de madres con educación superior en los centros poblados vs.la proporción de niños con niveles de anemia severa o moderada (circulos rojos), anemia leve (circulos azules) y sin anemia (circulos verdes).

La Figura 5.4 indica que para los centros poblados en los que se presenta anemia severa y moderada existen centros poblados en los que menos de 60 % de niños en dichos centros poblados consumen legumbres. Para los categorías de anemia leve y sanos se observa que existen centros poblados con alto consumo de legumbres en los niños. Es decir menor proporción de consumo de legumbres en los niños en los centros poblados implica una mayor propoción de niños con anemia severa/moderada y leve en los centros poblados.

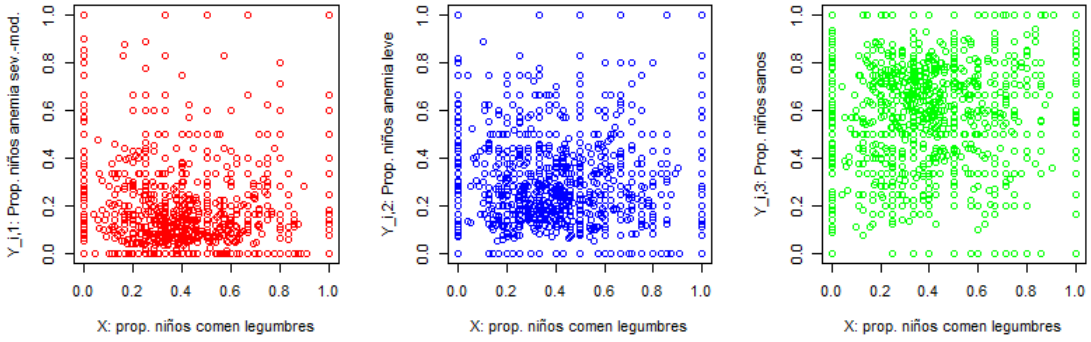


Figura 5.4: Diagramas de dispersión de la proporción de niños que comen legumbres/granos por centro poblado vs.la proporción de niños con niveles de anemia severa o moderada (circulos rojos), anemia leve (circulos azules) y sin anemia (circulos verdes.

En la Figura 5.5 se puede constatar que en varios centros poblados los niños con anemia severa y moderada tienen una menor proporción de consumo de grasas que las categorías de anemia leve y sanos.

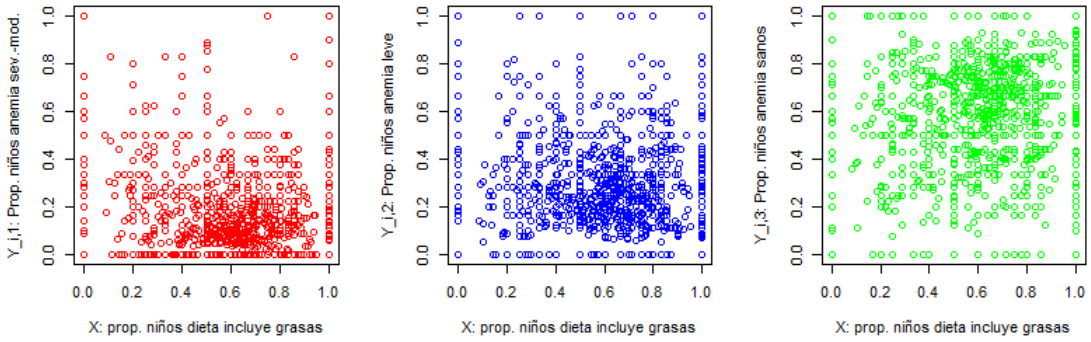


Figura 5.5: Diagramas de dispersión de la proporción de niños que comen grasas por centro poblado vs.la proporción de niños con niveles de anemia severa o moderada (circulos rojos), anemia leve (circulos azules) y sin anemia (circulos verdes).

Las variables identificadas en el análisis descriptivo coinciden con las opiniones expertos en nutrición que manifiestan que una mala alimentación puede ocasionar que un niño pase de un estado sin anemia a un tener anemia leve o moderada; aunque es más fácil que un niño que ha nacido con anemia leve en condiciones socio económicas vulnerables pase a un nivel moderado o severo.

Finalmente, también se puede observar que las proporciones de anemia en un centro poblado están en el intervalo [0, 1] en vez de (0,1) en donde se define la v.a. con distribución de Dirichlet. Para remediar este problema de una forma muy simple, si $y_{ic}^{\star}$ es el valor observado de la proporción de anemia en el i-ésimo centro poblado se puede usar la siguiente transformación propuesta por Smithson and Verkuilen (2006), $y_{ic} = \frac{y_{ic}^{\star}(n-1)+\frac{1}{2}}{n}$.

5.3. Modelo de regresión Dirichlet bayesiano

En este estudio la variable aleatoria $\boldsymbol{Y}$ representa los niveles de anemia, $Y_{i\bullet}$ corresponde a los niveles de anemia del i-ésimo centro poblado del Perú para $i = 1, 2, \ldots, n = 1559$, tal que $Y_{i\bullet} = (Y_{i,1}, Y_{i,2}, Y_{i,3})$, es decir $c = 1, 2, 3$, donde $c = 1$ es la categoría de referencia, y representa a los niveles de anemia severo y moderado, $c = 2$ representa a la categoría de anemia leve y $c = 3$ representa a la categoría sin anemia. Se asume que $Y_{i\bullet}$'s son independientes para diferentes centros poblados. Luego, se asume que la variable respuesta $\boldsymbol{Y}$ consiste en datos composicionales que siguen una distribución Dirichlet.

Sea $\boldsymbol{X}$ covariables que representan características de los centros poblados en el Perú. En esta tesis se estudia la relación de dependencia de la variable de respuesta $\boldsymbol{Y}$, respecto de una o

más variables explicativas que son componentes de $\boldsymbol{X}$. En general, se tiene que

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,2} & \cdots & x_{1,Q} \\ 1 & x_{2,2} & \cdots & x_{2,Q} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,2} & \cdots & x_{n,Q} \end{bmatrix} = \begin{bmatrix} x_{1\bullet} \\ x_{2\bullet} \\ \vdots \\ x_{n\bullet} \end{bmatrix},$$

donde $\mathbf{x}_{i\bullet}$ representa las covariables del i-ésimo centro poblado.

Se asume que $Q = 2$ si se incluye solo una covariable o $Q = 3$ si se incluye dos covariables, etc. Es decir, se asume que la primera columna de $\boldsymbol{X}$ está compuesta por un vector de unos, correspondiente a los interceptos y las columnas de covariables.

Bajo el enfoque bayesiano, Sennhenn-Reulen (2018) introducen distribuciones a priori objetivas $\boldsymbol{\beta}$ con distribución normal, de tal forma que cuando se calcule la distribución a posteriori, la mayoría o toda la inferencia surgirá de la verosimilitud. Según lo explicado en el capítulo 3, el modelo completo incluyendo la asignación de las distribuciones a priori es dado por:

$$\begin{aligned} Y_{i\bullet} &= (Y_{i,1}, Y_{i,2}, Y_{i,3}), \quad i = 1, ..., n \\ Y_{i\bullet}|\boldsymbol{\beta}, \gamma &\overset{iid}{\sim} Dir(\boldsymbol{\mu}_i, \exp(\gamma)). \\ \beta_{q,c} &\sim N(0, 1000), 1 = 1, 2, 3, \quad q = 1, ..., Q \\ \gamma &\sim N(0, 1000), \end{aligned}$$

donde $\boldsymbol{\mu}_i = (\mu_{i,1}, \mu_{i,2}, \mu_{i,3})$ y según (3.3), se tiene:

$$\mu_{i,c} = \text{logit}^{-1}(\mathbf{x}_{i\bullet}\boldsymbol{\beta}_{\bullet c}),$$

donde $\beta_{\bullet c}$ representa el vector de coeficientes de regresión de la categoría c, que pertenece a la matriz de coeficientes de regresión definida por:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \\ \vdots & \vdots & \vdots \\ \beta_{Q,1} & \beta_{Q,2} & \beta_{Q,3} \end{bmatrix} = \begin{bmatrix} \beta_{\bullet 1} & \beta_{\bullet 2} & \beta_{\bullet 3} \end{bmatrix},$$

donde dado que la categoría $c = 3$ es tomada como referencia, $\beta_{\bullet 3} = 0$.

Luego, la función de verosimilitud está dada por:

$$L(\boldsymbol{\beta}, \gamma; \mathbf{Y}) = \prod_{i=1}^{n} \left\{ \frac{\Gamma(\sum_{c=1}^{3} \exp(\gamma)\mu_{ic})}{\prod_{c=1}^{3} \Gamma(\exp(\gamma)\mu_{i,c})} \prod_{c=1}^{3} y_c^{(\exp(\gamma)\mu_{i,c}-1)} \right\},$$

Luego, la función de densidad conjunta a posteriori es dada por:

$$\begin{aligned} p(\boldsymbol{\beta}, \gamma|\mathbf{Y}) &\propto L(\boldsymbol{\beta}, \gamma; \mathbf{Y}) p(\boldsymbol{\beta}, \gamma), \\ &\propto L(\boldsymbol{\beta}, \gamma; \mathbf{Y}) p(\boldsymbol{\beta}) p(\gamma). \end{aligned}$$

En este modelo de regresión bayesiano se estiman los parámetros $\boldsymbol{\beta}$ y γ a partir de la distribución condicional a posteriori $p(\boldsymbol{\beta}, \gamma|\boldsymbol{Y})$. En el Cuadro 5.1 constan las covariables identificadas en el análisis descriptivo que se utilizan para ajustar 5 modelos de regresión.

Cuadro 5.1: Modelos considerados para ajuste

Modelo	**Covariables (X)**
1.	Proporción de niños alimentados con legumbres/granos.
2.	Proporción de niños que incluyen grasas en su dieta.
3.	Edad promedio (en meses) de los niños.
4.	Proporción de madres con educación superior.
5.	Proporción de madres con educación superior. Edad promedio (en meses de los niños).

Para estimar los parámetros mencionados se utiliza una variante de HMC a través de Rstan.

5.4. Diagnósticos de convergencia y autocorrelación

Se muestran los diagnósticos gráficos de convergencia y autocorrelación del modelo 3; estos resultados son análogos a los presentados por los otros cuatro modelos.

Los gráficos de traza de la parte inferior muestran cuatro cadenas con 1000 iteraciones cada una por cada parámetro luego de realizar `burn in`. Las cadenas realizan pasos largos y atraviesan las distribuciones rápidamente lo que indica que podrían haber alcanzado convergencia. Se puede concluir que las cadenas se mezclan bien.

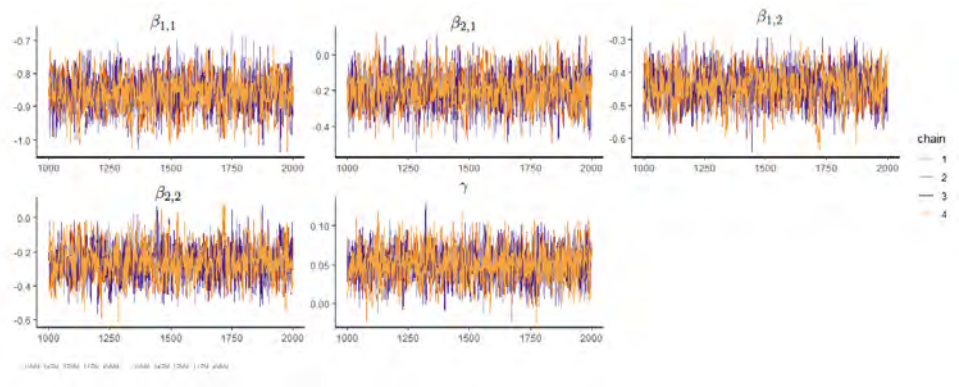


Figura 5.6: Convergencia de las cadenas en el modelo 3.

La Figura 5.7 muestra como la autocorrelación desciende rápidamente alrededor de cero para todos los parámetros, lo que confirma que la mezcla de las cadenas es eficiente. No fue necesario aplicar *thinning* sobre los valores obtenidos en la cadena a fin de reducir la autocorrelación.

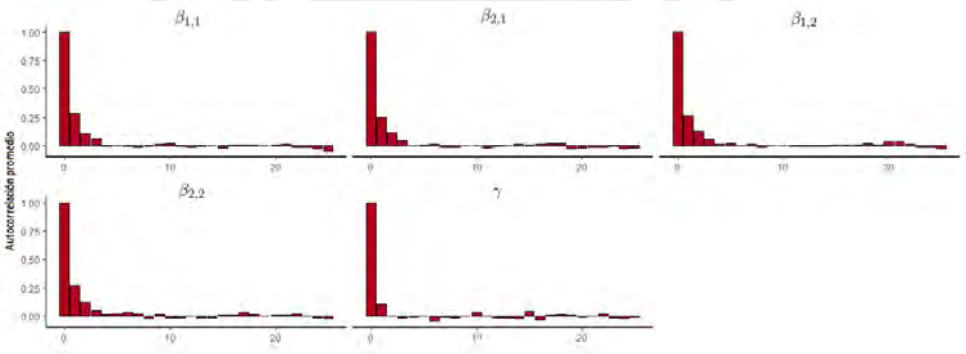


Figura 5.7: Autocorrelación de las muestras a posteriori en modelo 3.

Finalmente se analiza el test estadístico MCMC de Gelman y Rubin (Gelman et al., 2003), conocido como factor de reducción de escala potencial $\hat{R}$ que hace una comparación de las variaciones dentro de la cadena y entre las cadenas; es similar a un análisis de la varianza. En la sección 5.5 se verifica que se obtuvo $\hat{R}$=1 para todos los parámetros de los 5 modelos propuestos, lo que confirma la convergencia de las cadenas.

5.5. Estimación de parámetros

Previamente mencionamos que en $Y_{i,c}$, $c = 1, 2, 3$ categorías, donde $c = 1$ se refiere a los niveles de anemia severo y moderado, $c = 2$ a la categoría de anemia leve y $c = 3$ a la categoría sin anemia; esta última categoría es utilizada como categoría de referencia para la estimación de parámetros.

Los parámetros $\boldsymbol{\beta}$ y γ se estiman por inferencia bayesiana. Se examinan cinco modelos de regresión, la media a posteriori (MAP) de los coeficientes $\boldsymbol{\beta}$ y γ y los intervalos de credibilidad

para determinar si son significativos.

Según los resultados en el Cuadro 5.2, los intervalos de credibilidad de $\boldsymbol{\beta}$ del modelo 1 son significativos; lo que indica que la variable $X_{1\bullet}$: Proporción de niños alimentados con legumbres/granos en un centro poblado sirve para explicar los diferentes niveles de anemia y no anemia en dichos centros poblados. En particular, $e^{\beta_{2,1}}$ indica que para los centros poblados con niños alimentados con legumbres el odds de la proporción de centros poblados sin anemia disminuye en 17.8 % al pasar a la categoría anemia leve y $e^{\beta_{2,2}}$ muestra que el odds de la proporción de centros poblados sin anemia disminuye en 22.9 % al a la categoría anemia media y grave.

Cuadro 5.2: Resumen de resultados para el del modelo 1 con covariable $X_{1\bullet}$:proporción de niños alimentados con legumbres/granos. Media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

Parámetro	**Inferencia bayesiana**			
	2.5 %	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	-0.966	-0.861	-0.758	1.001
$\beta_{2,1}$	-0.383	-0.196	-0.001	1.002
$\beta_{1,2}$	-0.549	-0.444	-0.345	1.003
$\beta_{2,2}$	-0.448	-0.260	-0.067	1.002
γ	0.009	0.051	0.092	1.000

Según los resultados en el Cuadro 5.3, los intervalos de credibilidad de $\boldsymbol{\beta}$ y γ del modelo indican que $X_{2\bullet}$: proporción de niños que incluyen grasas en sus alimentos en un centro poblado si es significativa.

Cuadro 5.3: Resumen de resultados para el del modelo 2 con covariable $X_{1\bullet}$: proporción de niños que incluyen grasas en sus alimentos. Media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

Parámetro	**Inferencia bayesiana**			
	2.5 %	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	-0.915	-0.776	-0.638	1.003
$\beta_{2,1}$	-0.459	-0.267	-0.074	1.003
$\beta_{1,2}$	-0.531	-0.388	-0.247	1.002
$\beta_{2,2}$	-0.458	-0.262	-0.060	1.001
γ	0.011	0.052	0.094	1.000

Según los resultados en el Cuadro 5.4, los intervalos de credibilidad de $\beta_{2,1}$ y $\beta_{2,2}$ del modelo 3 son significativos, luego $X_{3\bullet}$: Edad promedio (en meses) del niño sirve para explicar la proporción del nivel de anemia leve, severo/moderado en los centros poblados.

Cuadro 5.4: Resumen de resultados para el del modelo 3 con covariable $X_{1\bullet}$: edad promedio (en meses) del niño. Media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

Parámetro	**Inferencia bayesiana**			
	2.5 %	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	-0.249	-0.023	0.198	1.000
$\beta_{2,1}$	-0.039	-0.032	-0.024	1.000
$\beta_{1,2}$	-0.086	0.135	0.363	1.001
$\beta_{2,2}$	-0.032	-0.025	-0.018	1.000
γ	-0.036	0.004	0.044	1.000

Según los resultados en el Cuadro 5.5, los intervalos de credibilidad de los coeficientes $\beta_{2,1}$ y $\beta_{2,2}$ del modelo 4 no son significativos, es decir que $X_{4\bullet}$: proporción de madres que tienen instrucción superior en los centros poblados no sirve para explicar el nivel de anemia severo/moderado.

Cuadro 5.5: Resumen de resultados para el del modelo 4 con con covariable $X^2_{1\bullet}$: proporción de madres que tienen instrucción superior. Media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

Parámetro	**Inferencia bayesiana**			
	2.5 %	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	-1.004	-0.933	-0.863	1.000
$\beta_{2,1}$	-1.459	-0.518	0.390	1.000
$\beta_{1,2}$	-0.631	-0.563	-0.495	1.000
$\beta_{2,2}$	-1.897	-0.965	0.038	1.000
γ	-0.059	-0.017	0.025	1.000

Según los resultados en el Cuadro 5.6 los intervalos de credibilidad de los coeficientes $\beta_{4,1}$ y $\beta_{4,2}$ del modelo 5 son significativos; por lo tanto $X_{3\bullet}$: la edad promedio (en meses) del niño en los centros poblados; si explica la proporción de los niveles de anemia; pero no la $X_{1\bullet}$: proporción de madres que tienen instrucción superior en los centros poblados, ni $X_{2\bullet} = X^2_{1\bullet}$.

Cuadro 5.6: Resumen de resultados para el del modelo 5 con covariables $X_{1\bullet}$: proporción de madres que tienen instrucción superior, $X_{2\bullet} = X_{1\bullet}^2$, $X_{3\bullet}$: edad promedio (en meses) del niño. Media a posteriori (MAP) , Intervalos de credibilidad (2.5, 97.5) %, diagnóstica de convergencia Rhat $\hat{R}$.

title of Table

Parámetro	**Inferencia bayesiana**			
	2.5 %	**MAP**	**97.5 %**	**$\hat{R}$**
$\beta_{1,1}$	-0.274	-0.047	0.197	1.000
$\beta_{2,1}$	-0.799	0.312	1.407	1.000
$\beta_{3,1}$	-2.719	-0.745	1.133	1.000
$\beta_{4,1}$	-0.039	-0.031	-0.024	1.000
$\beta_{1,2}$	-0.156	0.081	0.314	1.000
$\beta_{2,2}$	-1.131	-0.041	1.045	1.001
$\beta_{3,2}$	-2.619	-0.656	1.281	1.000
$\beta_{4,2}$	-0.030	-0.022	-0.015	1.000
γ	-0.028	0.014	0.057	1.000

5.6. Evaluación

Para evaluar la capacidad predictivade de los modelos de regresión se utiliza el WAIC, estudiado en el capítulo 2, que estima el número adecuado de parámetros para evitar un sobreajuste del modelo.

En segundo lugar se utiliza el error medio cuadrático MSE que se define como la diferencia cuadrática promedio entre los valores observados $\mathbf{Y}$ y las predicciones $\hat{\mathbf{Y}}$.

El Cuadro 5.7 indica que los modelos 1, 2 y 3 presentan un MSE más bajo de 0.066 y 0.065 respectivamente; y el WAIC para el modelo 3 con -11712.85 .

Cuadro 5.7: Criterios de selección de modelos.

Modelo	**MSE**	**WAIC**
1	0.066	-9866.23
2	0.065	-9868.30
3	0.067	**-11712.85**
4	0.069	-10812.64
5	0.098	-10747.70

5.7. Interpretación de parámetros del modelo seleccionado

Se realiza la interpretación de $e^{\boldsymbol{\beta}}$ con los coeficientes de odds (OR) del modelo 3 donde se obtuvieron los mejores resultados en la evaluación. Se tiene en cuenta lo siguiente:

- $\beta_{q,c} > 0$ implica $e^{\beta_{q,c}} > 1$ y el odds incrementa con X_i.
- $\beta_{q,c} < 0$ implica $e^{\beta_{q,c}} > 1$ y el odds disminuye con X_i.

El modelo 3 $e^{\beta_{2,1}}$ indica que cuando la edad promedio de los niños aumenta, el odds de la proporción de anemia disminuye en 3.15 % cuando la categoría es anemia leve, comparado con el odds de la proporción de anemia cuando la categoría es sin anemia. Y $e^{\beta_{2,2}}$ muestra que cuando la edad promedio de los niños aumenta en un mes, el odds de la proporción de anemia disminuye en 2.47 % cuando la categoría es anemia media y grave, comparado con el odds de la proporción de anemia cuando de la categoría de referencia.

Capítulo 6

Conclusiones

6.1. Conclusiones

En esta tesis se estudió acerca del modelamiento de una regresión Dirichlet bayesiano para datos composicionales que siguen una distribución Dirichlet; estos datos consisten en vectores de proporciones que surgen cuando se clasifican n sujetos en categorías disjuntas y se registran las frecuencias relativas resultantes; es decir que se encuentran en un intervalo acotado que suman uno.

En la regresión se utilizó la parametrización alternativa de la distribución Dirichlet donde se modela la media de la variable respuesta $\boldsymbol{\mu}$ y un parámetro de precisión θ_i. La perspectiva bayesiana utilizada se planteó con distribuciones a priori no informativas.

Se llevó a cabo un proceso de simulación de datos con la correspondiente recuperación de parámetros con tres escenarios considerando el modelo de regresión Dirichlet con enfoque bayesiano e inferencia clásica. Se utilizó el software libre RStan. La calidad de las predicciones obtenidas se evaluó con el MSE.

Se realizaron réplicas de las simulaciones de los parámetros $\boldsymbol{\beta}$ y γ del escenario 1 con diferentes tamaños de muestra n y se validó la capacidad del modelo para recuperar los parámetros ya que se obtuvieron valores pequeños de sesgo y MSE.

Se ajustaron varios modelos de regresión bayesiana Dirichlet para modelar los niveles de prevalencia de anemia en el Perú: no anemia, leve y moderado-severo. Se obtuvieron modelos en los cuales los intervalos de credibilidad de los coeficientes $\boldsymbol{\beta}$ resultaron significativos. En base a los resultados de MSE y WAIC.

6.2. Sugerencias para investigaciones futuras

Continuar con investigaciones que se enfoquen en:

- Comparar la efectividad del modelo de regresión Dirichlet bayesiano con otros modelos que podrían utilizarse para modelar proporciones.

- Determinar las circunstancias óptimas para modelar datos composicionales; por ejemplo al evaluar si es conveniente que las proporciones de cada individuo i hayan sido generadas a partir de muestras grandes o muestras homogéneas.
- Incluir otras covariables en los modelos.
- Analizar la distribución espacial de la anemia en el Perú.

Bibliografía

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*, Chapman and Hall Ltd.

Alcazar, L. (2012). *Impacto Económico de la Anemia en el Perú*, Impresiones y Ediciones Arteta E.I.R.L.

Amaral, M., Paulino, C. y Muller, P. (2019). *Computational Bayesian Statistics an Introduction*, Cambridge University Press.

Camargo, A., Stern, J. y M., L. (2012). Estimation and model selection in dirichlet regression, *American Institute of Physics* pp. 206–213.

Colegio Médico del Perú (2018). La Anemia en el Perú, *Reporte de Políticas de Salud* pp. 1–19.

Cribari, F. y Zeileis, A. (2010). Beta regression in r, *Journal of Statistical Sofware* pp. 1–24.

Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *J Appl Stat* pp. 1–18.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models(comment on article by browne and draper), *Bayesian Analysis* pp. 516–530.

Gelman, A., Carlin, J. y Stern, H. (2003). *Bayesian Data Analysis*, Chapman and Hall.

Gujarati, D. y Porter, D. (2010). *Econometría*, McGraw-Hill.

Hijazi, R. (2003). *Analysis of Compositional Data Using Dirichlet Covariate Models*, PhD thesis, American University, Washington, D.C.

Hijazi, R. y Jernigan, R. (2007). Modeling compositional data using dirichlet regression models, *Journal of Applied Probability and Statistics* pp. 77–91.

Hoffman, M. y Gelman, A. (2011). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, *Journal of Machine Learning Research* **15**: 1–30.

Kass, R. y Wasserman, L. (1994). Formal rules for selecting prior distributions, *Journal of the American Statistical Association* **91**: 1343–1370.

Maier, M. (2014). Dirichlet Regression for Compositional Data in R, *Institute for Statistics and Mathematics* pp. 1–25.

Moghadas, S. y Star, L. (2010). The rol of mathematical modelling in public health planning and dedision making, *Purple Paper* pp. 1–6.

Rawlins, J., Pantula, S. y Dickey, D. (1986). *Applied Regression Analysis: A research tool*, Springer.

Sennhenn-Reulen, H. (2018). Bayesian regression for a dirichlet distributed response using stan.

Stan Development Team (2019). *Stan Reference Manual.*
URL: *https://mc-stan.org/*

Tsagris, M. y Stewart, C. (2017). A dirichlet regression model for compositional data with zeros. https://arxiv.org/abs/1410.5011.

van der Merwe, S. (2018). A method for bayesian regression modelling of composition data.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*, Springer.

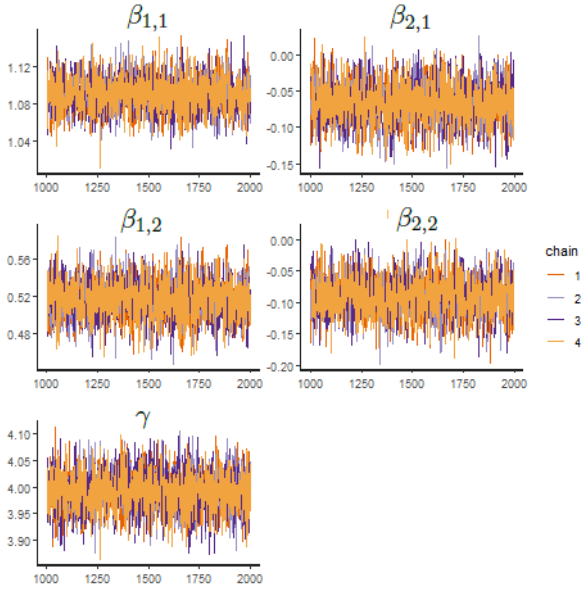
Wang, K., Liang, G. y Tang, M. (2011). *Dirichlet and Related Distributions*, John Wiley and Sons.

World Health Organization (2008). Worldwide prevalence of anaemia 1993-2005, pp. 1–35.

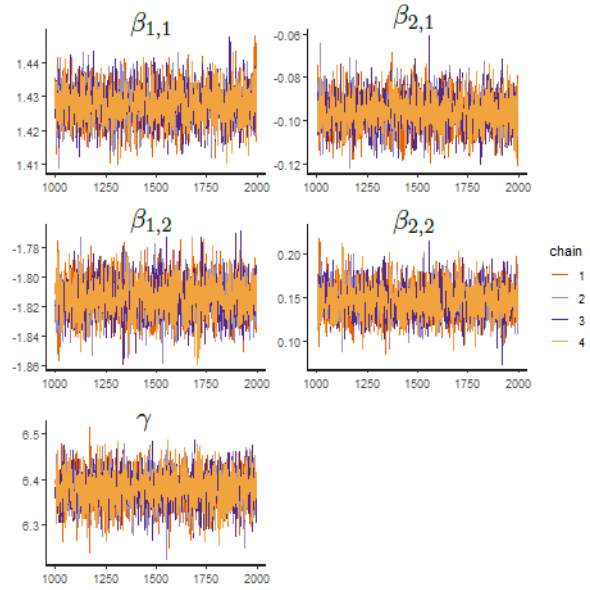
Zavaleta, N. y Astete, L. (2017). Efecto de la anemia en el desarrollo infantil: consecuencias a largo plazo, *Revista Peruana de Medicina Experimental y Salud* **34**: 77–91.

Diagnósticos de convergencia

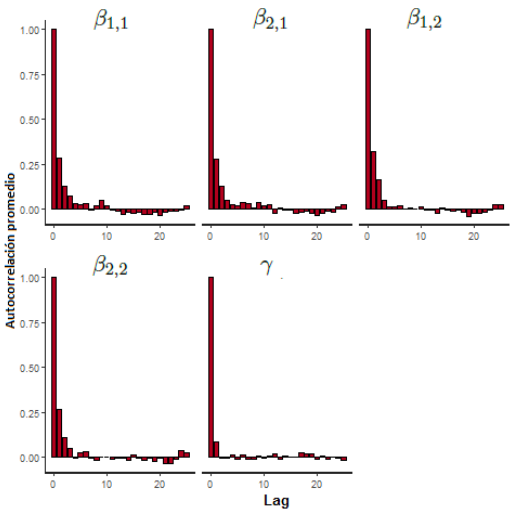
1. Diagnósticos de convergencia y autocorrelación en la simulación



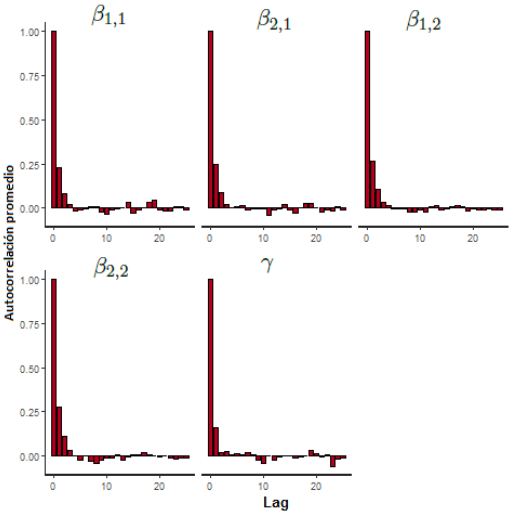
(a) Convergencia de las cadenas para el escenario 2.



(b) Convergencia de las cadenas para el escenario 3.



(c) Gráficos de autocorrelación de las muestras a posteriori para cada parámetro estimado en el escenario 2.



(d) Gráficos de autocorrelación de las muestras a posteriori para cada parámetro estimado en el escenario 3.

Código en R y Rstan de la aplicación

2. Código en R de la aplicación del modelo de regresión Dirichlet bayesiano seleccionado

```
#############################################################################
## descargando bases de datos en formato spss                              #
#############################################################################

dir.save <- getwd()
library(dplyr)
library(foreign)


mom_dataset    = read.spss(paste("RECH5.SAV", sep = ''), to.data.frame =
 TRUE)
child_dataset  = read.spss(paste("REC44.SAV", sep = ''), to.data.frame =
  TRUE)
home1_dataset  = read.spss(paste("RECH0.SAV", sep = ''), to.data.frame =
 TRUE)
home2_dataset  = read.spss(paste("RECH1.SAV", sep = ''), to.data.frame =
 TRUE)
home3_dataset  = read.spss(paste("RECH4.SAV", sep = ''), to.data.frame =
 TRUE)
home4_dataset  = read.spss(paste("RECH23.SAV", sep = ''), to.data.frame =
TRUE)


##home3_dataset$SH15N   nivel de educacion
id <- paste(substr(as.character(home3_dataset$HHID), 7, 15),
        home3_dataset$IDXH4, sep = '')
home3_dataset$newHHID <- id


id <- paste(substr(as.character(mom_dataset$HHID), 7, 15), mom_dataset$HA0,
        sep = '')

merge.mom_dataset <- left_join(mom_dataset, home3_dataset, by = "newHHID")
names(merge.mom_dataset)
dim(merge.mom_dataset)


lactancia1_dataset = read.spss(paste("REC41.SAV", sep = ''), to.data.frame =
 TRUE)


salud1_dataset = read.spss(paste("REC42.SAV", sep = ''), to.data.frame = TRUE)


#############################################################################
## juntando bases de datos                                                 #
#############################################################################
```

```
names(salud1_dataset)
dim(salud1_dataset)
child_dataset  <-
  inner_join(child_dataset, salud1_dataset, by = "CASEID")


dim(lactancia1_dataset)
dim(child_dataset)
child_dataset <-
  data.frame(
    child_dataset,
    M5 = lactancia1_dataset$M5,
    M55A = lactancia1_dataset$M55A,
    M55G = lactancia1_dataset$M55G
  )
names(child_dataset)
#child_dataset <- child_dataset[,-32]

### child dataset
HHI = substr(as.character(child_dataset$CASEID), 7, 15)
child_dataset <- cbind(child_dataset, HHID = HHI)

names(child_dataset)

#merge data home with datachild
home1_dataset$HV005
HH <- substr(as.character(home1_dataset$HHID), 7, 15)
home1_dataset$HHID <- HH

dim(child_dataset)
dim(home1_dataset)


merge.data.child  <-
  inner_join(child_dataset, home1_dataset, by = "HHID")
dim(merge.data.child)

HHID = substr(as.character(merge.mom_dataset$HHID.x), 7, 15)
merge.mom_dataset$HHID <- HHID

dim(child_dataset)
dim(merge.mom_dataset)
merge.data1  <-
  inner_join(merge.data.child, merge.mom_dataset, by = "HHID")



##############################################################################
# calculo de casos y proporciones de anemia en centros poblados por código   #
##############################################################################


tb <- table(merge.data1$nomccpp, merge.data1$HW57)
tb <- as.data.frame.matrix(tb)
tb$rowSum = rowSums(tb)

tb <-tb[tb$rowSum > 0, c("Severe", "Moderate", "Mild", "Not anemic")]

tb <- as.matrix(tb)
tb <- as.table(tb)
head(tb)
```

```
class(tb)

rowSum = rowSums(tb)
anemia.data <-
  data.frame(
    nomccpp = as.character(row.names(tb)),
    severe.moderate = tb[, 1] + tb[, 2],
    mild = tb[, 3],
    non = tb[, 4]
  )

prop.anemia.data <-
  data.frame(
    nomccpp = as.character(row.names(tb)),
    severe.moderate = (tb[, 1] + tb[, 2]) / rowSum,
    mild = tb[, 3] / rowSum,
    non = tb[, 4] / rowSum
  )

dim(prop.anemia.data)

prop.anemia.data[1:5,]

##############################################################################
# bases de datos de niveles de anemia                                        #
##############################################################################

ind_sev_mod <-
  which(merge.data1$HW57 == 'Moderate' |
          merge.data1$HW57 == 'Severe')
sev_mod <- merge.data1[ind_sev_mod,]

ind_mild_mod <- which(merge.data1$HW57 == 'Mild')
mild_mod <- merge.data1[ind_mild_mod,]

ind_non_mod <- which(merge.data1$HW57 == 'Not anemic')
non_mod <- merge.data1[ind_non_mod,]


##############################################################################
# Análisis descriptivo (Diagramas de dispersión)                             #
##############################################################################

### Edad en meses
resage <-
  aggregate(merge.data1$HW1, list(merge.data1$nomccpp), mean, na.rm = TRUE)
names(resage) <- c('nomccpp', 'HW1')

prop.anemia.data1 <-
  full_join(prop.anemia.data, resage , by = "nomccpp")
head(prop.anemia.data1)
write.csv(prop.anemia.data1,file="base5.csv")

par(mfrow = c(1, 3))
plot(
  prop.anemia.data1$HW1,
  prop.anemia.data1$severe.moderate,
  col = 'red',
  xlab = 'X:edad en meses',
  ylab = 'Y_i,1:Prop. niños anemia sev.-mod.',
  ylim = c(0, 1),
  xlim = c(10, 50)
)
```

```
plot(
  prop.anemia.data1$HW1,
  prop.anemia.data1$mild,
  col = 'blue',
  xlab = 'X:edad en meses',
  ylab = 'Y_i,2:Prop. niños anemia leve',
  ylim = c(0, 1),
  xlim = c(10, 50)
)
plot(
  prop.anemia.data1$HW1,
  prop.anemia.data1$non,
  col = 'green',
  xlab = 'X:edad en meses',
  ylab = 'Y_i,3:Prop. niños sanos',
  ylim = c(0, 1),
  xlim = c(10, 50)
)
mtext(
  "Edad media en niños vs prop. anemia",
  side = 3,
  line = -3,
  outer = TRUE
)

# Librerías a utilizar

library(DirichletReg) # Para dar valores iniciales al modelo de regresión
                                          # bayesiano.
library(rstan)
library(bayesplot) # Para graficar los intervalos de credibilidad
library("rstudioapi")
library(gridExtra) # Para acomodar los gráficos


##############################################################################
## fit 1                                                                    #
##############################################################################
# Se cargan datos

library(DirichletReg)
y = cbind(prop.anemia.data1$severe.moderate,prop.anemia.data1$mild,
          prop.anemia.data1$non)
x = prop.anemia.data1$HW1

#x = prop.anemia.data1$HW1
indNA = which(is.na(y[,1])==TRUE)
y <- y[-indNA,]
x <- x[-indNA]
#indNAx = which(is.na(x)==TRUE)
#y <- y[-indNAx,]
#x <- x[-indNAx]

X = cbind(1, x)

datos = data.frame(Y=y,X = x)

datos$Y<- DR_data(y)
datos$Y

head(datos)
```

```
# Se eliminan missing values
datos <- na.omit(datos)
colnames(datos)


## Se convierten las covariables en una matriz de diseño:
X <- as.matrix(model.matrix(lm(Y ~ X , data = datos)))
X <- matrix(nrow = nrow(X), ncol = ncol(X), data = as.numeric(X))


## Se define la variable de respuesta Y:

Y <- datos$Y

## Para el modelo con una distribución a priori N(0,1000^2)
## Se presenta la información como una lista, con sd_prior=1000
D1 <- list(N = nrow(Y), ncolY = ncol(Y), ncolX = ncol(X),
           X = X, Y = Y, sd_prior = 1000)

## Se estiman los parámetros con la distribución a
priori N(0,1000^2) utilizando la función
## "sampling" del paquete "Rstan":


fit1 <- sampling(prg, data = D1, chains = 4, iter = 2000, cores = 4,
                 control = list(adapt_delta = 0.95, max_treedepth = 20),
                 refresh = 100,seed=1)

## Se utiliza "extract" para acceder a las muestras a
posteriori obtenidas con la distribución
## a priori N(0,1000^2):

B <- extract(fit1)$beta
save(B, file = "B.RData")


## Con summary extraemos las estimaciones de los
parámetros en fit1 con una priori N(0,1000^2)

summary(fit1, c("beta","theta"))$summary

## Con "summary" extraemos las estimaciones de los
parámetros en fit1 con una priori N(0,1000^2)
fit1.coef<- summary(fit1, c("beta","theta"))$summary
fit1.coef
(exp.fit1 <- (exp(fit1.coef)))



# EL código siguiente fue implementado por Holger Sennhenn-Reulen (2018)
# Bayesian Regression for a Dirichlet DistributedResponse using Stan.  Arxiv.

## Se cargan las siguientes librerías

library("DirichletReg") # Para dar valores iniciales al
modelo de regresión bayesiano.

library("rstan")

stan_code <- "
data {
int<lower=1> N; // número total de observaciones
int<lower=2> ncolY; // número de categorías
```

```
int<lower=2> ncolX; // número de predictores
matrix[N,ncolX] X; // matriz de diseño de las predictoras
matrix[N,ncolY] Y; // variable de respuesta
real sd_prior; // desviación estándar de la distribución priori
}
parameters {
matrix[ncolY-1,ncolX] beta_raw; // coeficientes
real theta;
}
transformed parameters{
real exptheta = exp(theta);
matrix[ncolY,ncolX] beta; // coeficientes
for (l in 1:ncolX) {
beta[ncolY,l] = 0.0;
}
for (k in 1:(ncolY-1)) {
for (l in 1:ncolX) {
beta[k,l] = beta_raw[k,l];
}
}
}
model { // distribución a priori:
theta~normal(0,sd_prior);
for (k in 1:(ncolY-1)) {
for (l in 1:ncolX) {
beta_raw[k,l]~normal(0,sd_prior);
}
}
for (n in 1:N) { // verosimilitud
vector[ncolY] logits;
for (m in 1:ncolY){
logits[m] = X[n,] * transpose(beta[m,]);
}
transpose(Y[n,])~dirichlet(softmax(logits) * exptheta);
}
}
"
## Stan compila el código previamente introducido

prg <- stan_model(model_code = stan_code)
```