PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

FUSIÓN DE EFECTOS PARA MODELOS DE REGRESIÓN CON RESPUESTA POSITIVA BAJO UN ENFOQUE BAYESIANO

TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN ESTADÍSTICA

AUTOR

Andie Bryan Dongo Román

ASESOR

Cristian Luis Bayes Rodriguez

Setiembre, 2021

Dedicatoria

A cada una de las personas que me animaron a perseguir mis sueños y dar este paso tan importante, con todo cariño.

"Mucha gente pequeña,
en lugares pequeños,
haciendo cosas pequeñas,
puede cambiar el mundo"

Eduardo Galeano

Agradecimentos

Al profesor Cristian Bayes, por su paciencia, comprensión e indispensable guía en la construcción de este trabajo.

A mis padres, mi hermana y mis hermanos, quienes me motivan a ser una mejor persona cada día y a bregar por una sociedad mejor.

A los profesores y profesoras de la Maestría en Estadística de la Pontificia Universidad Católica del Perú, de quienes no sólo me llevo grandes aprendizajes, sino también gratos recuerdos. Su amabilidad y dedicación me inspiran.

A mi familia y mis amistades, siempre cuento con su apoyo en cada aventura.

Resumen

El presente trabajo tiene como objetivo adaptar el modelo bayesiano para fusión de efectos presentado por Pauger y Wagner (2019), de tal manera que sea adecuado para modelos de regresión con respuesta positiva bajo una distribución gamma. El modelo plantea como distribución a priori de los coeficientes de cada covariable cualitativa a una normal multivariada, deducida a partir de una distribución a priori spike y slab para la diferencia de cada par de efectos, cuya matriz de precisión permite conocer qué niveles pueden fusionarse. La estructura de la matriz de precisión depende de un hiperparámetro que permite estimar las probabilidades de fusión a posteriori entre cada par de niveles, con las cuales se pueden agrupar aquellos niveles con efectos similares mediante la función de pérdida de Binder. La estimación a posteriori del modelo es realizada con métodos MCMC utilizando el programa JAGS en R.

Se aplicó la metodología a un conjunto de datos reales extraído de la Encuesta Nacional de Hogares (ENAHO) del año 2019, donde se pudo verificar la existencia de una brecha salarial por etnicidad en los entrevistados de la macro región sur del Perú. Así mismo, se incluyó en el caso aplicativo a la interacción entre los efectos de la etnicidad y el sexo, revelándose que la brecha por género existente es mayor en la población aymara y en la no indígena, en comparación con la población quechua.

Palabras-clave: Inferencia bayesiana, métodos bayesianos, modelos para datos positivos, fusión de efectos, covariables nominales y ordinales, spike y slab, matriz dispersa, métodos de Montecarlo vía cadenas de Markov (MCMC), JAGS.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones preliminares

Cuando se construye un modelo de regresión es común encontrar que algunas de las covariables son cualitativas, las cuales pueden presentar dos o más niveles o categorías. Después de estimar los parámetros del modelo se podría encontrar que, para una determinada covariable cualitativa, alguno o algunos de sus niveles no tengan efecto significativo sobre la variable dependiente o que algunos de sus niveles tengan efectos significativos similares entre sí, por lo que podrían fusionarse. El desafío está en probar, en cada covariable, la equivalencia de los efectos de determinado grupo de niveles para fusionarlos y así obtener un modelo más simple.

Desde el enfoque frecuentista, existen métodos para seleccionar variables dentro de un modelo como el Lasso (Tibshirani, 1996) o Elastic net (Zou y Hastie, 2005), los cuales se basan en restricciones en la estimación de los coeficientes. Por otro lado, desde el enfoque bayesiano, algunos métodos logran la selección de variables definiendo una distribución a priori spike y slab en los coeficientes (Mitchell y Beauchamp, 1988; Ishwaran y Rao, 2005), o aplicando los métodos Lasso o Elastic net desde un enfoque bayesiano (Park y Casella, 2008; Li y Lin, 2010). Sin embargo, estos métodos solo permiten fusionar los efectos que tienden a cero con el nivel de referencia, mas no a aquellos efectos diferentes de cero pero similares al efecto de otro u otros niveles.

También existen métodos que permiten la fusión de efectos (iguales o diferentes de cero) en covariables ordinales, donde la fusión solo es posible para los efectos de niveles adyacentes. Como ejemplos de estos métodos están el Fused Lasso (Tibshirani et al., 2005), desde el enfoque frecuentista, y su respectiva versión bayesiana (Kyung et al., 2010).

Para la fusión de efectos en covariables nominales existen pocas alternativas. Desde el enfoque frecuentista, Bondell y Reich (2009) proponen una variación del Fused Lasso; mientras que Gertheiss y Tutz plantean alternativas basadas en el método Lasso con penalidades para covariables ordinales y nominales (Gertheiss y Tutz, 2009; Tutz y Gertheiss, 2016). Desde el enfoque bayesiano se tienen algunas propuestas como la de Dellaportas y Tarantola (2005), que analiza la dependencia de variables categóricas en modelos log-lineales, y Malsiner-Walli et al. (2018), que definen como distribución a priori de los coeficientes de una covariable cualitativa a una mixtura finita de normales. Pauger y Wagner (2019) presentan otra alternativa desde un enfoque bayesiano, en la cual asignan una distribución a priori normal multivariada a los coeficientes de una covariable cualitativa, cuya matriz de covarianza es construida de

tal manera que induce a que niveles con efectos similares tengan una alta correlación y, por lo tanto, puedan fusionarse.

La metodología presentada por Pauger y Wagner (2019) se basa en un modelo de regresión lineal, cuya variable respuesta se distribuye como una normal. En el presente trabajo se adaptará dicha metodología para que se ajuste a un modelo de regresión con respuesta positiva.

1.2. Objetivos

El objetivo general de la tesis es analizar, implementar y aplicar una metodología bayesiana que permita identificar la equivalencia de los efectos de un conjunto de niveles de una covariable cualitativa sobre una respuesta positiva en un modelo de regresión.

De manera más específica, se buscará:

- Explorar y analizar el método propuesto en Pauger y Wagner (2019) para la fusión de categorías en covariables cualitativas.
- Plantear una adaptación de la propuesta de Pauger y Wagner (2019) para que sea aplicable en modelos de regresión con respuesta positiva.
- Implementar un método de inferencia bayesiana para el modelo usando algoritmos MCMC en el programa JAGS.
- Realizar un estudio de simulación para analizar el desempeño del modelo según el valor de los hiperparámetros.
- Aplicar la metodología a un conjunto de datos reales para modelar el ingreso por hora de la población peruana de la macro región sur y evaluar la posible existencia de una brecha salarial por etnicidad.

1.3. Organización del trabajo

En el capítulo 2 se analizará el método de fusión de efectos propuesto por Pauger y Wagner (2019) y se adaptará a un modelo de regresión con respuesta positiva bajo una distribución gamma. Se definirá la estructura que debe tener el modelo y la forma de las distribuciones a priori para los coeficientes de las covariables cualitativas y sus respectivos hiperparámetros. Finalmente, se explicará cómo el método ayuda a cumplir con el objetivo de fusionar niveles con efectos similares.

El capítulo 3 explicará el método de estimación de los parámetros e hiperparámetros del modelo: primero se buscará la distribución a posteriori del modelo propuesto y luego se describirá brevemente a los métodos MCMC como una alternativa para inferencia bayesiana y su implementación en el programa JAGS.

En el capítulo 4 se presentará un método para la selección del modelo final mediante la optimización de la función de pérdida de Binder sobre las probabilidades de fusión a posteriori de los efectos.

El capítulo 5 buscará, por medio de simulaciones, entender las propiedades y el funcionamiento del modelo, enfocado en el efecto de los hiperparámetros sobre el desempeño del modelo estimado. Así mismo, se buscarán cuáles son los valores más adecuados para estos hiperparámetros.

En el capítulo 6 se utilizará un caso real para ilustrar la metodología descrita. Se tomó como referencia los datos utilizados por Baca (2019) para estimar el ingreso por hora de los habitantes peruanos de la macro región sur en función de distintas covariables. Se estimará si alguna covariable cualitativa tiene categorías cuyos efectos puedan ser fusionados, poniendo énfasis en la etnicidad para evaluar la existencia de una brecha salarial.

Finalmente, en el capítulo 7 se expondrán los principales hallazgos del presente trabajo. Se discutirán las ventajas y debilidades de la metodología propuesta, y también se presentarán algunas recomendaciones para futuros trabajos referentes a la fusión de efectos.

En los anexos se presentarán algunas demostraciones, resultados en mayor detalle e información complementaria que aporte en el desarrollo y entendimiento del presente trabajo.

Capítulo 2

Modelo de fusión de efectos para covariables cualitativas

El método presentado en Pauger y Wagner (2019) permite estimar si las categorías de una covariable cualitativa pueden ser fusionadas o no, en un modelo de regresión con respuesta normal. Para adaptar dicho modelo a uno con respuesta positiva, se tomará de referencia el trabajo de Sal y Rosas et al. (2019) para modelar los ingresos del personal de salud, en el cual consideran una distribución gamma para la variable respuesta.

Para poder explicar mejor el método, se planteará un caso sencillo considerando solo una covariable cualitativa; posteriormente el modelo puede ser generalizado para un mayor número de covariables.

2.1. Especificación del modelo

Se considera un modelo de regresión con variable respuesta Y_i $(i = 1, 2, ..., n)$ de distribución gamma con media μ_i y parámetro de precisión ω:

$$f(y_i) = \frac{(\omega/\mu_i)^{\omega}\, y_i^{\omega-1} \exp\{-y_i\omega/\mu_i\}}{\Gamma(\omega)}, \quad y_i > 0,$$

que será denotado por $Y_i \sim Gamma(\mu_i, \omega)$; y una covariable cualitativa Z_i con $c+1$ niveles, la cual puede ser nominal u ordinal. Para representar los efectos de la covariable se tomará al primer nivel $(k = 0)$ como referencia y definiremos para cada nivel restante $(k = 1, ..., c)$ una variable indicadora $X_{k,i}$, a través de la cual se determinará el efecto respectivo de cada nivel k cuando $k \geq 1$. El modelo quedará planteado de la siguiente manera:

- Componente aleatorio:

$$Y_i \sim Gamma(\mu_i, \omega), \tag{2.1}$$

- Componente sistemático:

$$\eta_i = \beta_0 + \sum_{k=1}^{c} \beta_k X_{k,i},$$

- Función de enlace:

$$\log(\mu_i) = \eta_i,$$

donde η_i es el predictor lineal, β_0 es el intercepto y β_k es el efecto fijo del k-ésimo nivel de la covariable cualitativa Z.

2.2. Distribución a priori para fusión de efectos

Para poder realizar la inferencia bayesiana del modelo, primero es necesario asignar distribuciones a priori a todos los parámetros que se consideren variables aleatorias. Escoger una distribución a priori no es una tarea irrelevante, pues se debe tener en cuenta las características subyacentes que se quieran controlar en el modelo a través de dicha distribución.

En este caso, el principal objetivo es saber si dos o más niveles de una covariable cualitativa pueden fusionarse o no. En términos del modelo, se puede definir un conjunto de variables aleatorias condicionalmente independientes a partir de las diferencias entre cada par de efectos de una covariable $\theta_{kj} = \beta_k - \beta_j$, para $0 \leq j < k \leq c$, a las cuales Pauger y Wagner (2017) les asigna una distribución a priori spike y slab de la siguiente forma:

$$\theta_{kj}|\delta_{kj}, \tau^2 \sim \delta_{kj} N(0, \tau^2\gamma) + (1 - \delta_{kj}) N\left(0, \frac{1}{r}\tau^2\gamma\right), \tag{2.2}$$

donde τ^2 es un hiperparámetro que determina la magnitud de la varianza de la diferencia de efectos de dos niveles de una covariable cualitativa θ_{kj}, γ es una constante y δ_{kj} es un hiperparámetro que indica si la diferencia de los efectos tiende a cero (spike), por lo que ambos niveles podrían ser fusionados, o si la diferencia de los efectos de los niveles es significativa (slab), por lo que los efectos deberían ser considerados diferentes:

$$\delta_{kj} = \begin{cases} 1\,, & \text{si } \beta_k \text{ y } \beta_j \text{ tienen efectos diferentes,} \\ 0\,, & \text{si } \beta_k \text{ y } \beta_j \text{ tienen efectos similares,} \end{cases} \tag{2.3}$$

r es una constante que reduce la varianza de la diferencia de los efectos cuando ésta tiende a cero ($\delta_{kj} = 0$), por lo que debe tomar valores relativamente altos para definir adecuadamente la parte spike de la distribución a priori. Además, en la subsección 2.2.1, se mostrará que r está relacionada con la correlación parcial de los efectos de cada par de niveles que pueden fusionarse.

Si se define $\kappa_{kj} = \delta_{kj} + r(1 - \delta_{kj})$, entonces (2.2) podría representarse de una forma más compacta:

$$\theta_{kj}|\delta_{kj}, \tau^2 \sim N\left(0, \frac{1}{\kappa_{kj}}\tau^2\gamma\right). \tag{2.4}$$

A partir de (2.4) se puede demostrar que el vector de efectos de la covariable cualitativa $\boldsymbol{\beta} = (\beta_1, ..., \beta_c)^\intercal$ tiene una distribución a priori normal multivariada (ver apéndice A):

$$\boldsymbol{\beta} \mid \boldsymbol{\delta}, \tau^2 \sim N_c(\mathbf{0}, \mathbf{B}(\tau^2, \boldsymbol{\delta})), \tag{2.5}$$

donde la matriz de covarianza de los efectos, definida por

$$\mathbf{B}(\boldsymbol{\delta}, \tau^2) = \gamma\tau^2\mathbf{Q}^{-1}(\boldsymbol{\delta}), \tag{2.6}$$

depende de los hiperparámetros $\boldsymbol{\delta}$ y τ^2. El vector $\boldsymbol{\delta}$ contiene a los indicadores binarios δ_{kj} que

definen la equivalencia o diferencia de los efectos para cada par de niveles β_k y β_j (2.3), y la matriz $\mathbf{Q}(\boldsymbol{\delta})$ determina la estructura de la precisión a priori de los coeficientes de regresión, cuyos elementos dependerán de las restricciones que la covariable pueda tener. Concretamente, la estructura de la matriz $\mathbf{Q}(\boldsymbol{\delta})$ dependerá de qué pares de efectos puedan ser fusionados o no, lo cual a su vez depende del tipo de covariable cualitativa que se esté estudiando: en las covariables nominales no existe ningún tipo de restricción en la estructura de la matriz de precisión; es decir, cualquier par de efectos podría ser fusionado. Por otro lado, en las covariables ordinales existe la restricción de que solo los efectos adyacentes pueden ser fusionados.

2.2.1. Estructura para la fusión de efectos sin restricciones

Como se mencionó en (2.3), el vector $\boldsymbol{\delta}$ está compuesto por los indicadores binarios δ_{kj} para cada par de efectos, donde $0 \leq j < k \leq c$. Al no haber restricciones, dicho vector definirá la estructura de la matriz de precisión $\mathbf{Q}(\boldsymbol{\delta})$, simétrica y de dimensiones $c \times c$:

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1c} \\ q_{21} & q_{22} & \cdots & q_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ q_{c1} & q_{c2} & \cdots & q_{cc} \end{pmatrix},$$

donde los elementos de $\mathbf{Q}(\boldsymbol{\delta})$ se especifican de la siguiente manera:

$$q_{kj} = \begin{cases} -\kappa_{kj} & , \text{ si } k \neq j, \\ \sum_{j \neq k} \kappa_{kj} = \underbrace{\kappa_{k0} + \kappa_{k1} + ... + \kappa_{k,j-1} + \kappa_{k,j+1} + ... + \kappa_{kc}}_{\text{c términos}} & , \text{ si } k = j, \end{cases} \tag{2.7}$$

quedando la matriz $\mathbf{Q}(\boldsymbol{\delta})$ definida de la siguiente forma:

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} \sum_{j \neq 1} \kappa_{1j} & -\kappa_{12} & \cdots & -\kappa_{1c} \\ -\kappa_{21} & \sum_{j \neq 2} \kappa_{2j} & \cdots & -\kappa_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ -\kappa_{c1} & -\kappa_{c2} & \cdots & \sum_{j \neq c} \kappa_{cj} \end{pmatrix},$$

donde los elementos κ_{kj} son determinados por:

$$\kappa_{kj} = \kappa_{jk} = \delta_{kj} + r(1 - \delta_{kj}). \tag{2.8}$$

De esta manera, cuando dos niveles tengan efectos diferentes ($\delta_{kj} = 1$), se tiene el valor de $\kappa_{kj} = 1$; y cuando los efectos de dos niveles pueden ser fusionados ($\delta_{kj} = 0$), entonces $\kappa_{kj} = r$. El valor de r describe la correlación entre dos niveles que pueden ser fusionados, lo que refuerza el hecho de que debe tener un valor relativamente elevado. A r se le denominará el ratio de precisión.

Ratio de precisión (r)

De (2.6), se puede ver que la matriz de estructura $\mathbf{Q}(\boldsymbol{\delta})$ debe ser una matriz definida positiva y, por lo tanto, invertible (ver prueba en Pauger y Wagner, 2017).

Los elementos fuera de la diagonal de la matriz de precisión $\mathbf{Q}(\boldsymbol{\delta})$ brindan información sobre la correlación parcial a priori entre dos efectos β_k y β_j, tal que $k \neq j$:

$$Cor(\beta_k, \beta_j \mid \boldsymbol{\beta_{/kj}}) = -\frac{q_{kj}}{\sqrt{q_{kk}q_{jj}}} = \frac{\kappa_{kj}}{\sqrt{q_{kk}q_{jj}}}, \tag{2.9}$$

donde $\boldsymbol{\beta_{/kj}}$ indica que los efectos diferentes de β_k y β_j se mantienen constantes. Los elementos de la diagonal de la matriz de precisión $\mathbf{Q}(\boldsymbol{\delta})$ determinan la precisión parcial a priori del efecto β_k, dado el valor de los efectos restantes, de la siguiente manera:

$$Prec(\beta_k \mid \boldsymbol{\beta_{/k}}) = \frac{q_{kk}}{\gamma\tau^2}, \tag{2.10}$$

donde $\boldsymbol{\beta_{/k}}$ indica que los efectos diferentes de β_k se mantienen constantes (ver apéndice B).

En (2.9) se puede ver que el valor de las correlaciones parciales a priori entre dos niveles β_k y β_j dependerá del indicador binario δ_{kj}, pues tendrá valores altos (proporcionales a r) cuando $\delta_{kj} = 0$ (puede haber fusión) o tendrá valores bajos cuando $\delta_{kj} = 1$ (efectos diferentes).

Como se observa en (2.10), el valor de la precisión parcial a priori depende del valor del elemento de la diagonal q_{kk} (2.7), el cual a su vez es la suma de c términos que dependen del los indicadores binarios correspondientes a determinado nivel j (2.8). En ese sentido, se puede encontrar que el mínimo valor de $Prec(\beta_j \mid \boldsymbol{\beta_{/j}})$ es $\frac{c}{\gamma\tau^2}$, cuando todos los términos $\delta_{kj} = 1$ (para $k \neq j$). Del mismo modo, se puede encontrar que el máximo valor de $Prec(\beta_j \mid \boldsymbol{\beta_{/j}})$ es $r\frac{c}{\gamma\tau^2}$, cuando todos los términos $\delta_{kj} = 0$. De aquí, r viene a ser el ratio entre la precisión parcial a priori máxima y mínima.

Por otro lado, cabe señalar que Pauger y Wagner (2019) fijan el valor $\gamma = c/2$, de tal manera que la precisión parcial a priori no dependa del número de niveles (c) y tenga valores entre $\frac{2}{\tau^2}$ y $r\frac{2}{\tau^2}$.

Caso ilustrativo con matriz de estructura sin restricciones

Consideremos una covariable con siete niveles ($c = 6$) y valor de la constante $\gamma = 3$. Supongamos que al ratio de precisión r se le asigna un valor arbitrario relativamente grande $r = 10000$ y que los efectos β_2 y β_5 de la covariable puedan ser fusionados, es decir, la distribución a priori conjunta de (β_2, β_5) se encuentra concentrada cerca de $\beta_2 = \beta_5$. Entonces, a partir de (2.8), los elementos del vector $\boldsymbol{\delta}$ quedarán definidos como se muestra en el cuadro 2.1:

(kj)	δ_{kj}	κ_{kj}
(10)	1	$1 + 10000(1-1) = 1$
(20)	1	$1 + 10000(1-1) = 1$
(30)	1	$1 + 10000(1-1) = 1$
(40)	1	$1 + 10000(1-1) = 1$
(50)	1	$1 + 10000(1-1) = 1$
(60)	1	$1 + 10000(1-1) = 1$
(21)	1	$1 + 10000(1-1) = 1$
(31)	1	$1 + 10000(1-1) = 1$
(41)	1	$1 + 10000(1-1) = 1$
(51)	1	$1 + 10000(1-1) = 1$
(61)	1	$1 + 10000(1-1) = 1$
(32)	1	$1 + 10000(1-1) = 1$
(42)	1	$1 + 10000(1-1) = 1$
(52)	0	$0 + 10000(1-0) = 10000$
(62)	1	$1 + 10000(1-1) = 1$
(43)	1	$1 + 10000(1-1) = 1$
(53)	1	$1 + 10000(1-1) = 1$
(63)	1	$1 + 10000(1-1) = 1$
(54)	1	$1 + 10000(1-1) = 1$
(64)	1	$1 + 10000(1-1) = 1$
(65)	1	$1 + 10000(1-1) = 1$

Cuadro 2.1: Cálculo de δ_{kj} y κ_{kj} para el ejemplo sin restricciones.

Luego, en base a (2.7) se pueden calcular los elementos de la diagonal de la matriz $\mathbf{Q}(\boldsymbol{\delta})$:

$$
\begin{aligned}
q_{11} &= \kappa_{10} + \kappa_{12} + \kappa_{13} + \kappa_{14} + \kappa_{15} + \kappa_{16} = 1+1+1+1+1+1 = 6 \\
q_{22} &= \kappa_{20} + \kappa_{21} + \kappa_{23} + \kappa_{24} + \kappa_{25} + \kappa_{26} = 1+1+1+1+10000+1 = 10005 \\
q_{33} &= \kappa_{30} + \kappa_{31} + \kappa_{32} + \kappa_{34} + \kappa_{35} + \kappa_{36} = 1+1+1+1+1+1 = 6 \\
q_{44} &= \kappa_{40} + \kappa_{41} + \kappa_{42} + \kappa_{43} + \kappa_{45} + \kappa_{46} = 1+1+1+1+1+1 = 6 \\
q_{55} &= \kappa_{50} + \kappa_{51} + \kappa_{52} + \kappa_{53} + \kappa_{54} + \kappa_{56} = 1+1+10000+1+1+1 = 10005 \\
q_{66} &= \kappa_{60} + \kappa_{61} + \kappa_{62} + \kappa_{63} + \kappa_{64} + \kappa_{65} = 1+1+1+1+1+1 = 6.
\end{aligned}
$$

Finalmente, la estructura de la matriz de precisión a priori quedará definida por:

$$
\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix}
6 & -1 & -1 & -1 & -1 & -1 \\
-1 & 10005 & -1 & -1 & -10000 & -1 \\
-1 & -1 & 6 & -1 & -1 & -1 \\
-1 & -1 & -1 & 6 & -1 & -1 \\
-1 & -10000 & -1 & -1 & 10005 & -1 \\
-1 & -1 & -1 & -1 & -1 & 6
\end{pmatrix}.
$$

Si además se considera que la distribución marginal a priori de β_4 está concentrada en $\beta_4 = 0$, es decir β_4 podría ser fusionada con la categoría de referencia, la matriz $\mathbf{Q}(\boldsymbol{\delta})$ quedaría de la siguiente forma:

$$\delta_{04} = 0 \Rightarrow \kappa_{04} = 10000,$$

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} 6 & -1 & -1 & -1 & -1 & -1 \\ -1 & 10005 & -1 & -1 & -10000 & -1 \\ -1 & -1 & 6 & -1 & -1 & -1 \\ -1 & -1 & -1 & 10005 & -1 & -1 \\ -1 & -10000 & -1 & -1 & 10005 & -1 \\ -1 & -1 & -1 & -1 & -1 & 6 \end{pmatrix}.$$

Asumiendo el valor del hiperparámetro $\tau^2 = 9$, de (2.6) la matriz de covarianza sería:

$$\mathbf{B}(\tau^2, \boldsymbol{\delta}) = \begin{pmatrix} 5.79 & 1.93 & 1.93 & 0 & 1.93 & 1.93 \\ 1.93 & 3.86 & 1.93 & 0 & 3.86 & 1.93 \\ 1.93 & 1.93 & 5.79 & 0 & 1.93 & 1.93 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1.93 & 3.86 & 1.93 & 0 & 3.86 & 1.93 \\ 1.93 & 1.93 & 1.93 & 0 & 1.93 & 5.79 \end{pmatrix}.$$

Como se observa, la varianza y las covarianzas de β_4 se aproximan a cero, lo que implica que β_4 tiende a ser una constante con valor igual a cero, por lo que su efecto se fusionaría con el nivel de referencia. Por otro lado, la covarianza entre β_2 y β_5 es igual a sus respectivas varianzas debido a que aproximadamente $\beta_2 = \beta_5$, por lo que ambas categorías podrían ser fusionadas.

A partir de la matriz de covarianza es posible calcular la matriz de correlación, la cual se está representando mediante el diagrama de calor de la figura 2.1. Se puede observar que la correlación entre β_4 y los demás efectos es prácticamente nula; además, la correlación entre β_2 y β_5 es casi perfecta.

2.2.2. Estructura para la fusión de efectos con restricciones

Pueden existir diversos casos en los que la fusión de efectos solo es posible en algunos pares específicos de niveles, según la naturaleza de la covariable. Para indicar que dos niveles no tienen posibilidad de ser fusionados, por ejemplo β_k y β_j, se podría fijar el valor del respectivo indicador en $\delta_{kj} = 1$ o asignar el valor de cero al correspondiente elemento de la matriz $\mathbf{Q}(\boldsymbol{\delta})$. A la primera opción ($\delta_{kj} = 1$) se le denomina restricción suave, la cual indica que aún podría existir cierta relación entre β_k y β_j; mientras que al segundo caso ($q_{kj} = 0$) se le denomina restricción dura, pues implica que exista independencia condicional entre β_k y β_k (Rue y Held, 2005).

La implementación de una restricción suave es directa, siguiendo las pautas de la subsección anterior; sin embargo, la restricción dura involucra alterar la estructura de la matriz $\mathbf{Q}(\boldsymbol{\delta})$, al vector $\boldsymbol{\delta}$ y a la constante γ. Para la restricción dura, será necesario especificar un nuevo vector $\boldsymbol{\zeta}$, el cual indica si existe ($\zeta_{kj} = 1$) o no ($\zeta_{kj} = 0$) la posibilidad de evaluar la fusión para cada par de niveles de una covariable. En ese sentido, solo será necesario definir

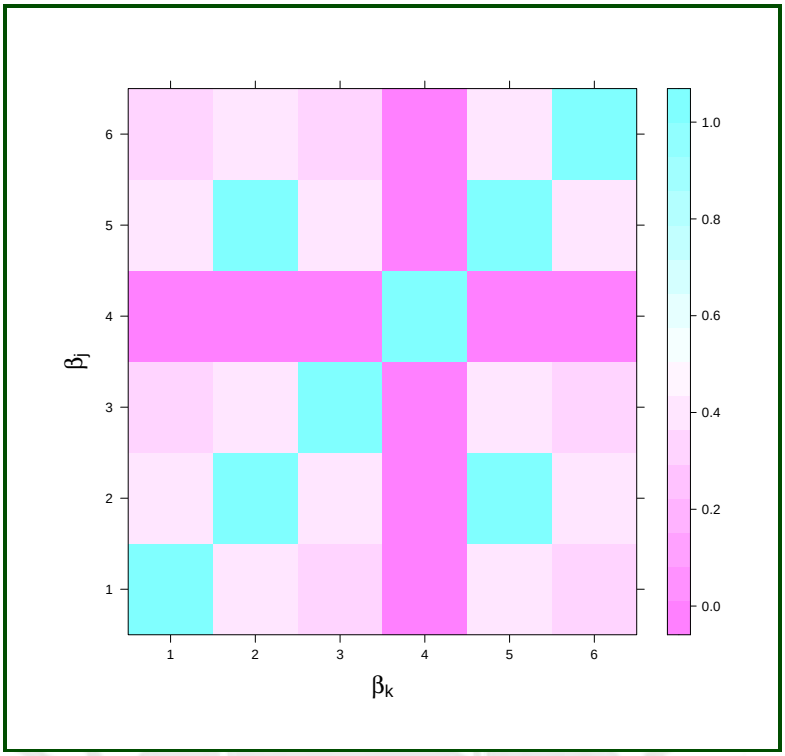


Figura 2.1: Correlaciones (β_k, β_j) para el ejemplo sin restricciones.

los valores de δ_{kj} para aquellos pares de niveles donde $\zeta_{kj} = 1$.

Los valores de la matriz de precisión ubicados fuera de la diagonal se definirían de la siguiente forma (para $k \neq j$):

$$q_{kj} = \begin{cases} -\kappa_{kj}\,, & \text{si } \zeta_{kj} = 1, \\ 0 \quad\;\; , & \text{si } \zeta_{kj} = 0, \end{cases}$$

y los elementos de la diagonal, de la siguiente manera:

$$q_{kk} = \begin{cases} \kappa_{k0} - \sum_{k \neq j} q_{kj}\,, & \text{si } \zeta_{k0} = 1, \\ -\sum_{k \neq j} q_{kj} \quad\;\; , & \text{si } \zeta_{k0} = 0. \end{cases} \tag{2.11}$$

Caso ilustrativo para covariables ordinales

Las covariables ordinales son un caso particular de restricción, pues por su naturaleza solo se puede evaluar la posibilidad de fusión en niveles adyacentes (Gertheiss y Tutz, 2009). En este caso, los elementos del vector $\boldsymbol{\zeta}$ se definirán como:

$$\zeta_{kj} = \begin{cases} 1\,, & \text{si } j = k - 1, \\ 0\,, & \text{caso contrario.} \end{cases} \tag{2.12}$$

Por lo tanto, el vector $\boldsymbol{\delta}$ tendrá c elementos $(\delta_{01}, ..., \delta_{c-1,c})$ y la matriz $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ sería de la

siguiente forma:

$$\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta}) = \begin{pmatrix} \kappa_{10} + \kappa_{12} & -\kappa_{12} & 0 & \cdots & 0 & 0 \\ -\kappa_{21} & \kappa_{21} + \kappa_{23} & -\kappa_{23} & \cdots & 0 & 0 \\ 0 & -\kappa_{32} & \kappa_{32} + \kappa_{34} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \kappa_{c-1,c-2} + \kappa_{c-1,c} & -\kappa_{c-1,c} \\ 0 & 0 & 0 & \cdots & -\kappa_{c,c-1} & \kappa_{c,c-1} \end{pmatrix}.$$

Como se observa en (2.11), para este caso el máximo valor del elemento de la diagonal q_{kk} es $2r$, cuando todos los términos $\delta_{kj} = 1$ (para $k \neq j$). Pauger y Wagner (2019) fijan el valor de la constante $\gamma = 1$, de tal manera que la precisión parcial a priori tenga un rango similar al caso sin restricciones.

Por ejemplo, considere una covariable ordinal con siete niveles ($c = 6$), el valor de la constante $\gamma = 1$ y que al ratio de precisión r se le asigna un valor arbitrario relativamente grande $r = 10000$. En primer lugar, debemos definir a los elementos del vector $\boldsymbol{\zeta}$ según (2.12), de tal manera que indiquen que solo es posible evaluar la fusión de efectos en niveles adyacentes. Luego, considerar que se puede fusionar el efecto β_1 con el nivel de referencia y, por otra parte, el efecto β_3 puede ser fusionado con β_4. Entonces, los valores de los vectores $\boldsymbol{\zeta}$ y $\boldsymbol{\delta}$ quedan definidos como muestra el cuadro 2.2.

(kj)	ζ_{kj}	δ_{kj}	κ_{kj}
(10)	1	0	$0 + 10000(1 - 0) = 10000$
(20)	0	-	-
(30)	0	-	-
(40)	0	-	-
(50)	0	-	-
(60)	0	-	-
(21)	1	1	$1 + 10000(1 - 1) = 1$
(31)	0	-	-
(41)	0	-	-
(51)	0	-	-
(61)	0	-	-
(32)	1	1	$1 + 10000(1 - 1) = 1$
(42)	0	-	-
(52)	0	-	-
(62)	0	-	-
(43)	1	0	$0 + 10000(1 - 0) = 10000$
(53)	0	-	-
(63)	0	-	-
(54)	1	1	$1 + 10000(1 - 1) = 1$
(64)	0	-	-
(65)	1	1	$1 + 10000(1 - 1) = 1$

Cuadro 2.2: Definición de ζ_{kj} y cálculo de δ_{kj} y κ_{kj} para el ejemplo con covariable ordinal.

A partir de (2.11), los elementos de la diagonal de la matriz $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ quedan definidos por:

$$\begin{aligned}
q_{11} &= \kappa_{10} + \kappa_{12} = 10000 + 1 = 10001 \\
q_{22} &= \kappa_{21} + \kappa_{23} = 1 + 1 = 2 \\
q_{33} &= \kappa_{32} + \kappa_{34} = 1 + 10000 = 10001 \\
q_{44} &= \kappa_{43} + \kappa_{45} = 10000 + 1 = 10001 \\
q_{55} &= \kappa_{54} + \kappa_{56} = 1 + 1 = 2 \\
q_{66} &= \kappa_{65} = 1.
\end{aligned}$$

Entonces, la estructura de la matriz de precisión a priori quedará definida de la siguiente manera:

$$\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta}) = \begin{pmatrix}
10001 & -1 & 0 & 0 & 0 & 0 \\
-1 & 2 & -1 & 0 & 0 & 0 \\
0 & -1 & 10001 & -10000 & 0 & 0 \\
0 & 0 & -10000 & 10001 & -1 & 0 \\
0 & 0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & 0 & -1 & 1
\end{pmatrix}.$$

De (2.6), asumiendo el valor del hiperparámetro $\tau^2 = 9$, la matriz de covarianza será:

$$\mathbf{B}(\tau^2, \boldsymbol{\zeta}, \boldsymbol{\delta}) = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 9 & 9 & 9 & 9 & 9 \\
0 & 9 & 18 & 18 & 18 & 18 \\
0 & 9 & 18 & 18 & 18 & 18 \\
0 & 9 & 18 & 18 & 27 & 27 \\
0 & 9 & 18 & 18 & 27 & 36
\end{pmatrix}.$$

Se observa que la varianza y covarianzas de β_1 se aproximan a cero, lo que implica que este efecto tiende a ser una constante con valor igual a cero, por lo que se fusionaría con el nivel de referencia. Por otro lado, las varianzas de β_3 y β_4 tienen valores muy cercanos a su covarianza, lo que indica que los efectos de β_3 y β_4 son aproximadamente iguales, por lo que ambas categorías también podrían ser fusionadas.

A partir de la matriz de covarianza se puede calcular la matriz de correlación, la cual se representa mediante el diagrama de calor de la figura 2.2, donde se observa que prácticamente no existe correlación entre β_1 y los demás efectos. Por otro lado, se muestra que la correlación entre β_3 y β_4 es casi perfecta.

2.3. Especificación de la distribución a priori del modelo

Basado en la propuesta de Pauger y Wagner (2019), pero considerando ahora una variable respuesta de distribución gamma, la estructura de la distribución a priori que asumiremos para el modelo será:

$$p(\omega, \beta_0, \boldsymbol{\beta}) = p(\omega)p(\beta_0)p(\boldsymbol{\beta} \mid \tau^2, \boldsymbol{\delta})p(\tau^2)p(\boldsymbol{\delta}),$$

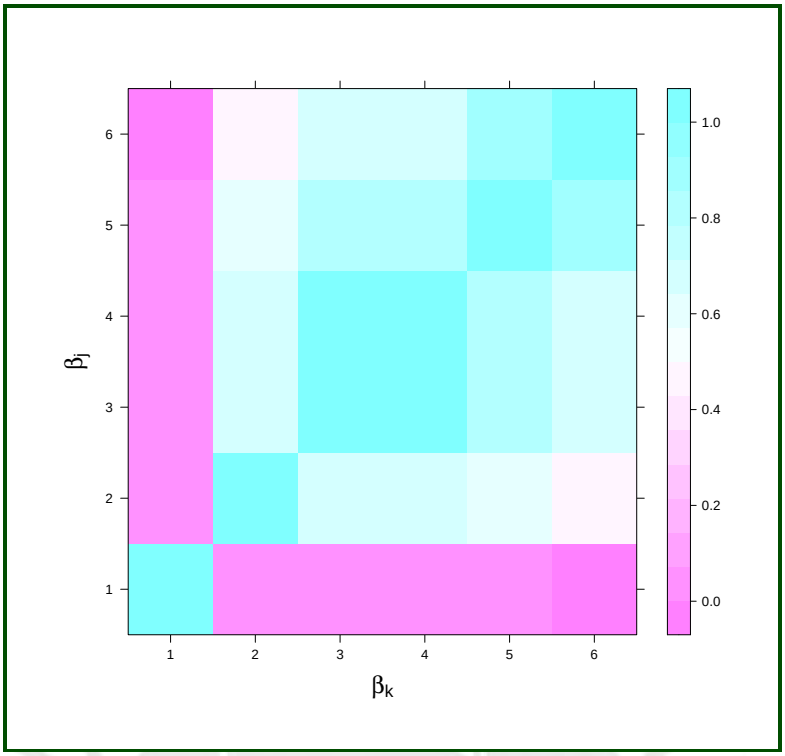


Figura 2.2: Correlaciones (β_k, β_j) para el ejemplo con covariable ordinal.

donde al parámetro de precisión ω y al intercepto β_0 se les asignarán distribuciones a priori propias $p(\omega) \sim Gamma(s_0, S_0)$ y $p(\beta_0) \sim N(0, M_0)$, respectivamente; siendo s_0, S_0 y M_0 constantes. Al vector de efectos de la covariable cualitativa de c niveles $\boldsymbol{\beta} = (\beta_1, ..., \beta_c)^\intercal$ se le asignará la distribución a priori normal multivariada definida en (2.5), donde el parámetro que describe la varianza de los efectos τ^2 tendrá como distribución a priori una gamma-inversa, definida de la siguiente manera:

$$\tau^2 \sim Inv - Gamma(g_0, G_0), \tag{2.13}$$

donde g_0 y G_0 son constantes.

2.3.1. Priori para la variable indicadora de fusión de efectos

Para definir la distribución a priori del vector $\boldsymbol{\delta}$, es usual asumir independencia condicional entre sus elementos δ_{kj}, por lo que se tienen diversas alternativas. Pauger y Wagner (2019) proponen una distribución a priori definida de la siguiente manera:

$$p(\boldsymbol{\delta}) \propto |\boldsymbol{Q}(\boldsymbol{\delta})|^{-1/2} r^{\sum_{k \neq j}(1-\delta_{kj})/2},$$

debido a que simplifica el cálculo de la priori conjunta de los efectos de regresión de la covariable y los indicadores de fusión $p(\boldsymbol{\beta}, \boldsymbol{\delta}|\tau^2)$. Esto permite reforzar las propiedades del ratio de precisión r cuando la covariable es nominal, donde la estructura del vector de indicadores de fusión de efectos $\boldsymbol{\delta}$ no tiene restricciones.

Por otro lado, las autoras también señalan que para el caso de selección de variables, a los elementos de $\boldsymbol{\delta}$ se les suele asignar una distribución a priori $p(\delta_{kj} = 1) = \lambda$, donde λ puede ser un valor fijo o se le puede asignar una distribución a priori $\lambda \sim Beta(\upsilon_0, \phi_0)$. Esta definición también es válida para fusión de efectos.

Dado que uno de los objetivos del presente trabajo es implementar la estimación de la distribución a posteriori en el programa JAGS, el cual no necesita de simplicidad en la forma analítica, se asignará a los indicadores de fusión de efecto la siguiente distribución a priori:

$$p(\delta_{kj} = 1) = \lambda \sim Beta(\upsilon_0, \phi_0), \tag{2.14}$$

donde υ_0 y ϕ_0 son constantes.

2.4. Definición de hiperparámetros para el modelo

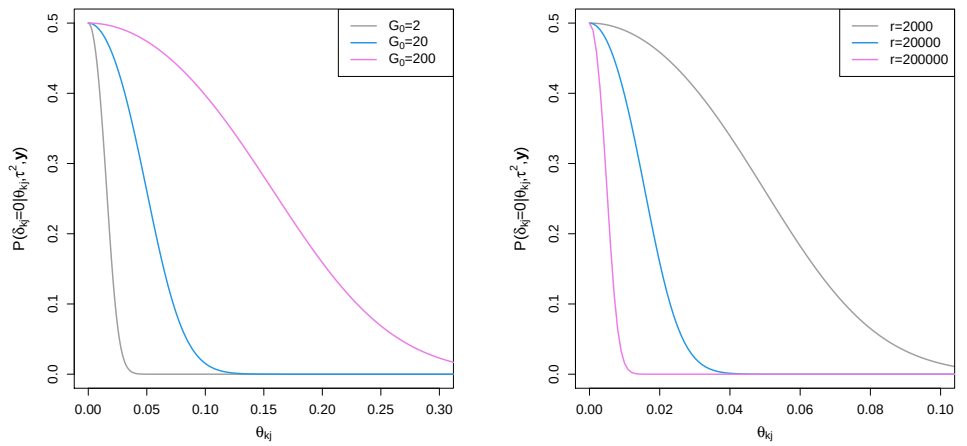
A los hiperparámetros de las distribuciones a priori de β_0, ω y λ se les podrían asignar valores de tal forma que las distribuciones a priori tiendan a ser no informativas.

La definición de los hiperparámetros de τ^2 y del ratio de precisión r debe realizarse de manera más cuidadosa, pues sus valores afectan a la distribución a priori de las diferencias de los efectos (2.2). Para este fin se utilizó de referencia la metodología presentada por Pauger y Wagner (2019), quienes fijan el valor de $g_0 = 5$ siguiendo la pauta de Fahrmeir et al. (2010) para que las colas de la spike y slab sean lo suficientemente pesadas y así evitar problemas en la confluencia de las cadenas del algoritmo MCMC al estimar δ_{kj} condicionado a θ_{kj}.

Para el caso de G_0 y r, se debe de tener en cuenta que estos hiperparámetros también afectan a la probabilidad de fusión de efectos a posteriori (3.4), lo cual es una de las principales cualidades del modelo planteado. La figura 2.3 muestra que, para valores fijos de θ_{kj}, G_0 se relaciona con la probabilidad de fusión de manera directa y r, de manera indirecta. El gráfico permite observar cuáles son los valores de los hiperparámetros que pueden ser más convenientes para evitar errores al fusionar efectos que son muy distintos o al no fusionar efectos que son muy similares. En ese sentido, pareciera que lo más conveniente es tomar valores pequeños de G_0 y tomar valores grandes de r; cuidando que los valores asignados no afecten la definición de la distribución priori de las diferencias de los efectos.

Para definir el valor de G_0, Pauger y Wagner (2019) intentaron asignarle una distribución hiperpriori exponencial con media fija; sin embargo, en sus estudios de simulación encontraron que los resultados eran equivalentes a fijar los valores de G_0. En el caso del ratio de precisión r, las autoras hicieron un planteamiento similar, encontrando que asignar una distribución hiperpriori exponencial no era muy favorable, pues tendía a fusionar todos los pares de efectos.

En el capítulo 5 se buscarán los valores más adecuados de los hiperparámetros r y G_0 para el modelo presentado en (2.1) mediante un estudio de simulación; que estará enfocado en el objetivo de evaluar la fusión de efectos en base a determinados indicadores.



(a) Diferentes valores de G_0 con $r = 20000$. (b) Diferentes valores de r con $G_0 = 5$.

Figura 2.3: Probabilidades de fusión a posteriori para distintos valores de G_0 y r con g_0=5 y $\gamma = 6$.

Capítulo 3

Método de inferencia bayesiana

El modelo de fusión de efectos presentado por Pauger y Wagner (2019) se encuentra implementado en el paquete *effectFusion* de R; sin embargo, éste se limita a trabajar solo en modelos con respuesta bajo una distribución normal o binomial.

En el presente trabajo se plantea utilizar un modelo con variable respuesta de distribución gamma, tal como se define en (2.1), de tal manera que sea apropiado para una respuesta positiva. Para el modelo, se considerará una distribución a priori definida por:

$$p(\omega, \beta_0, \boldsymbol{\beta}, \boldsymbol{\delta}, \lambda, \tau^2) = p(\omega)p(\beta_0)p(\boldsymbol{\beta}|\tau^2, \boldsymbol{\delta}, \lambda)p(\tau^2)p(\boldsymbol{\delta}|\lambda)p(\lambda), \tag{3.1}$$

donde ω es el parámetro precisión de la variable respuesta, β_0 el intercepto, $\boldsymbol{\beta}$ los efectos de la covariable cualitativa, τ^2 y $\boldsymbol{\delta}$ son los hiperparámetros de la distribución a priori spike y slab que controla la fusión de efectos, y λ es un hiperparámetro que describe al complemento de la probabilidad de fusión de efectos.

A continuación evaluaremos cuál es el método más conveniente para la inferencia del modelo planteado; ya sea analíticamente, si la distribución a posteriori tiene forma conocida, o si es necesario utilizar algún método MCMC, el cual puede ser implementado en el programa JAGS.

3.1. Distribución a posteriori

Por el teorema de Bayes (Gelman et al., 2013), la distribución a posteriori de los parámetros de un modelo satisface:

$$p(\boldsymbol{\vartheta} \mid \mathbf{y}) \propto p(\boldsymbol{\vartheta})p(\mathbf{y} \mid \boldsymbol{\vartheta}), \tag{3.2}$$

donde $\boldsymbol{\vartheta} = (\omega, \beta_0, \boldsymbol{\beta}^\intercal, \tau^2, \boldsymbol{\delta}^\intercal, \lambda)^\intercal$ son los parámetros del modelo, $p(\boldsymbol{\vartheta})$ la distribución a priori de los parámetros y $p(\mathbf{y} \mid \boldsymbol{\vartheta})$ la verosimilitud. De (2.1), la verosimilitud quedaría definida de la siguiente forma:

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\vartheta}) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\omega)} \left(\frac{\omega}{\mu_i}\right)^{\omega} y_i^{\omega-1} \exp\left\{-\left(\frac{\omega}{\mu_i}\right) y_i\right\} \\ &= \frac{1}{\Gamma(\omega)^n} \omega^{n\omega} \exp\left\{-\omega \sum_{i=1}^{n} \frac{y_i}{\mu_i}\right\} \prod_{i=1}^{n} \left(\frac{y_i^{\omega}}{y_i \mu_i^{\omega}}\right). \end{aligned}$$

Luego, de (3.1), (2.5), (2.13) y (2.14) en (3.2), tendemos la siguiente expresión para la dis-

tribución a posteriori:

$$
\begin{aligned}
p(\boldsymbol{\vartheta} \mid \mathbf{y}) \propto & \left[\frac{S_0^{s_0}}{\Gamma(s_0)}\omega^{s_0-1}\exp\{-S_0\omega\}\right]\left[\frac{1}{\sqrt{2\pi}M_0^{1/2}}\exp\left\{-\frac{\beta_0^2}{2M_0}\right\}\right] \\
& \left[\frac{1}{(2\pi)^{c/2}|\mathbf{B}(\tau^2,\boldsymbol{\delta})|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^\intercal\mathbf{B}(\tau^2,\boldsymbol{\delta})^{-1}\boldsymbol{\beta})\right\}\right] \\
& \left[\frac{G_0^{g_0}}{\Gamma(g_0)}\frac{1}{\tau^{2(g_0+1)}}\exp\left\{-\frac{G_0}{\tau^2}\right\}\right] \\
& \left[\prod_{k\neq j}\lambda^{\delta_{kj}}(1-\lambda)^{1-\delta_{kj}}\frac{\Gamma(\upsilon_0+\phi_0)}{\Gamma(\upsilon_0)\Gamma(\phi_0)}\lambda^{\upsilon_0-1}(1-\lambda)^{\phi_0-1}\right] \\
& \left[\frac{1}{\Gamma(\omega)^n}\omega^{n\omega}\exp\left\{-\omega\sum_{i=1}^{n}\frac{y_i}{e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}}}\right\}\prod_{i=1}^{n}\left(\frac{y_i^\omega}{y_i(e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}})^\omega}\right)\right] \\
\propto & \ \omega^{s_0-1+n\omega}\exp\{-S_0\omega\}\frac{1}{\Gamma(\omega)^n}\exp\left\{-\omega\sum_{i=1}^{n}\frac{y_i}{e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}}}\right\}\prod_{i=1}^{n}\left(\frac{y_i^\omega}{y_i(e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}})^\omega}\right) \\
& \exp\left\{-\frac{\beta_0^2}{2M_0}\right\}\frac{1}{(\tau^2)^{c/2+g_0+1}}\exp\left\{-\frac{1}{2}\frac{(\boldsymbol{\beta}^\intercal\mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\beta})}{\gamma\tau^2}\right\}\exp\left\{-\frac{G_0}{\tau^2}\right\} \\
& |\boldsymbol{Q}(\boldsymbol{\delta})|^{1/2}\prod_{k\neq j}\left(\lambda^{\delta_{kj}+\upsilon_0-1}(1-\lambda)^{-\delta_{kj}+\phi_0}\right). \qquad (3.3)
\end{aligned}
$$

Como se observa, la distribución a posteriori no tiene forma conocida. En este caso, conviene buscar las distribuciones condicionales completas para evaluar la posibilidad de aplicar el método del muestreador de Gibbs, un método MCMC fácil de implementar (Hoff, 2009).

3.2. Distribuciones condicionales completas

A partir de la distribución a posteriori encontrada en (3.3) se hallarán las distribuciones condicionales completas.

Distribución condicional completa para β_0

La distribución condicional completa del intercepto β_0 no tiene una forma conocida:

$$
p(\beta_0 \mid \omega, \boldsymbol{\beta}, \boldsymbol{\delta}, \lambda, \tau^2, \mathbf{y}) \propto \exp\left\{-\frac{\beta_0^2}{2M_0} - \omega\sum_{i=1}^{n}\frac{y_i}{e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}}}\right\}\prod_{i=1}^{n}\left(\frac{1}{(e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}})^\omega}\right).
$$

Cabe mencionar que también hubiera sido válido definir a β_0 como una constante.

Distribución condicional completa para $\boldsymbol{\beta}$

La distribución condicional completa del vector de efectos de la covariable $\boldsymbol{\beta}$ tampoco muestra una forma conocida:

$$
p(\boldsymbol{\beta} \mid \omega, \beta_0, \boldsymbol{\delta}, \lambda, \tau^2, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\frac{(\boldsymbol{\beta}^\intercal\mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\beta})}{\gamma\tau^2} - \omega\sum_{i=1}^{n}\frac{y_i}{e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}}}\right\}\prod_{i=1}^{n}\left(\frac{1}{(e^{\beta_0+\boldsymbol{\beta}\boldsymbol{x_i}})^\omega}\right),
$$

a diferencia del modelo con variable respuesta normal planteado por Pauger y Wagner (2019), donde la distribución condicional completa tiene forma de una normal.

Distribución condicional completa para ω

La distribución condicional completa del parámetro de precisión de la variable respuesta ω tampoco tiene una forma conocida:

$$p(\omega \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{\delta}, \lambda, \tau^2, \mathbf{y}) \propto \omega^{S_0-1+n\omega} \frac{1}{\Gamma(\omega)^n} \exp\left\{-S_0\omega - \omega \sum_{i=1}^{n} \frac{y_i}{e^{\beta_0+\boldsymbol{\beta x_i}}}\right\} \prod_{i=1}^{n} \left(\frac{y_i^{\omega}}{(e^{\beta_0+\boldsymbol{\beta x_i}})^{\omega}}\right).$$

Distribución condicional completa para $\boldsymbol{\delta}$

La distribución condicional completa del hiperparámetro que indica fusión de efectos $\boldsymbol{\delta}$ está dada por:

$$\begin{aligned} p(\boldsymbol{\delta} \mid \omega, \beta_0, \boldsymbol{\beta}, \lambda, \tau^2, \mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\frac{(\boldsymbol{\beta}^{\intercal}\mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\beta})}{\gamma\tau^2}\right\} |\boldsymbol{Q}(\boldsymbol{\delta})|^{1/2} \prod_{k\neq j} \left(\lambda^{\delta_{kj}}(1-\lambda)^{1-\delta_{kj}}\right) \\ &\propto |\boldsymbol{Q}(\boldsymbol{\delta})|^{1/2} \prod_{k\neq j} \left(\lambda^{\delta_{kj}}(1-\lambda)^{1-\delta_{kj}}\right) \exp\left(-\frac{1}{2\gamma\tau^2}(\beta_k-\beta_j)^2(\delta_{kj}+r(1-\delta_{kj}))\right). \end{aligned}$$

Luego, la distribución a priori condicional para cada indicador de fusión δ_{kj}, dado los valores de los demás indicadores $\delta_{/kj}$ para todo $0 \leq j < k \leq c$, es:

$$p(\delta_{kj} = 1 \mid \delta_{/kj}, \omega, \beta_0, \boldsymbol{\beta}, \lambda, \tau^2, \mathbf{y}) = \frac{1}{1 + L_{kj} \exp\left(-\frac{r-1}{2\gamma\tau^2}(\beta_k-\beta_j)^2\right)},$$

donde:

$$L_{kj} = \frac{|\boldsymbol{Q}(\boldsymbol{\delta}|\delta_{kj}=0)|^{1/2}}{|\boldsymbol{Q}(\boldsymbol{\delta}|\delta_{kj}=1)|^{1/2}}.$$

Finalmente, podemos notar que la distribución condicional completa del indicador δ_{kj} queda de la siguiente forma:

$$\delta_{kj} \mid \delta_{/kj}, \omega, \beta_0, \boldsymbol{\beta}, \lambda, \tau^2, \mathbf{y} \sim Bernoulli\left(\frac{1}{1 + L_{kj} \exp\left(-\frac{r-1}{2\gamma\tau^2}(\beta_k-\beta_j)^2\right)}\right). \tag{3.4}$$

Distribución condicional completa para τ^2

La distribución condicional completa del hiperparámetro τ^2 es proporcional a la siguiente expresión:

$$p(\tau^2 \mid \omega, \beta_0, \boldsymbol{\beta}, \boldsymbol{\delta}, \lambda, \mathbf{y}) \propto \left(\frac{1}{\tau^2}\right)^{g_0+c/2+1} \exp\left\{-\frac{1}{\tau^2}\left(G_0 + \frac{(\boldsymbol{\beta}^{\intercal}\mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\beta})}{2\gamma}\right)\right\},$$

la cual tiene forma de una distribución gamma-inversa:

$$\tau^2 \mid \omega, \beta_0, \boldsymbol{\beta}, \boldsymbol{\delta}, \lambda, \mathbf{y} \sim Inv - Gamma\left(g_0 + c/2, G_0 + \frac{(\boldsymbol{\beta}^\intercal \mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\beta})}{2\gamma}\right).$$

Distribución condicional completa para λ

La distribución condicional completa del hiperparámetro λ queda definida por:

$$p(\lambda \mid \omega, \beta_0, \boldsymbol{\beta}, \tau^2, \boldsymbol{\delta}, \mathbf{y}) \propto \lambda^{\sum_{k\neq j}(\delta_{kj}+\upsilon_0-1)}(1-\lambda)^{-\sum_{k\neq j}(\delta_{kj}+\phi_0)},$$

la cual tiene forma de una distribución beta:

$$\lambda \mid \omega, \beta_0, \boldsymbol{\beta}, \tau^2, \boldsymbol{\delta}, \mathbf{y} \sim Beta\left(\sum_{k\neq j}(\delta_{kj}+\upsilon_0-1)+1, (\delta_{kj}+\phi_0)+1\right).$$

Como muestran los resultados, al escoger una distribución gamma para la variable respuesta, no todas las distribuciones condicionales completas tendrán forma conocida; a diferencia del modelo con variable respuesta normal planteado por Pauger y Wagner (2019), donde todas las distribuciones condicionales completas son conocidas y es posible utilizar el método del muestreador de Gibbs. Sin embargo, aún es posible implementar dicho método utilizando internamente el algoritmo de Metrópolis Hastings o el Slice sampling para aquellos parámetros que no tienen distribución condicional completa conocida.

3.3. Inferencia con métodos MCMC en JAGS

El programa BUGS (Bayesian Inference Using Gibbs Sampling) y sus derivados integran los algoritmos Metrópolis Hastings y Slice sampling dentro del muestreador de Gibbs para la estimación de la distribución a posteriori mediante métodos MCMC (Gelman et al., 2013).

JAGS (Just Another Gibbs Sampler) es un programa de código abierto escrito en C++ que estima la distribución a posteriori de manera similar a BUGS (incluso utilizan el mismo lenguaje), con la ventaja de contar con una mayor cantidad de funciones y tener la posibilidad de construir otras que se requieran. JAGS fue diseñado para trabajar de manera integrada con R, de tal forma que resulte fácil la implementación de modelos complejos, gracias a lo intuitivo del lenguaje BUGS, en un entorno que permite un trabajo modular (Plummer et al., 2003).

Dado que el modelo para fusión de efectos con respuesta de distribución gamma no está implementado en el programa R, JAGS se presenta como una alternativa relativamente sencilla, pues basta con tener clara la estructura del modelo. Para la implementación del modelo en JAGS se utilizó la librería *dclone* de R, pues su función *jags.parfit* permite realizar los procesos en forma paralela y reducir considerablemente el tiempo de estimación. El código del modelo de fusión está disponible en el apéndice C.

Capítulo 4

Selección del modelo final

Una vez realizada la estimación del modelo por métodos MCMC, cada indicador de fusión δ_{kj} tendrá un valor entre 0 y 1, el cual definirá la probabilidad de no fusión a posteriori entre dos efectos k y j. Sin embargo, lo que se necesita es saber si debe haber o no una fusión para cada par de efectos y, posteriormente, estimar el modelo final escogido.

4.1. Función de pérdida de Binder

Para poder especificar cuáles son los niveles que pueden fusionarse o no y cumplir con el objetivo de seleccionar el modelo final, se requiere definir una función de pérdida adecuada. Pauger y Wagner (2019) sugieren utilizar un caso especial de la función de pérdida de Binder utilizada en el estudio de Lau y Green (2007) para agrupamiento de observaciones basado en modelos bayesianos. La función de pérdida de Binder penaliza el agrupamiento incorrecto de un par de elementos, lo cual ocurre cuando dos elementos no debían agruparse y son asignadas al mismo grupo, o cuando dos elementos debían agruparse y son asignadas a diferentes grupos. Para este estudio, cada elemento corresponde a un efecto, y un agrupamiento incorrecto corresponde a clasificar la diferencia de efectos como falsamente negativa, cuando dos efectos se agrupan incorrectamente, o falsamente positiva, cuando dos efectos se mantienen separados incorrectamente.

La función de pérdida de Binder es dada por:

$$\mathcal{L}(\mathbf{z}, \mathbf{z}^*) = \sum_{j \neq k} \left(\ell_1 \mathbf{I}_{\{z_k = z_j\}} \mathbf{I}_{\left\{z_k^* \neq z_j^*\right\}} + \ell_2 \mathbf{I}_{\{z_k \neq z_j\}} \mathbf{I}_{\left\{z_k^* = z_j^*\right\}} \right),$$

donde los elementos del vector $\mathbf{z}$ indican la configuración (agrupamiento o no agrupamiento) real de cada par de efectos; es decir, si el efecto j y el efecto k están asignado al mismo grupo, entonces $z_j = z_k$. Por otro lado, los elementos del vector $\mathbf{z}^*$ denotan la configuración obtenida a partir de la estimación del modelo; mientras que ℓ_1 y ℓ_2 representan los costos de clasificación incorrecta. Según Pauger y Wagner (2019), si consideramos ℓ_1=ℓ_2, el esperado de la función de pérdida de Binder a posteriori resultará:

$$E\left(\mathcal{L}(\mathbf{z}, \mathbf{z}^* | \mathbf{y})\right) = \sum_{j \neq k} |\mathbf{I}_{\left\{z_k^* = z_j^*\right\}} - \pi_{kj}|, \tag{4.1}$$

donde $\pi_{kj} = P(z_k = z_j | \mathbf{y})$ denota el elemento (kj) de la matriz de similaridad a posteriori. El estimador que optimice el esperado de la función de pérdida de Binder (acción óptima

bayesiana) puede ser determinado minimizando la siguiente expresión:

$$\sum_{j \neq k} \mathbf{I}_{\{z_k^* = z_j^*\}} \left(\frac{1}{2} - \pi_{kj} \right).$$

Para minimizar esta expresión, Lau y Green (2007) utilizan el algoritmo implementado en la función *minbinder* del paquete *mmclust* de R, el cual requiere como entrada a una matriz de similaridad.

4.2. Fusión de efectos mediante la función de pérdida de Binder

Para determinar el modelo final en base al esperado de la función de pérdida de Binder (4.1), la optimización debe ser realizada para cada covariable por separado, considerando como estimador de los elementos de la matriz de similaridad a posteriori π_{kj} al complemento de la probabilidad de no fusión a posteriori obtenida a partir de las muestras MCMC; es decir,

$$\hat{\pi}_{kj} = 1 - \frac{1}{M} \sum_{m=1}^{M} \delta_{kj}^{(m)}.$$

De esta manera, se logra encontrar una alternativa objetiva para determinar qué niveles deben ser agrupados y cuáles no.

En el caso de las covariables ordinales, debido a la estructura de la matriz de precisión, solo tendremos valores para los términos δ_{kj} tales que $|k - j| = 1$, es decir, solo para los efectos adyacentes. Para completar la matriz de similaridad a posteriori, se plantea calcular los términos faltantes en función de los resultados de todos los pares de efectos consecutivos involucrados. Por ejemplo, para cada muestra MCMC: si $\delta_{12} = 0$ y $\delta_{23} = 0$, entonces definiremos $\delta_{13} = 0$, en caso contrario $\delta_{13} = 1$.

Por último, luego de definir los agrupamientos correspondientes, se estimará el modelo final considerando a cada nivel o grupo de niveles (según corresponda) como una variable indicadora (dummy) con una distribución a priori normal no informativa $N(0, B_0)$, donde B_0 es una constante.

Capítulo 5

Estudio de simulación

A fin de evaluar el desempeño del modelo y buscar los valores más adecuados para los hiperparámetros G_0 y r, se utilizará como referencia el estudio de simulación presentado por Pauger y Wagner (2019). Para ello, se compara el agrupamiento de niveles de un modelo teórico con el resultado del modelo final seleccionado a partir de un conjunto de datos simulados.

5.1. Configuración de las simulaciones

Se generaron aleatoriamente 100 conjuntos de datos, cada uno con una dimensión de $n = 500$ observaciones que se ajusten al modelo de regresión gamma definido en (2.1) con parámetro $\omega = 20$, intercepto $\beta_0 = 0$ y una matriz de diseño fija. Se consideraron cuatro covariables ordinales y cuatro covariables nominales, donde cada tipo de covariable está conformado por dos covariables de ocho niveles y dos de cuatro niveles. Los coeficientes de regresión teóricos fueron definidos por $\beta_1 = (0, 1, 1, 2, 2, 4, 4)^\intercal$, $\beta_2 = (0, 0, 0, 0, 0, 0, 0)^\intercal$, $\beta_3 = (0, -2, 2)^\intercal$ y $\beta_4 = (0, 0, 0)^\intercal$ para las covariables ordinales; mientras que $\beta_5 = (0, 1, 1, 1, 1, -2, -2)^\intercal$, $\beta_6 = (0, 0, 0, 0, 0, 0, 0)^\intercal$, $\beta_7 = (0, 2, 2)^\intercal$ y $\beta_8 = (0, 0, 0)^\intercal$, para las covariables nominales. Los niveles de las variables predictoras (indicadoras o dummy) fueron generados con probabilidades de $(0.1, 0.1, 0.2, 0.05, 0.2, 0.1, 0.2, 0.05)$ y $(0.1, 0.4, 0.2, 0.3)$ para las covariables con ocho y cuatro niveles, respectivamente.

Para configurar el modelo de fusión de efectos, basados en lo mencionado en la subsección 2.4, se especificó una distribución a priori normal con varianza $M_0 = 10000$ para el intercepto; mientras que para el parámetro ω se definió una distribución a priori gamma con parámetros $s_0 = 0.001$ y $S_0 = 0.001$. Para el hiperparámetro que describe la varianza de los efectos τ^2, se especificó una distribución a priori gamma-inversa con $g_0 = 5$ fijo; mientras que para G_0, se buscará el valor más adecuado para el modelo comparando los resultados para ciertos valores fijos $(20, 100, 200, 300)$ o asignando una distribución hiperpriori Exponencial con valor esperado $E(G_0) = 20$. Por último, buscaremos el valor óptimo del ratio de precisión r asignándole valores fijos $(200, 2000, 10000, 20000)$ y evaluando con cuál de ellos se tiene un mejor desempeño.

Para las estimaciones por MCMC en JAGS, se consideraron 5000 iteraciones para el proceso de adaptación, paso previo en el cual se calibran ciertos parámetros para generar cadenas más eficientes. Luego se procedió a generar cuatro cadenas, descartando las 15000 primeras simulaciones en cada una. Finalmente, se tomaron en cuenta a las siguientes 10000

iteraciones (en cada cadena) para las estimaciones.

Se debe considerar que cuando la variable respuesta se distribuye como una gamma, el proceso de adaptación y generación de cadenas suele ser más costoso que en un modelo de regresión con variable respuesta normal.

5.2. Calibración de hiperparámetros

Como se mencionó en la sección anterior, se generaron 100 simulaciones para cada valor a evaluar de los hiperparámetros G_0 y r.

Para comparar los resultados, se consideró un análogo a la matriz de confusión en la que se comparan los indicadores de fusión teóricos δ'_{kj}, planteados al inicio de la simulación, contra los valores producto del modelo de fusión y su respectiva selección del modelo mediante la función de pérdida de Binder. La definición detallada de los indicadores que se desprenden a partir de la matriz de confusión está disponible en el apéndice D; sin embargo, hay dos indicadores en los cuales se enfocará la evaluación: la tasa de falsos positivos (FPR) y la tasa de falsos negativos (FNR). Los falsos positivos ocurren cuando dos efectos que debieron fusionarse no lo hacen, lo cual puede que no traiga muchas complicaciones en el desempeño del modelo; sin embargo, se estaría considerando un modelo más complejo y, en consecuencia, menos eficiente en términos computacionales. Por otro lado, los falsos negativos corresponden a los efectos que debieron estar separados pero que al final fueron fusionados, lo cual puede significar que nuestro modelo dé resultados sesgados y con ciertas deficiencias predictivas. En ese sentido, el objetivo debe ser priorizar la reducción de la tasa de falsos negativos, manteniendo a la reducción de falsos positivos en un segundo plano.

Los resultados del cuadro 5.1 muestran las tasas FNR y FPR obtenidas para distintos valores de G_0 fijando el valor $r = 20000$. La tabla indica que se tienen buenos resultados para la tasa FNR en valores de G_0 entre 20 y 300; sin embargo, se puede observar que a mayor valor de G_0 la tasa FPR disminuye, tanto en covariables ordinales (h:$1-4$) como en nominales (h:$5-8$). Notar que el incremento de la tasa FPR para valores pequeños de G_0 es mucho mayor en las covariables nominales; sobre todo en la covariable de ocho niveles sin efecto $h = 6$. Para encontrar el valor más adecuado del hiperparámetro G_0 se debe buscar reducir las tasas FNR y FPR, procurando tomar valores pequeños de G_0 debido a su relación con la probabilidad de fusión y con la precisión de la diferencia de efectos, como se vio en la sección 2.4. En ese sentido, se escogió el valor de $G_0 = 100$, pues solo muestra una tasa FPR relativamente alta para la covariable $h = 6$; sin embargo, debemos tener en cuenta que el indicador principal es la tasa FNR y que la tasa FPR va en un segundo plano, por lo que un $FPR_6 = 8.39\,\%$ aún es aceptable. Por otro lado, al igual que en el artículo de Pauger y Wagner (2019), se puede observar que definir una hiperpriori exponencial con $E(G_0) = 20$ tiene un desempeño similar a fijar el valor de $G_0 = 20$.

El cuadro 5.2 muestra los resultados de las tasas FNR y FPR para distintos valores del ratio de precisión r fijando el valor de $G_0 = 100$. En esta tabla se puede apreciar que la tasa FNR es muy buena para cualquier valor de r entre 200 y 20000; mientras que la tasa FPR suele aumentar conforme se incrementa el valor de r. Además, se observa que el incremento de la tasa FPR para covariables ordinales (h:$1-4$) es bastante pequeño,

	$G_0 = 20$		$G_0 = 100$		$G_0 = 200$		$G_0 = 300$		$G_0 \sim \mathcal{E}(1/20)$	
h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h
1	0	1	0	0.25	0	0	0	0	0	0.75
2	-	0	-	0	-	0	-	0	-	1.57
3	0	0	0	0	0	0	0	0	0	0
4	-	0	-	0	-	0	-	0	-	1
5	0	1.62	0	0	0	0	0	0	0	1.25
6	-	18.04	-	8.39	-	3.36	-	2.61	-	18.64
7	0	0.5	0	0	0	0	0	0	0	0
8	-	0	-	0	-	0	-	0	-	0

Cuadro 5.1: Resultados para $r = 20000$ y distintos valores de G_0 (escenario 1).

mientras que para las covariables nominales (h:$5-8$), el incremento es mayor. Considerando que debemos buscar valores altos de r para evitar problemas de convergencia y debido a su relación con la probabilidad de fusión (ver sección 2.4), podríamos concluir que se deberían considerar valores entre $r = 10000$ y $r = 20000$, pues en este intervalo los niveles de las tasas FNR y FPR aún son bastante aceptables.

	$r = 200$		$r = 2000$		$r = 10000$		$r = 20000$	
h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h
1	0	0	0	0	0	0.25	0	0.25
2	-	0	-	0	-	0	-	0
3	0	0	0	0	0	0	0	0
4	-	0	-	0	-	0	-	0
5	0	0	0	0	0	0	0	0
6	-	0	-	0.25	-	3.82	-	9.07
7	0	0	0	0	0	0	0	0
8	-	0	-	0	-	0	-	0

Cuadro 5.2: Resultados para $G_0 = 100$ y distintos valores de r (escenario 1).

Cabe mencionar que para valores más pequeños de G_0 y más grandes de r (mayor precisión) que los considerados en la presente simulación se encontraron problemas de convergencia. Por otro lado, en escenarios con diferentes valores del parámetro ω, la convergencia de los parámetros también podría verse afectada.

5.3. Simulación de modelos de baja precisión

Para evaluar el comportamiento de los hiperparámetros dentro de un modelo gamma con un parámetro ω con valores bajos, similar al caso aplicativo que se presentará en el capítulo 6, se replicó el procedimiento de simulación anterior considerando como valor de $\omega = 1.2$.

Los cuadros 5.3 y 5.4 muestran que las tendencias descritas en la sección 5.2 se mantienen parcialmente bajo este escenario. En el cuadro 5.3 se observa que a mayor valor del hiperparámetro G_0, la tasa FPR se reduce en las covariables ordinales (h:$1-4$); sin embargo, en el caso de las covariables nominales, la tasa FPR no muestra una tendencia del todo clara. Aún así, los más conveniente sigue siendo considerar valores cercanos a G_0=100.

El cuadro 5.4 muestra que las tendencias de las tasas se mantienen similares a las descritas

	$G_0 = 100$		$G_0 = 200$		$G_0 = 300$	
h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h
1	0	0.75	0	0.75	0	0.25
2	-	1.25	-	0	-	0
3	0	1	0	0	0	0
4	-	1.5	-	1	-	1
5	0.1	1	0.1	0.88	0.1	2.12
6	-	14.36	-	16.71	-	18.14
7	0	1	0	0	0	0
8	-	0.67	-	0.67	-	0.67

Cuadro 5.3: Resultados para $r = 20000$ y distintos valores de G_0 (escenario 2).

en el cuadro 5.2: a mayor valor del ratio de precisión r, la tasa FPR se incrementa tanto en covariables nominales como en ordinales. Siguiendo el criterio utilizado en el escenario anterior, también convendría tomar valores entre $r = 10000$ y $r = 20000$.

	$r = 200$		$r = 2000$		$r = 10000$		$r = 20000$	
h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h	FNR_h	FPR_h
1	0.54	0.5	0	0.5	0	0.25	0	0.75
2	-	0	-	0.25	-	0.5	-	0.25
3	0	0	0	0	0	1	0	1
4	-	0	-	0	-	1.5	-	1.5
5	0.3	0.5	0.1	1.5	0.1	1.38	0.1	1.38
6	-	0.25	-	12.93	-	16	-	15.14
7	0	0	0	0	0	0	0	0.5
8	-	0	-	0.67	-	0.67	-	0.67

Cuadro 5.4: Resultados para $G_0 = 100$ y distintos valores de r (escenario 2).

En este escenario se han presentado problemas de convergencia para valores del hiperparámetro G_0 menores a 100, cosa que no ocurría en el primer escenario. Cabe señalar que aunque se haya incrementado el número iteraciones a 30000 en cada cadena, el problema de convergencia persiste. Dado que en una aplicación real el valor del parámetro ω no es conocido, es muy importante verificar la convergencia de los parámetros del modelo estimado.

Capítulo 6

Aplicación del modelo de fusión a un análisis de regresión sobre la brecha salarial en la macro región sur del Perú

En el presente capítulo se aplicará la metodología expuesta a un conjunto de datos reales; desde el planteamiento de un modelo completo que tome en cuenta todos los niveles de cada covariable cualitativa por separado, la estimación de las probabilidades de fusión a posteriori y la selección el modelo final mediante la optimización de la función de pérdida de Binder.

Se plantea evaluar la existencia de una brecha salarial por etnicidad en trabajadores y trabajadoras de la macro región sur del Perú, para lo cual tomamos de referencia el conjunto de datos utilizado por Baca (2019) para estudiar el efecto de la calidad educativa universitaria sobre la brecha étnica de ingresos. Para la evaluación de la brecha salarial, nos basamos en la metodología utilizada por Sal y Rosas et al. (2019), quienes utilizan una distribución gamma para modelar los ingresos de personal de medicina y enfermería del Perú (variable respuesta positiva), y miden la brecha salarial por género en función del coeficiente estimado para dicha covariable.

6.1. Conjunto de datos

Basado en el estudio de Baca (2019), quien analiza la brecha salarial por etnicidad en Perú, utilizaremos datos de la Encuesta Nacional de Hogares (ENAHO) del año 2019, la cual es administrada por el Instituto Nacional de Estadística e Informática (INEI). Para el caso aplicativo, se está considerando a las personas mayores de 25 años (el autor sugiere que a esa edad se suele terminar los estudios superiores), que tengan alguna ocupación remunerada, que pertenezcan a algún grupo étnico según su lengua materna (no se consideran a aquellos con lengua materna extranjera o con limitaciones para el habla) y que no se dediquen a servicios elementales domésticos o que trabajen en organizaciones extraterritoriales. A diferencia de Baca (2019), quien considera el total de las encuestas, el presente caso está enfocado en la macro región sierra sur del país, que comprende territorios de los departamentos de Apurimac, Arequipa, Cusco, Moquegua, Puno y Tacna; debido a la representatividad de habitantes quechuas y aymaras en dicha macro región. Finalmente, la tabla de datos utilizada en el análisis está conformada por 5442 encuestas.

Las covariables potenciales planteadas por Baca (2019) son el grupo étnico, el estado civil, el sexo, el estrato geográfico (urbano o rural), el tipo de vivienda, la calidad educativa,

los años de educación, el tiempo de experiencia en la empresa en la que trabaja, el tiempo de experiencia potencial (número de años desde que culminó o dejó los estudios), el número de empleados de la empresa en la que trabaja, el sector al que pertenece la empresa, el tipo de ocupación que desempeña y la pertenencia a un sindicato. Para mayor detalle sobre las variables de estudio, revisar el apéndice E.

Cabe señalar que Baca (2019) dicotomizó casi todas las covariables cualitativas consideradas, agrupando las categorías a criterio propio. El presente trabajo muestra una alternativa de agrupamiento de efectos basada en una metodología que depende del conjunto de datos analizado y de la estructura del modelo (bayesiano) planteado.

Dado que el fin del modelo de regresión en estudio es la inferencia de sus parámetros, no se están considerando a los factores de ponderación de la encuesta en las estimaciones.

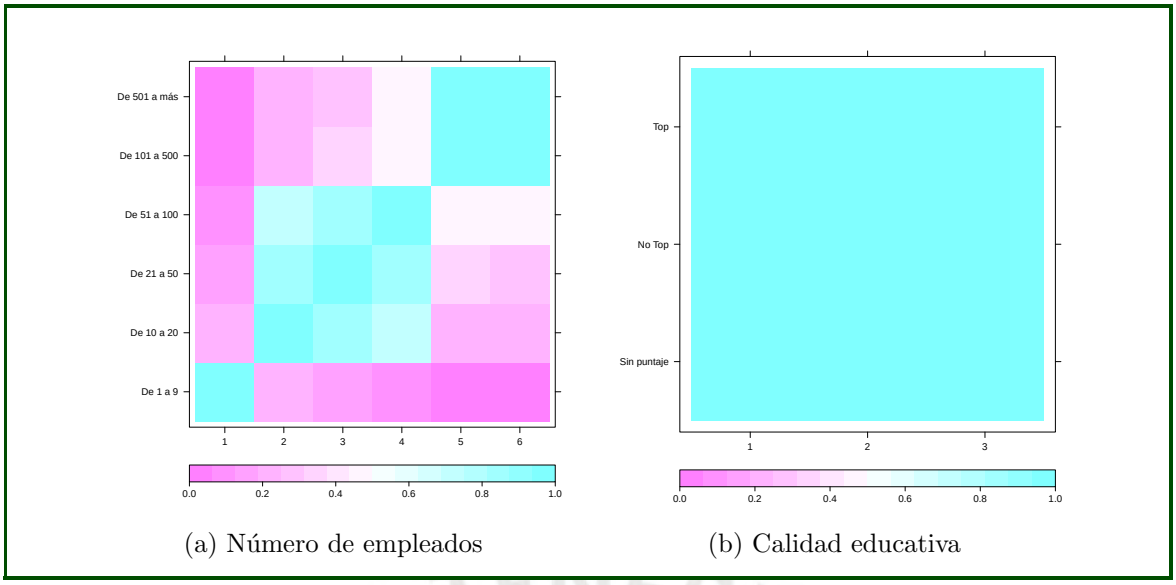
6.2. Estimación del modelo completo

Para especificar el modelo de fusión de efectos, se seleccionaron los valores de los hiperparámetros según los resultados obtenidos en el estudio de simulación (capítulo 5). Para toda covariable cualitativa, ya sea nominal u ordinal, se fijaron los valores de $g_0 = 5$ y $G_0 = 200$; además se consideró como alternativas del ratio de precisión a $r = 10000$ y $r = 20000$, entre las que se escogió la segunda opción ($r = 20000$) según sus valores del DIC (ver apéndice F.1).

Para la estimación de las distribuciones a posteriori mediante el programa JAGS (que usa métodos MCMC), se consideró primero un proceso de adaptación con 10000 iteraciones. Luego, se generaron cuatro cadenas, descartando las primeras 20000 iteraciones y tomando las 20000 posteriores para las estimaciones. La evaluación de la convergencia de las cadenas se hizo de forma visual mediante los *traceplots* (ver apéndice F.3).

El exponencial de la estimación puntual de los coeficientes del modelo completo y sus respectivos intervalos HPD, se muestran en el cuadro 6.2. Sin embargo, en esta etapa lo más relevante es revisar las estimaciones de las probabilidades de fusión a posteriori, pues a partir de ellas se decide cuáles son los efectos que deben fusionarse.

Para visualizar la probabilidad de fusión a posteriori de los efectos de cada covariable podemos hacer uso de mapas de calor, los cuales muestran cuáles son los efectos que tienen una alta probabilidad de fusión cuando el color del casillero tiende a ser celeste, mientras que el tono rosa indica una baja probabilidad de fusión. La figura 6.1 muestra el mapa de calor de las probabilidades de fusión a posteriori para las covariable ordinales: el número de empleados de la empresa y la calidad educativa según el ranking Web of Science. En el caso del número de empleados, el gráfico sugiere que pueden existir tres grupos en función de las probabilidades de fusión a posteriori: las empresas que tengan hasta 9 empleados como máximo (referencia), las empresas que tengan de 10 a 100 empleados y las empresas que con más de 100 empleados; lo que indica que los niveles dentro de cada grupo tendrían efectos similares sobre el ingreso por hora. Por otro lado, respecto a la calidad educativa, podemos ver que todos los niveles tienen una alta probabilidad de fusión y formarían un solo grupo; es decir, la calidad educativa de los entrevistados no tendría un efecto relevante sobre sus niveles de ingreso.



(a) Número de empleados (b) Calidad educativa

Figura 6.1: Probabilidades de fusión a posteriori para las covariables ordinales del modelo.

La figura 6.2 muestra los mapas de calor de las probabilidades de fusión a posteriori para las covariables nominales. En el caso del estado civil, podemos ver que los convivientes, casados y separados o divorciados tienen una alta probabilidad de fusión a posteriori; por lo que estos niveles conformarían un grupo. Para el sector productivo, hay indicios de que los sectores primario y secundario podrían ser agrupados. En el caso de la ocupación, si bien no es del todo claro, se puede ver que los profesionales, jefes y directivos, operadores y transportistas tienen una alta probabilidad de fusión a posteriori; así como también los que trabajan en construcción con los que trabajan en servicios elementales. Respecto al tipo de vivienda, todos los niveles tienen una alta probabilidad de fusión a posteriori; es decir, ningún nivel de la covariable tiene un efecto diferente sobre el ingreso por hora de los encuestados. Finalmente, se puede ver que los grupos étnicos tienen una baja probabilidad de fusión a posteriori, por lo que se podría decir que los efectos son diferentes entre sí y que existe una brecha salarial dada por esta covariable, sobre todo respecto a aquellos que tienen al castellano como lengua materna (no indígenas).

6.3. Selección del modelo final

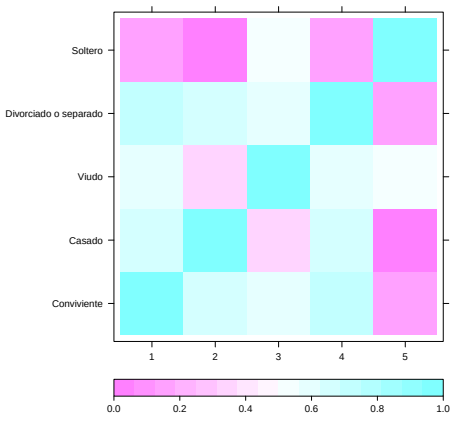
En la sección 6.2 se estimaron las probabilidades de fusión a posteriori y se pudo dar algunos alcances sobre qué niveles podrían fusionarse; sin embargo, debido a la subjetividad del análisis visual, es conveniente apoyarnos en alguna metodología cuantitativa que permita discernir mejor sobre algunas agrupaciones que no sean fácilmente distinguibles en el mapa de calor, como puede ser el caso de la covariable ocupación. Para este fin, en el capítulo 4 se presentó una alternativa más objetiva, la cual consiste en agrupar los niveles minimizando la función de pérdida de Binder.

El cuadro 6.1 muestra los resultados de la optimización de la función de pérdida de Binder; donde la segunda columna indica cuáles son los efectos de cada covariable que deben fusionarse asignándolos a un mismo grupo (representado por un número). En él se puede

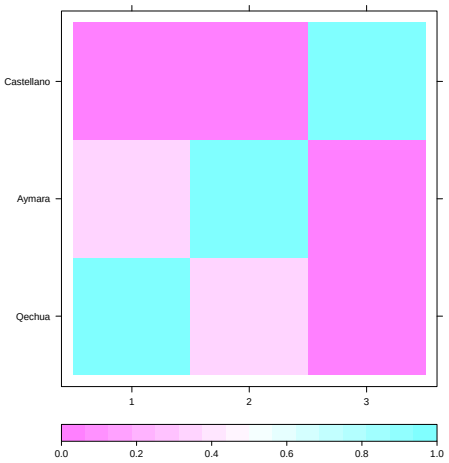
apreciar, por ejemplo, que la covariable ocupación finalmente posee cuatro grupos, reforzando los resultados del análisis visual mediante mapas de calor. Para las demás covariables, los resultados también coinciden con la interpretación gráfica de las probabilidades de fusión a posteriori dadas en las sección anterior.

	Agrupación
Número de empleados	
De 1 a 9 (referencia)	1
De 10 a 20	2
De 21 a 50	2
De 51 a 100	2
De 101 a 500	3
De 501 a más	3
Calidad educativa	
Sin puntaje (referencia)	1
No top	1
Top	1
Estado civil	
Conviviente (referencia)	1
Casado	1
Viudo	2
Divorciado o separado	1
Soltero	2
Grupo étnico	
Qechua (referencia)	1
Aymara	2
Castellano	3
Sector	
Primario (referencia)	1
Secundario	1
Terciario	2
Ocupación	
Profesionales (referencia)	1
Técnicos	2
Jefes y directivos	1
Servicios y vendedores	1
Agricultura, pecuario, etc.	3
Construcción	4
Operadores y transportes	1
Servicios elementales	4
Tipo de vivienda	
Otro (referencia)	1
Propia	1
Alquilada	1

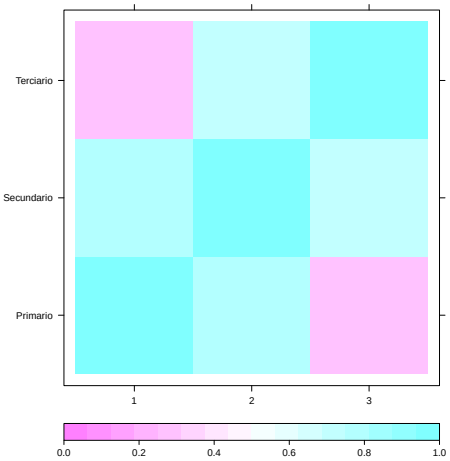
Cuadro 6.1: Agrupación de efectos bajo el modelo estimado de fusión.



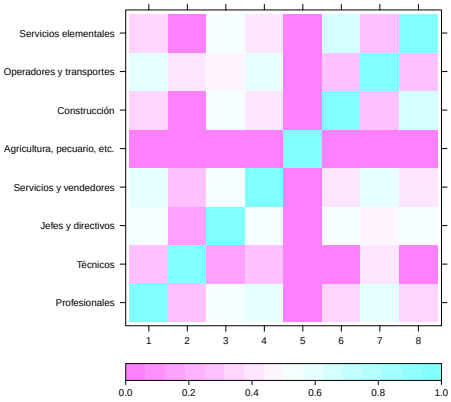
(a) Estado civil



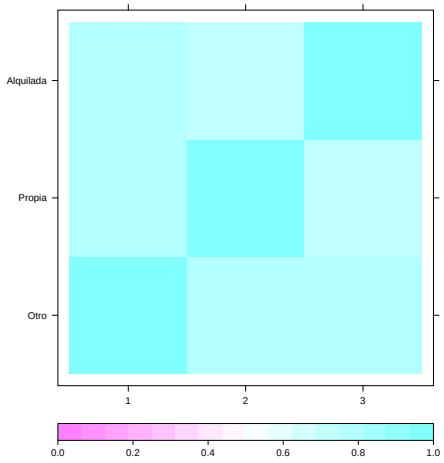
(b) Grupo étnico



(c) Sector



(d) Ocupación



(e) Vivienda

Figura 6.2: Probabilidades de fusión a posteriori para las covariables nominales del modelo.

Finalmente, tenemos mayor certeza de qué efectos pueden ser fusionados y cuáles no, confirmando también la existencia de una brecha salarial en función del grupo étnico, pues vemos que cada nivel tiene un efecto diferente sobre el ingreso por hora. Sin embargo, para saber qué grupo étnico presenta un mayor o menor efecto sobre el ingreso, bastaría con revisar la estimación a posteriori de sus coeficientes. Por otro lado, dado que los niveles dentro de la calidad educativa y del tipo de vivienda tienen efectos muy similares, incluyendo a sus respectivos niveles de referencia, ambas covariables pueden ser excluidas del modelo final seleccionado.

6.4. Estimación del modelo final

Luego de haber estimado el modelo completo, calculando las probabilidades de fusión a posteriori, y de haber agrupado niveles según los resultados en base a la función de pérdida de Binder; se procede a estimar el modelo final con efectos fusionados.

Para las estimaciones del modelo final seleccionado, primero se usaron 5000 iteraciones para el proceso de adaptación; luego se generaron cuatro cadenas, descartando las primeras 15000 iteraciones y tomando las 10000 posteriores para las estimaciones (ver apéndice F.4). A comparación del modelo completo, no fueron necesarias tantas iteraciones para lograr la convergencia del modelo final debido a la mayor simplicidad en su estructura.

El modelo final estimado se muestra en el cuadro 6.2, el cual presenta 15 efectos significativos (diferentes del nivel de referencia) y diferentes entre sí, en comparación con los 30 efectos que considera el modelo completo. En el modelo final se puede apreciar, por ejemplo, que los efectos del tamaño de la empresa se han fusionado mostrando solo dos valores diferentes a la referencia; lo mismo ocurre para las demás covariables con niveles que debían ser fusionados. De acuerdo a la selección del modelo final, la calidad educativa y el tipo de vivienda fueron excluidos debido a que todos sus niveles debían fusionarse.

En el modelo final (cuadro 6.2) se puede ver que la brecha salarial por grupo étnico identificada favorece a los no indígenas, pues su efecto indica que en promedio presentan mayores niveles de ingreso respecto a los quechuas y aymaras (28 % mayor que los quechuas), siendo los aymaras aquellos que presentan el menor ingreso promedio (15 % menor que los quechuas). Por otro lado, el trabajar en una empresa con mayor número de empleados tiene un efecto positivo sobre el ingreso por hora de lo entrevistados. Respecto al estado civil, resulta interesante saber que los viudos y los solteros tienen un efecto negativo sobre sus niveles de ingreso en comparación a los demás estados civiles (convivientes, casados y divorciados o separados). En cuanto al sector productivo, aquellos que trabajen en el sector terciario tienen en promedio un menor ingreso que aquellos que se desempeñan en los sectores primario o secundario. Los entrevistados dedicados a una ocupación técnica tienen en promedio un ingreso mayor que los demás; mientras que aquellos con menores ingresos en promedio son los que trabajan en construcción, servicios elementales y, sobre todo, los productores agropecuarios (75 % menos que los profesionales). Respecto al género, se puede ver que en promedio los ingresos de las entrevistadas mujeres es 43 % menor al de los varones, lo cual también podría interpretarse como indicios de una brecha de género. Por otro lado, aquellos que residen en un espacio urbano tienen en promedio un mayor nivel de ingresos. Otro aspecto interesante

	Modelo completo			Modelo final		
	e^{β}	Intervalo HPD (95 %)		e^{β}	Intervalo HPD (95 %)	
Número de empleados						
De 10 a 20	1.241	1.004	1.444	1.260	1.136	1.400
De 21 a 50	1.301	1.069	1.538	1.260	1.136	1.400
De 51 a 100	1.372	1.132	1.676	1.260	1.136	1.400
De 101 a 500	1.601	1.421	1.811	1.651	1.533	1.783
De 501 a más	1.619	1.478	1.766	1.651	1.533	1.783
Calidad educativa						
No top	1.035	0.968	1.108	-	-	-
Top	1.033	0.957	1.122	-	-	-
Estado civil						
Casado	1.051	0.994	1.117	-	-	-
Viudo	0.940	0.843	1.038	0.812	0.764	0.867
Divorciado o separado	0.998	0.924	1.078	-	-	-
Soltero	0.825	0.741	0.909	0.812	0.764	0.867
Grupo étnico						
Aymara	0.873	0.786	0.968	0.846	0.783	0.917
Castellano	1.288	1.207	1.378	1.280	1.202	1.362
Sector						
Secundario	0.943	0.806	1.044	-	-	-
Terciario	0.879	0.773	0.998	0.825	0.762	0.895
Ocupación						
Técnicos	1.207	1.040	1.419	1.280	1.140	1.441
Jefes y directivos	0.950	0.847	1.061	-	-	-
Servicios y vendedores	0.989	0.887	1.098	-	-	-
Agricultura, pecuario, etc.	0.367	0.303	0.437	0.346	0.310	0.386
Construcción	0.884	0.752	1.015	0.800	0.739	0.865
Operadores y transportes	1.026	0.912	1.160	-	-	-
Servicios elementales	0.884	0.779	1.001	0.800	0.739	0.865
Tipo de vivienda						
Propia	1.015	0.954	1.084	-	-	-
Alquilada	0.985	0.936	1.035	-	-	-
Sexo						
Mujer	0.660	0.622	0.700	0.666	0.629	0.702
Estrato geográfico						
Urbano	1.315	1.231	1.407	1.308	1.226	1.395
Sindicato						
Sí pertenece	1.227	1.067	1.421	1.233	1.074	1.418
Años de educación	1.390	1.328	1.454	1.390	1.331	1.448
Experiencia en la empresa	1.076	1.041	1.115	1.070	1.034	1.107
Experiencia potencial	0.947	0.903	0.991	0.968	0.928	1.008

Cuadro 6.2: Estimación de efectos a posteriori.

es que los entrevistados que pertenecen a un sindicato tienen en promedio un mayor ingreso que aquellos que no pertenecen a alguno. Entre las variables cuantitativas, los años de educación es la variable con mayor efecto positivo sobre los ingresos, seguido de el tiempo de experiencia en la empresa; mientras que la experiencia general o potencial tendría un efecto

negativo o no significativo sobre los ingresos, según su intervalo HPD.

Los efectos estimados en el modelo final son coherentes con los efectos obtenidos en el modelo completo; así mismo, la media a posteriori del parámetro de precisión ω resultó similar en ambos modelos (ver apéndice F.5).

6.5. Interacción de covariables

Si bien el estudio de Baca (2019) no considera interacciones de las covariables, el modelo propuesto permite incluir todas las que se requieran. En esta sección se incluirán, a modo de ejemplo, la interacción entre el sexo y el grupo étnico, para profundizar los resultados del caso aplicativo, y una interacción con restricciones, para mostrar las consideraciones que ésto implica.

Para incluir una interacción dentro del modelo, se debe calcular una nueva covariable que considere la combinación de los niveles diferentes a los de referencia de ambas covariables, o agregar las interacciones directamente en la matriz de diseño. Para que el modelo evalúe la fusión de niveles de las interacciones, también se debe asignar una distribución a priori para el indicador de fusión de sus efectos, tal como se hace en las otras covariables cualitativas.

6.5.1. Interacción entre el sexo y el grupo étnico

Dentro de las covariables del caso de estudio presentado, consideramos que el sexo y el grupo étnico son las más relevantes y podría ser de interés incluir su interacción en el modelo. De esta manera, podremos profundizar un poco más sobre la brecha existente en función de los efectos dentro de la combinación de ambas covariables.

Dado que el grupo étnico y el sexo son covariables nominales, la matriz de estructura a priori de los efectos de la interacción no considerará restricciones ($\mathbf{Q}(\boldsymbol{\delta})$). En el caso de que la interacción comprenda alguna covariable con restricciones (por ejemplo: ordinal), se deberán tener en cuenta estas restricciones en la matriz de estructura.

Para las estimaciones de los modelos (completo y final) con interacción, se utilizaron los mismos valores de los hiperparámetros G_0 y r, y el mismo número de iteraciones descritos en las secciones 6.2 y 6.4.

El cuadro 6.3 muestra las estimaciones del modelo completo y del modelo final, el cual se basa en los grupos formados a partir de las probabilidades de fusión a posteriori estimadas (ver apéndice G). Se puede apreciar que la introducción de la interacción del grupo étnico y el sexo excluye adicionalmente del modelo al sector productivo y, además, fusiona los efectos de los grupos étnicos quechua y aymara. La interacción indica que la diferencia de ingresos entre quechuas y aymaras encontrada en el modelo sin interacción (cuadro 6.1) se debería al efecto negativo en contra de las mujeres dentro del grupo aymara, pues muestra que la brecha de ingresos por género presente (25 % menor respecto de los hombres) se hace mayor cuando las mujeres son aymaras o no indígenas (-26 % adicional).

	Modelo completo			Modelo final		
	e^{β}	Intervalo HPD (95 %)		e^{β}	Intervalo HPD (95 %)	
Número de empleados						
De 10 a 20	1.231	1.006	1.436	1.295	1.172	1.441
De 21 a 50	1.281	1.054	1.524	1.295	1.172	1.441
De 51 a 100	1.375	1.141	1.694	1.295	1.172	1.441
De 101 a 500	1.609	1.437	1.803	1.663	1.545	1.794
De 501 a más	1.623	1.493	1.775	1.663	1.545	1.794
Calidad educativa						
No top	1.033	0.969	1.108	-	-	-
Top	1.028	0.949	1.116	-	-	-
Estado civil						
Casado	1.049	0.992	1.116	-	-	-
Viudo	0.913	0.812	1.025	0.807	0.757	0.859
Divorciado o separado	0.994	0.919	1.071	-	-	-
Soltero	0.828	0.752	0.912	0.807	0.757	0.859
Grupo étnico						
Aymara	0.976	0.921	1.034	-	-	-
Castellano	1.443	1.340	1.549	1.464	1.366	1.566
Sector						
Secundario	0.962	0.868	1.054	-	-	-
Terciario	0.901	0.801	1.000	-	-	-
Ocupación						
Técnicos	1.186	1.030	1.383	1.303	1.157	1.459
Jefes y directivos	0.949	0.847	1.060	-	-	-
Servicios y vendedores	0.968	0.863	1.076	-	-	-
Agricultura, pecuario, etc.	0.373	0.316	0.438	0.411	0.378	0.446
Construcción	0.876	0.756	1.005	0.884	0.825	0.948
Operadores y transportes	1.020	0.903	1.156	-	-	-
Servicios elementales	0.883	0.779	0.995	0.884	0.825	0.948
Tipo de vivienda						
Propia	1.013	0.953	1.076	-	-	-
Alquilada	0.986	0.937	1.034	-	-	-
Sexo × Grupo étnico						
Mujer:Aymara	0.760	0.681	0.846	0.744	0.680	0.811
Mujer:Castellano	0.759	0.687	0.838	0.744	0.680	0.811
Sexo						
Mujer	0.761	0.705	0.825	0.754	0.703	0.812
Estrato geográfico						
Urbano	1.314	1.233	1.406	1.304	1.221	1.392
Sindicato						
Sí pertenece	1.227	1.067	1.420	1.265	1.102	1.457
Años de educación	1.399	1.338	1.465	1.387	1.331	1.446
Experiencia en la empresa	1.081	1.045	1.120	1.080	1.045	1.118
Experiencia potencial	0.949	0.907	0.993	0.954	0.917	0.994

Cuadro 6.3: Estimación de efectos a posteriori con interacción: Sexo × Grupo étnico.

6.5.2. Interacción con restricciones: número de empleados y grupo étnico

A modo de ilustrar el caso en que la matriz de estructura tenga restricciones ($\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$), se considerará la interacción entre el número de empleados y el grupo étnico, una covariable ordinal y nominal, respectivamente.

Para simplificar el cálculo de los términos faltantes de la matriz de similaridad a posteriori, debido a las restricciones, se agruparán los niveles del número de empleados según los resultados del cuadro 6.1. El cuadro 6.4 muestra los niveles de la interacción entre ambas covariables que se están considerando.

k	Número de empleados × Grupo étnico
0	De 1 a 9:Quechua (referencia)
1	De 10 a 100:Aymara
2	De 101 a más:Aymara
3	De 10 a 100:Castellano
4	De 101 a más:Castellano

Cuadro 6.4: Niveles de la interacción entre el número de empleados y el grupo étnico.

Las restricciones dadas por el numero de empleados (solo puede haber fusión en niveles adyacentes) se verán reflejadas en la interacción. El cuadro 6.5 muestra valores de los elementos del vector $\boldsymbol{\zeta}$, que indican cuáles son los niveles en los que es posible evaluar una fusión de efectos. En este caso, se podrá evaluar la fusión de efectos ($\zeta_{kj} = 1$) cuando los niveles k y j de la interacción están compuestos por niveles adyacentes o por el mismo nivel del número de empleados; en caso contrario, $\zeta_{kj} = 0$.

k	0	0	0	0	1	1	1	2	2	3
j	1	2	3	4	2	3	4	3	4	4
ζ_{kj}	1	0	1	0	1	1	1	1	1	1

Cuadro 6.5: Definición de ζ_{kj} para la interacción entre el número de empleados y el grupo étnico.

A partir del vector $\boldsymbol{\zeta}$, la estructura de la matriz de precisión a priori quedará definida de la siguiente forma:

$$\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta}) = \begin{pmatrix} q_{11} & -\kappa_{12} & -\kappa_{13} & -\kappa_{14} \\ -\kappa_{21} & q_{22} & -\kappa_{23} & -\kappa_{24} \\ -\kappa_{31} & -\kappa_{32} & q_{33} & -\kappa_{34} \\ -\kappa_{41} & -\kappa_{42} & -\kappa_{43} & q_{44} \end{pmatrix}, \tag{6.1}$$

donde los elementos de la diagonal se definen de la siguiente manera:

$$q_{kk} = \begin{cases} \kappa_{10} + \kappa_{30} - \sum_{k \neq j} q_{kj}\,, & \text{si } k \in \{1, 3\}\,, \\ -\sum_{k \neq j} q_{kj} & \text{, si } k \notin \{1, 3\}\,. \end{cases} \tag{6.2}$$

De (6.1) y (6.2), se puede observar que las restricciones de la interacción solo afectan a los elementos de la diagonal de la matriz de estructura, debido al reducido número de niveles. En el apéndice H se puede apreciar la forma de la matriz de estructura en caso se hubieran considerado todos los niveles del número de empleados en la interacción.

Para completar los elementos faltantes de la matriz de similaridad a posteriori, compuesta por las probabilidades de fusión a posteriori, se deben calcular los indicadores de fusión δ_{20} y δ_{40}. Para $k = 2, 4$, se asume lo siguiente:

$$\delta_{k0} \begin{cases} 0, & \text{si } (\delta_{10} = 0 \text{ y } \delta_{k1} = 0) \text{ ó } (\delta_{30} = 0 \text{ y } \delta_{k3} = 0), \\ 1, & \text{caso contrario.} \end{cases}$$

Tener en cuenta que cambiar estas condiciones podría alterar el resultado del modelo final.

Las estimaciones del modelo completo y final se realizaron bajo las mismas condiciones descritas en la subsección 6.5.1.

El cuadro 6.6 muestra las estimaciones del modelo completo y del modelo final (ver agrupación de efectos en el apéndice I). Se observa que al incluir la interacción del número de empleados y el grupo étnico, los coeficientes estimados del modelo son similares al caso sin restricciones (ver cuadro 6.2) a excepción del coeficiente del grupo étnico aymara, el cual incrementa su efecto negativo sobre los ingresos. La interacción indica que hay un efecto favorable sobre los ingresos para los entrevistados aymaras que laboran en empresas con 10 o más empleados (46 % mayor respecto de los otros grupos); compensando el incremento negativo del efecto global de los aymaras sobre los ingresos.

	Modelo completo			Modelo final		
	e^{β}	Intervalo HPD (95 %)		e^{β}	Intervalo HPD (95 %)	
Número de empleados						
De 10 a 100	1.383	1.163	1.711	1.214	1.092	1.349
De 101 a más	1.670	1.465	1.883	1.596	1.477	1.722
Calidad educativa						
No top	1.038	0.971	1.118	-	-	-
Top	1.038	0.960	1.128	-	-	-
Estado civil						
Casado	1.052	0.992	1.117	-	-	-
Viudo	0.936	0.840	1.040	0.812	0.763	0.865
Divorciado o separado	0.997	0.922	1.075	-	-	-
Soltero	0.825	0.748	0.916	0.812	0.763	0.865
Grupo étnico						
Aymara	0.857	0.773	0.963	0.795	0.731	0.865
Castellano	1.314	1.223	1.417	1.290	1.212	1.372
Sector						
Secundario	0.950	0.827	1.046	-	-	-
Terciario	0.880	0.781	0.993	0.816	0.751	0.883
Ocupación						
Técnicos	1.200	1.042	1.394	1.274	1.133	1.429
Jefes y directivos	0.951	0.847	1.072	-	-	-
Servicios y vendedores	0.989	0.889	1.108	-	-	-
Agricultura, pecuario, etc.	0.369	0.312	0.433	0.344	0.309	0.383
Construcción	0.879	0.745	1.007	0.794	0.734	0.859
Operadores y transportes	1.025	0.911	1.162	-	-	-
Servicios elementales	0.874	0.765	0.981	0.794	0.734	0.859
Tipo de vivienda						
Propia	1.014	0.950	1.077	-	-	-
Alquilada	0.984	0.936	1.035	-	-	-
Número de empleados × Grupo étnico						
De 10 a 100:Aymara	1.033	0.897	1.230	1.464	1.181	1.814
De 101 a más:Aymara	1.052	0.883	1.317	1.464	1.181	1.814
De 10 a 100:Castellano	0.899	0.694	1.048	-	-	-
De 101 a más:Castellano	0.942	0.818	1.084	-	-	-
Sexo						
Mujer	0.661	0.623	0.701	0.670	0.635	0.708
Estrato geográfico						
Urbano	1.319	1.235	1.410	1.311	1.227	1.401
Sindicato						
Sí pertenece	1.233	1.069	1.427	1.231	1.072	1.420
Años de educación	1.386	1.323	1.449	1.392	1.334	1.452
Experiencia en la empresa	1.077	1.040	1.115	1.072	1.037	1.110
Experiencia potencial	0.947	0.905	0.993	0.969	0.928	1.008

Cuadro 6.6: Estimación de efectos a posteriori con interacción: Número de empleados × Grupo étnico.

Capítulo 7

Conclusiones

7.1. Conclusiones

El método presentado permitió encontrar un modelo menos complejo, en términos del número de coeficientes, gracias a la fusión de efectos que eran irrelevantes o similares entre sí dentro de cada covariable categórica. La fusión fue posible debido a las propiedades del modelo a partir de la elección de una distribución a priori spike y slab para la diferencia de efectos en cada covariable (en consecuencia, una normal multivariada para los efectos); lo cual permitió controlar la estructura de la matriz de precisión de tal manera que se pudieran considerar restricciones para especificar qué pares de efectos podrían ser evaluados para una fusión. Estas restricciones fueron de utilidad en las covariables ordinales, las cuales solo permiten la fusión de niveles adyacentes.

Fue posible implementar el modelo bayesiano de fusión de efectos en el programa JAGS y adaptarlo a un modelo con respuesta de distribución gamma. Este modelo resulta útil en caso se tenga una variable respuesta que solo admita valores positivos, como es el caso del ingreso por hora. Cabe resaltar que la implementación del modelo en JAGS es relativamente sencilla, por lo que este método puede extenderse a modelos con otras distribuciones para la variable respuesta, con distinta función de enlace, o incluso a modelos aditivos generalizados.

En el modelo presentado se encontró que las tendencias de los ratios FNR y FPR en función de los valores de los hiperparámetros G_0 y r eran similares a los descritos por Pauger y Wagner (2019); sin embargo, para este tipo de modelo resulta más costoso llegar a la convergencia y, además, tanto la convergencia como los resultados de los ratios FNR y FPR pueden verse alterados según el valor del parámetro ω.

Para la selección del modelo final, se utilizaron las estimaciones de la probabilidad de fusión a posteriori de cada par de efectos como elementos de una matriz de similaridad. El modelo final es resultado de la agrupación de efectos que minimiza la función de pérdida de Binder para cada covariable. Las simulaciones mostraron que el método de selección del modelo funciona adecuadamente, tanto en covariables nominales como en ordinales, según el valor de los hiperparámetros que se escojan.

Por otro lado, se pudo aplicar el modelo de fusión a un conjunto de datos reales, en el cual se pudo apreciar que la elección del modelo final podría resultar complicado a partir de un análisis visual. La función de pérdida de Binder fue una herramientas de mucha utilidad para realizar esta tarea de manera objetiva, mostrando resultados coherentes con el análisis visual. Los coeficientes estimados del modelo final guardan relación con los del modelo completo,

con la ventaja de que el modelo final logra reducir la cantidad de parámetros y disminuir el valor del DIC (ver apéndice F.2).

Finalmente, en el caso aplicativo se pudo evidenciar la existencia de una brecha salarial en relación al grupo étnico, pues sus niveles no fueron fusionados en el modelo final. Respecto de los quechuas, se estimó un efecto favorable sobre el ingreso para los no indígenas (28 % mayor que los quechua) y un efecto negativo para los aymaras (15 % menor que los quechuas). Además, se explicó cómo incluir la interacción de covariables (con y sin restricciones) al modelo; planteando la interacción del grupo étnico y el sexo en el caso aplicativo, donde se encontró que la brecha salarial por género sería mayor en las mujeres no indígenas y en las aymaras.

7.2. Sugerencias para futuras investigaciones

- Sería interesante evaluar el desempeño predictivo del modelo final, tanto en el estudio de simulación, como en un caso real. Esta información puede ser relevante para la aplicación del modelo en casos donde el principal objetivo sea la predicción de datos.
- Evaluar la posibilidad de utilizar una alternativa a la función de pérdida de Binder para la selección del modelo final, manteniendo la objetividad que brinda el método y tratando de mantener o mejorar el desempeño del modelo final.
- Investigar si se pueden considerar otras distribuciones a priori para el vector indicador de fusión $\boldsymbol{\delta}$ y comparar los desempeños. En caso de encontrar una distribución a priori continua para $\boldsymbol{\delta}$, se podría implementar el modelo en el programa STAN y evaluar si así se puede reducir el tiempo de estimación del modelo.
- Profundizar la evaluación del efecto de los hiperparámetros G_0 y r sobre el desempeño del modelo, considerando una mayor cantidad de combinaciones de valores de G_0 y r. Así mismo, replicar este ejercicio para distintos valores del parámetro ω.
- Comparar los resultados obtenidos con otras técnicas de fusión de efectos, ya sea desde el enfoque frecuentista o bayesiano, para verificar que se mantiene la ventaja del método propuesto por Pauger y Wagner (2019), aún cuando el modelo tiene una variable respuesta con distribución diferente a la normal.
- Automatizar y optimizar los procesos del modelo implementado para poder difundirlo a través de un paquete en R.

Apéndice A

Distribución a priori del vector de efectos $\boldsymbol{\beta}$

Las propiedades referentes a la fusión de efectos de una covariable cualitativa a partir de la distribución a priori del vector de efectos $\boldsymbol{\beta}$ dado en (2.5), cuya matriz de precisión $\mathbf{Q}(\boldsymbol{\delta})$ está definida en (2.2.1); son consecuencia de haber especificado una priori spike y slab sobre la diferencia de efectos $\theta_{kj} = \beta_k - \beta_j$ para $0 \leq j < k \leq c$, tal como se detalla en (2.2).

A continuación, se presenta la demostración hecha por Pauger y Wagner (2017) con algunas precisiones adicionales para comprender mejor el planteamiento de la distribución a priori y sus propiedades para la fusión de efectos.

Se define un vector $\boldsymbol{\theta}$, de dimensiones $d \times 1$, que contiene a todas las diferencia de efectos θ_{kj} de una covariable cualitativa. Los primeros c elementos de $\boldsymbol{\theta}$ corresponden a un subvector $\boldsymbol{\theta_0}$, cuyos elementos se definen como $\theta_{k0} = \beta_k - \beta_0$ para $0 < k \leq c$. Siendo β_0 el efecto nivel de referencia de la covariable, su valor viene a ser igual a cero; por lo que se puede encontrar la equivalencia $\boldsymbol{\theta_0} = \boldsymbol{\beta}$. El subvector $\boldsymbol{\theta_{/0}}$ contiene a los últimos $d - c$ elementos del vector $\boldsymbol{\theta_0}$.

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{10} \\ \vdots \\ \theta_{c0} \\ \theta_{21} \\ \vdots \\ \theta_{c1} \\ \vdots \\ \theta_{c-1,c} \end{pmatrix} = \left(\begin{array}{c} \boldsymbol{\theta_0} \\ \hline \\ \boldsymbol{\theta_{/0}} \end{array}\right) = \left(\begin{array}{c} \boldsymbol{\beta} \\ \hline \\ \boldsymbol{\theta_{/0}} \end{array}\right)$$

Entonces, podemos escribir la restricción lineal $\theta_{kj} = \beta_k - \beta_j = \theta_{k0} - \theta_{j0}$ de tal forma que podemos expresar a $\boldsymbol{\theta_{/0}}$ como resultado de la combinación lineal de $\boldsymbol{\theta_0}$ a través de una matriz $\mathbf{U}$, tal que $\mathbf{U}\boldsymbol{\theta_0} - \boldsymbol{\theta_{/0}} = \mathbf{0}$.

Luego, a partir de (2.4), podemos expresar la distribución del vector $\boldsymbol{\theta}$ como una normal multivariada $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$; de tal manera que la distribución de $\boldsymbol{\theta}$ bajo la restricción $\mathbf{R}\boldsymbol{\theta} = \mathbf{0}$ sigue siendo una normal multivariada, con momentos:

$$E(\boldsymbol{\theta}|\mathbf{R}\boldsymbol{\theta} = \mathbf{0}) = \mathbf{0},$$
$$COV(\boldsymbol{\theta}|\mathbf{R}\boldsymbol{\theta} = \mathbf{0}) = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{R}^{\intercal}(\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\intercal})^{-1}\boldsymbol{R}\boldsymbol{\Sigma},$$

donde $\mathbf{R} = (\mathbf{U} - \mathbf{I})$, siendo $\mathbf{I}$ una matriz identidad de dimensiones $d - c \times 1$ (Rue y Held,

2005).

Para determinar la matriz de covarianza $\boldsymbol{\Omega} = COV(\boldsymbol{\theta}|\mathbf{R}\boldsymbol{\theta} = \mathbf{0})$, primero se puede dividir la estructura de la matriz $\boldsymbol{\Sigma}$ convenientemente:

$$\boldsymbol{\Sigma} = \gamma\tau^2 \begin{pmatrix} \boldsymbol{\Delta}_0 & \\ & \boldsymbol{\Delta}_{/0} \end{pmatrix} = \gamma\tau^2 \begin{pmatrix} \boldsymbol{\kappa}_0^{-1} & \\ & \boldsymbol{\kappa}_{/0}^{-1} \end{pmatrix}, \tag{A.1}$$

donde $\boldsymbol{\kappa}_0 = diag(\kappa_{10}, ..., \kappa_{c0})$ y $\boldsymbol{\kappa}_{/0}$ es la matriz diagonal con elementos κ_{kj}, donde $j \neq 0$. Entonces, en base a las siguientes expresiones:

$$\boldsymbol{\Sigma}\mathbf{R}^\intercal = \gamma\tau^2 \begin{pmatrix} \boldsymbol{\Delta}_0\mathbf{U}^\intercal \\ -\boldsymbol{\Delta}_{/0} \end{pmatrix}$$

$$(\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^\intercal)^{-1} = \frac{1}{\gamma\tau^2}\left(\mathbf{U}\boldsymbol{\Delta}_0\mathbf{U}^\intercal + \boldsymbol{\Delta}_{/0}\right)^{-1} = \frac{1}{\gamma\tau^2}\mathbf{W}$$

la estructura de la matriz $\boldsymbol{\Omega}$ quedaría definida de la siguiente forma:

$$\begin{aligned} \boldsymbol{\Omega} &= \gamma\tau^2 \left(\boldsymbol{\Delta} - \begin{pmatrix} \boldsymbol{\Delta}_0\mathbf{U}^\intercal \\ -\boldsymbol{\Delta}_{/0} \end{pmatrix} \mathbf{W} \left(\mathbf{U}\boldsymbol{\Delta}_0 - \boldsymbol{\Delta}_{/0} \right) \right) \\ &= \gamma\tau^2 \left[\begin{pmatrix} \boldsymbol{\Delta}_0 & \\ & \boldsymbol{\Delta}_{/0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Delta}_0\mathbf{U}^\intercal\mathbf{W}\mathbf{U}\boldsymbol{\Delta}_0 & -\boldsymbol{\Delta}_0\mathbf{U}^\intercal\mathbf{W}\boldsymbol{\Delta}_{/0} \\ -\boldsymbol{\Delta}_{/0}\mathbf{W}\mathbf{U}\boldsymbol{\Delta}_0 & \boldsymbol{\Delta}_{/0}\mathbf{W}\boldsymbol{\Delta}_{/0} \end{pmatrix} \right]. \end{aligned} \tag{A.2}$$

Dada la partición hecha sobre la matriz $\boldsymbol{\Sigma}$ (A.1), se puede ver que la submatriz superior izquierda $c \times c$ de $\boldsymbol{\Omega}$ (A.2) corresponde a la matriz de covarianza de los efectos $Cov(\boldsymbol{\beta}) = \mathbf{B}_0(\boldsymbol{\delta}, \tau^2)$, entonces

$$\mathbf{B}_0(\boldsymbol{\delta}, \tau^2) = \gamma\tau^2 \left(\boldsymbol{\Delta}_0 - \boldsymbol{\Delta}_0\mathbf{U}^\intercal\mathbf{W}\mathbf{U}\boldsymbol{\Delta}_0 \right),$$

donde la matriz de precisión $\mathbf{Q}(\boldsymbol{\delta})$ es dada por

$$\mathbf{Q}(\boldsymbol{\delta}) = \left(\boldsymbol{\Delta}_0 - \boldsymbol{\Delta}_0\mathbf{U}^\intercal\mathbf{W}\mathbf{U}\boldsymbol{\Delta}_0 \right)^{-1}.$$

Aplicando la formula de Woodbury (Higham, 2002), tenemos que

$$\begin{aligned} \mathbf{Q}(\boldsymbol{\delta}) &= \boldsymbol{\Delta}_0^{-1} + \mathbf{U}^\intercal\boldsymbol{\Delta}_{/0}^{-1}\mathbf{U} \\ &= \boldsymbol{\kappa}_0^{-1} + \mathbf{U}^\intercal\boldsymbol{\kappa}_{/0}^{-1}\mathbf{U} \end{aligned}$$

Sea $\mathbf{u}_k$ la $k-$ésima columna de la matriz $\mathbf{U}$, entonces los elementos fuera de la diagonal de $\mathbf{Q}(\boldsymbol{\delta})$ están dados por

$$q_{kj} = \mathbf{u}_k^\intercal\boldsymbol{\kappa}_{/0}\mathbf{u}_j = -\kappa_{kj},$$

donde cada par de vectores $\mathbf{u}_k$ y $\mathbf{u}_j$ no tienen elementos iguales a cero en las filas que corresponden a la restricción lineal para θ_{kj}; sino que en la correspondiente fila un vector tendrá un valor de 1 y el otro, de -1.

Por último, los elementos de la diagonal de $\mathbf{Q}(\boldsymbol{\delta})$ estan dados por

$$\begin{aligned} q_{kk} &= \kappa_{k0} + \mathbf{u}_k^\intercal \boldsymbol{\kappa}_{\boldsymbol{/0}} \mathbf{u}_k \\ &= \kappa_{k0} + \sum_{\substack{j=1 \\ j \neq k}}^{c} \kappa_{kj} \\ &= \sum_{j \neq k} \kappa_{kj} \end{aligned}$$

Apéndice B

Teorema 2.3 (Rue y Held, 2005)

Sea $\mathbf{x}$ un campo aleatorio de Gauss Markov respecto a $G = (\nu, \varepsilon)$ con media μ y matriz de precisión $\mathbf{P} > 0$, entonces podemos definir a la precisión parcial y la correlación parcial de la siguiente manera:

$$Prec(x_i|\mathbf{x}_{/i}) = P_{ii},$$
$$Cor(x_i, x_j|\mathbf{x}_{/ij}) = -\frac{P_{ij}}{\sqrt{P_{ii}P_{jj}}};\ \text{si } i \neq j.$$

Los elementos de la diagonal de $\mathbf{P}$ son las precisiones condicionales de x_i, dado los valores de las variables diferentes de x_i (denotados por $\mathbf{x}_{/i}$); mientras que los elementos fuera de la diagonal, con el escalamiento adecuado, brindan información sobre la correlación condicional entre x_i y x_j, tomando como fijos los valores de las variables diferentes de x_i y x_j (denotados por $\mathbf{x}_{/ij}$). Estos resultados pueden ser comparados con la interpretación de los elementos de la matriz de covarianza $\mathbf{\Sigma}$; dado que $Var(x_i) = \Sigma_{ii}$ y $Cor(x_i, x_j) = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$; la matriz de covarianza nos da información sobre la varianza marginal de x_i y la correlación marginal entre x_i y x_j. La interpretación marginal dada por $\mathbf{\Sigma}$ es intuitiva y directamente informativa, dado que reduce la dimensionalidad original de una distribución para poder interpretarla en una o dos dimensiones. La interpretación marginal a partir de $\mathbf{P}$ es más complicada, dado que se debe integrar respecto de $\mathbf{x}_{/i}$ o $\mathbf{x}_{/ij}$ a la distribución conjunta parametrizada en términos de $\mathbf{P}$; sin embargo, en términos matriciales esto es más sencillo, pues por definición $\mathbf{P}^{-1} = \mathbf{\Sigma}$, y Σ_{ii} generalmente depende de todos los elementos de $\mathbf{P}$, y viceversa.

Apéndice C

Código del modelo en JAGS

A continuación, el código para un modelo que considere una covariable ordinal, una nominal y una numérica:

```
M.fus_i <- function() {
  #Verosimilitud
  for(i in 1:n){
    y[i]~dgamma(omega,omega/exp(mu[i]))
    mu[i] = b0+inprod(X01[i,],beta01)+inprod(X02[i,],beta02)+inprod(X03[i,],beta03)

  }
  #####Priori
  ####################matriz_estructura ORDINAL 1
  pi1[1,2]~dbeta(1,1)
  delta1[1,2]~dbern(pi1[1,2])
  kappa1[1,2]<-delta1[1,2]+r*(1-delta1[1,2])
  kappa1[2,1]<-kappa1[1,2]
  for(k in 2:c1){
    pi1[k,k+1]~dbeta(1,1)
    delta1[k,k+1]~dbern(pi1[k,k+1])
    kappa1[k,k+1]<-(delta1[k,k+1]+r*(1-delta1[k,k+1]))
    kappa1[k+1,k]<-kappa1[k,k+1]
    Q1[k-1,k]<--kappa1[k,k+1]
    Q1[k,k-1]<--kappa1[k,k+1]
    for(j in (k+2):(c1+1)){
      kappa1[k,j]<-0
      kappa1[j,k]<-0
      Q1[k-1,j-1]<--kappa1[k,j]
      Q1[j-1,k-1]<--kappa1[k,j]
    }
  }
  for(k in 2:(c1+1)){
    kappa1[k,k]<-0
  }
  for(k in 2:(c1)){
    Q1[k-1,k-1]<-sum(kappa1[k,k-1],kappa1[k,k+1])
  }
  Q1[c1,c1]<-kappa1[c1,c1+1]
  ####################matriz_estructura NOMINAL 2
  for(j in 2:(c2+1)){
    pi2[1,j]~dbeta(1,1)
    delta2[1,j]~dbern(pi2[1,j])
    kappa2[1,j]<-delta2[1,j]+r*(1-delta2[1,j])
    kappa2[j,1]<-kappa2[1,j]
  }
  for(k in 2:c2){
    for(j in (k+1):(c2+1)){
```

```
      pi2[k,j]~dbeta(1,1)
      delta2[k,j]~dbern(pi2[k,j])
      kappa2[k,j]<-delta2[k,j]+r*(1-delta2[k,j])
      kappa2[j,k]<-kappa2[k,j]
      Q2[k-1,j-1]<--kappa2[k,j]
      Q2[j-1,k-1]<--kappa2[k,j]
    }
  }
  for(k in 2:(c2+1)){
    kappa2[k,k]<-0
  }
  for(k in 2:(c2+1)){
    Q2[k-1,k-1]<-sum(kappa2[k,1:(c2+1)])
  }
  ####################efectos
  beta01[1:c1]~dmnorm(rep(0,c1),Q1[1:c1,1:c1]*t1/gamma1)
  beta02[1:c2]~dmnorm(rep(0,c2),Q2[1:c2,1:c2]*t2/gamma2)
  beta03~dnorm(0,1/10000)
  ####################intercepto
  b0~dnorm(0,1/B0)
  ####################Precision
  omega~dgamma(s0,S0)
  ####################var.efectos
  t1~dgamma(g01,G01)
  tau1=1/t1
  t2~dgamma(g02,G02)
  tau2=1/t2
}

########################################
# Lista con los datos

d  = within(list(), {
  y = c(y)
  X01 = X01
  X02 = X02
  X03 = X03
  n = length(y)
  c1 = ifelse(is.null(dim(X01)),1,ncol(X01))
  c2 = ifelse(is.null(dim(X02)),1,ncol(X02))
  r=20000
  s0=0.001
  S0=0.001
  g01=5
  G01=100
  gamma1=1  #ordinal
  g02=5
  G02=100
  gamma2=c2/2  #nominal
  B0=10000
})

########################################
# Estimación

M.fus_i.r <- jags.parfit(cl, data=d, params=c("b0","omega", "beta01","beta02","tau1","
    tau2","delta1","delta2"),
          model=M.fus_i,
          n.chains=4, n.adapt=10000, n.update=20000, n.iter=20000, thin=4 )
```

Apéndice D

Indicadores de evaluación del modelo

Para evaluar la selección del modelo final en las simulaciones, se utilizaron indicadores a partir de una metodología similar al de una matriz de confusión para cada covariable; donde se comparan los indicadores de fusión planteados en el modelo teórico (δ'_{kj}) con los indicadores de fusión a posteriori (δ_{kj}):

		Modelo seleccionado: 0	Modelo seleccionado: 1
Modelo teórico	0	TN	FP
	1	FN	TP

Cuadro D.1: Matriz de confusión.

donde TP (verdaderos positivos) es el número de veces que se estimó correctamente que la diferencia de dos efectos era diferente de cero (no fusión), TN (verdaderos negativos) es el número de veces que se estimó correctamente las diferencias de efectos iguales a cero (fusión), FN (falsos negativos) son los casos en los que el modelo teórico indicaba que la diferencia de los efectos era diferente de cero y fueron fusionados en la estimación, mientras que FP (falsos positivos) son aquellos casos en los que la diferencia de efectos era igual a cero según el modelo teórico y fueron estimados como diferentes de cero.

A partir de estos valores se pueden construir distintos indicadores para evaluar la selección del modelo; tales como el ratio de verdaderos positivos (TPR), el ratio de verdaderos negativos (TNR), el ratio de falsos positivos (FPR), el ratio de falsos negativos (FNR), el valor predictivo positivo (PPV) y el valor predictivo negativo (NPV), los cuales se definen de la siguiente manera:

$$TPR = [TP/(TP + FN)]100\,\%$$
$$TNR = [TN/(TN + FP)]100\,\%$$
$$FPR = 1 - TNR$$
$$FNR = 1 - TPR$$
$$PPV = [TP/(TP + FP)]100\,\%$$
$$NPV = [TN/(TN + FN)]100\,\%$$

Apéndice E

Definición de variables para el caso aplicativo

La construcción de la variable respuesta y de algunas covariables se realizó siguiendo el procedimiento descrito por Baca (2019), con la diferencia de que las covariables nominales no fueron dicotomizadas a criterio del investigador.

Campo	Descripción	Tipo	Código	Etiqueta
ingreso_h	Ingreso por hora[1]	Numérica	-	
empleados	Número de empleados en la empresa donde trabaja	Ordinal	1	De 1 a 9
			2	De 10 a 20
			3	De 21 a 50
			4	De 51 a 100
			5	De 101 a 500
			6	De 501 a más
calidad	Calidad del centro de estudios de procedencia[2]	Nominal	1	Sin puntaje
			2	No top
			3	Top
P209	Estado civil	Nominal	1	Conviviente
			2	Casado
			3	Viudo
			4	Divorciado o separado
			5	Soltero
P300A	Grupo étnico según lengua materna	Nominal	1	Quechua
			2	Aymara
			3	Castellano
sector	Sector productivo	Nominal	1	Primario
			2	Secundario
			3	Terciario
ocupación	Tipo de ocupación	Nominal	1	Profesionales
			2	Técnicos
			3	Jefes, adiministrativos y directivos
			4	Servicios y vendedores

(sigue en la página siguiente)

Campo	Descripción	Tipo	Código	Etiqueta
			5	Agricultura, pecuario, etc.
			6	Construcción
			7	Operadores y transportes
			8	Servicios elementales
vivienda_t	Tipo de vivienda	Nominal	1	Otro
			2	Propia
			3	Alquilada
mujer	Sexo: Mujer	Dicotómica	1	Mujer
			0	Varón
area_u	Estrato geográfico urbano	Dicotómica	1	Urbano
			0	Rural
sindicato	Pertenece a un sindicato	Dicotómica	1	Sí pertenece
			0	No pertenece
añoseduz	Años de educación (estandarizado)[3]	Numérica	-	
expempz	Años de experiencia en la empresa donde trabaja (estandarizado)	Numérica	-	
exp_generalz	Años de experiencia potencial (estandarizado)[4]	Numérica	-	

(Fin de la tabla)

[1]Es el monto monetario por concepto de salarios, ganancias, honorarios y comisiones que recibe un individuo. Se hará uso de la preguntas P523 y P530 de la ENAHO.

[2]Se consideran top a las universidades ubicadas en el 30 % superior del ranking Web of Science 2020 de MINEDU, no top son las universidades en el 70 % inferior del ranking y sin puntuación son las universidades fuera del ranking u otros centros de estudio superior.

[3]Para construir esta variable se utilizarán el nivel educativo y los años de estudio cursados. Se utilizarán las siguientes fórmulas según nivel educativo alcanzado: Ninguno=0; Primaria=0+u; Secundaria=6+u; Superior técnica=11+u; Universitaria=11+u; Posgrado=16+u. Donde 'u' es el último año cursado en el nivel respectivo.

[4]Se calcula restando a la edad los años de educación y 6 años adicionales.

Apéndice F

Complementos del caso aplicativo

F.1. Elección del ratio de precisión para el modelo final

Para determinar el valor del ratio de precisión r a utilizar en el caso aplicativo se estimaron dos modelos; uno con $r = 10000$ y otro con $r = 20000$. Se escogerá valor de r que consiga el menor valor del DIC en el modelo final (resultado de la fusión de efectos).

	$r = 10000$	$r = 20000$
DIC	23237.29	23097.71
ρ_D	14.83	17.03

Cuadro F.1: DIC del modelo final para distintos valores de r.

F.2. Comparación del modelo completo y el modelo final

	Modelo completo	Modelo final
DIC	23120.32	23097.71
ρ_D	30.48	17.03

Cuadro F.2: DIC del modelo completo y del modelo final.

F.3. Convergencia del modelo de fusión de efectos

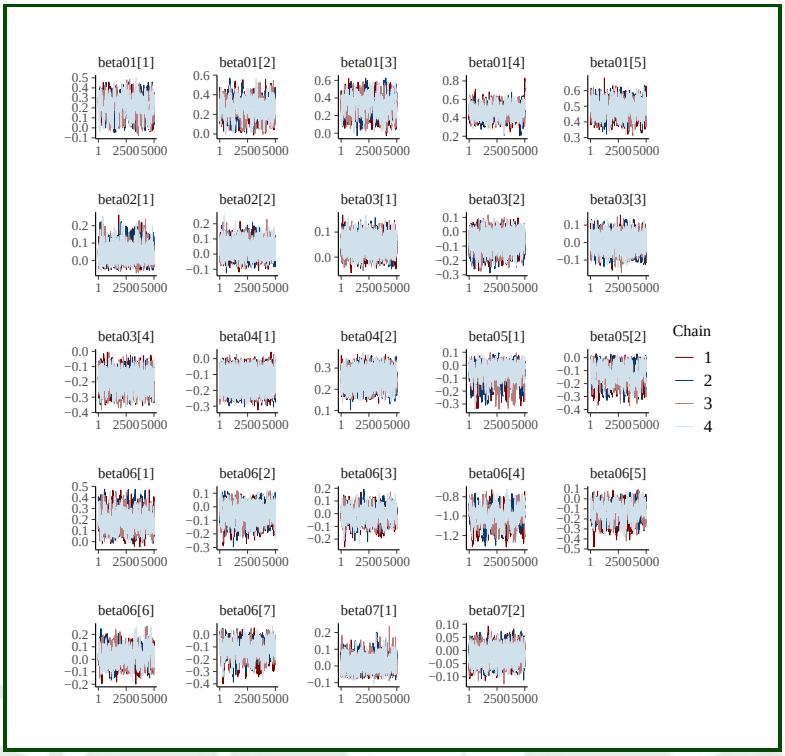


Figura F.1: Convergencia de los coeficientes $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_3}$, $\boldsymbol{\beta_4}$, $\boldsymbol{\beta_5}$, $\boldsymbol{\beta_6}$ y $\boldsymbol{\beta_7}$.

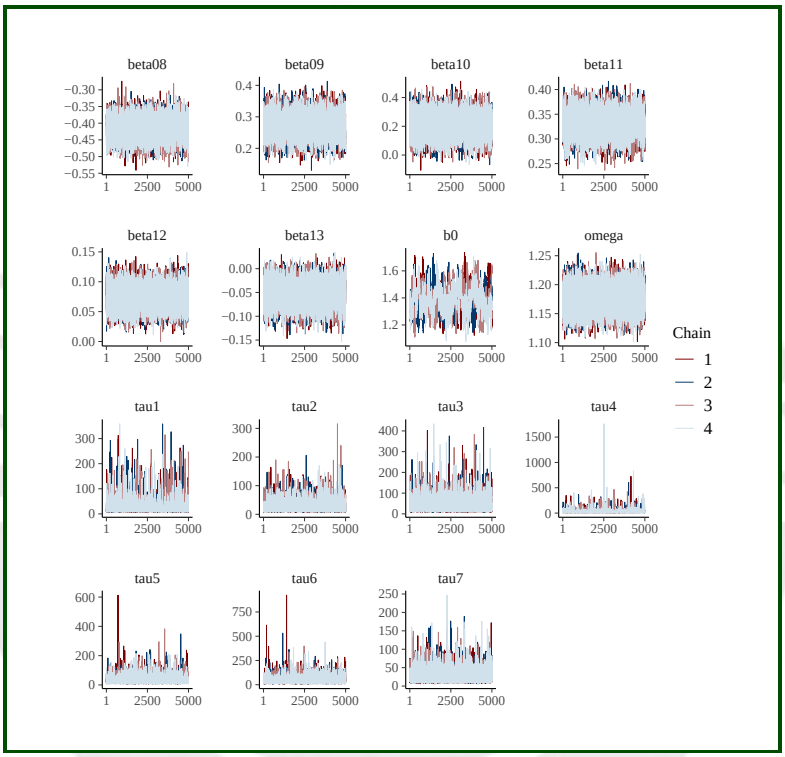


Figura F.2: Convergencia de los coeficientes β_8, β_9, β_{10}, β_{11}, β_{12}, β_{13}, β_0, ω y τ_h^2.

F.4. Convergencia del modelo final

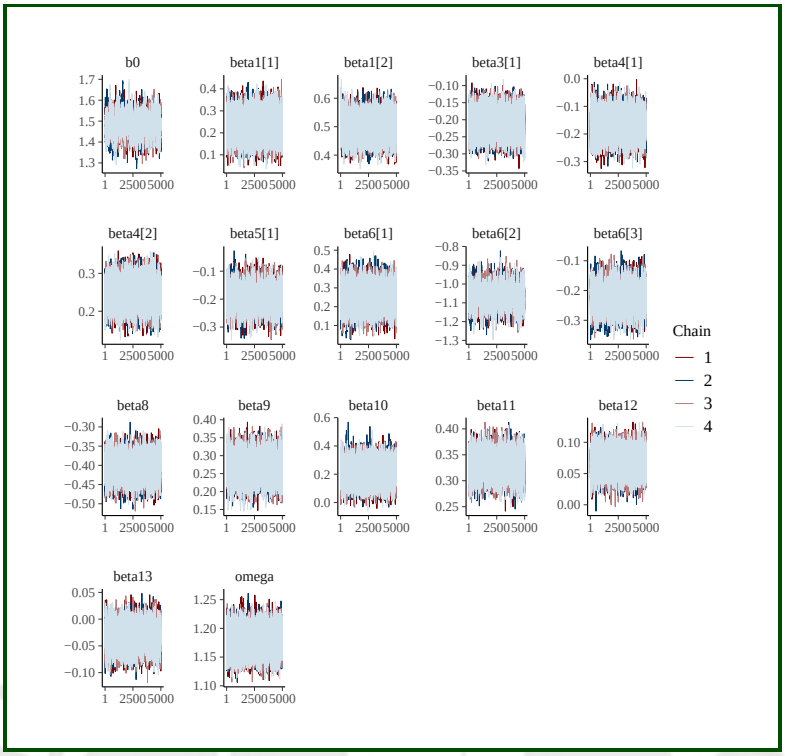


Figura F.3: Convergencia del modelo final.

F.5. Estimación de parámetros del modelo final

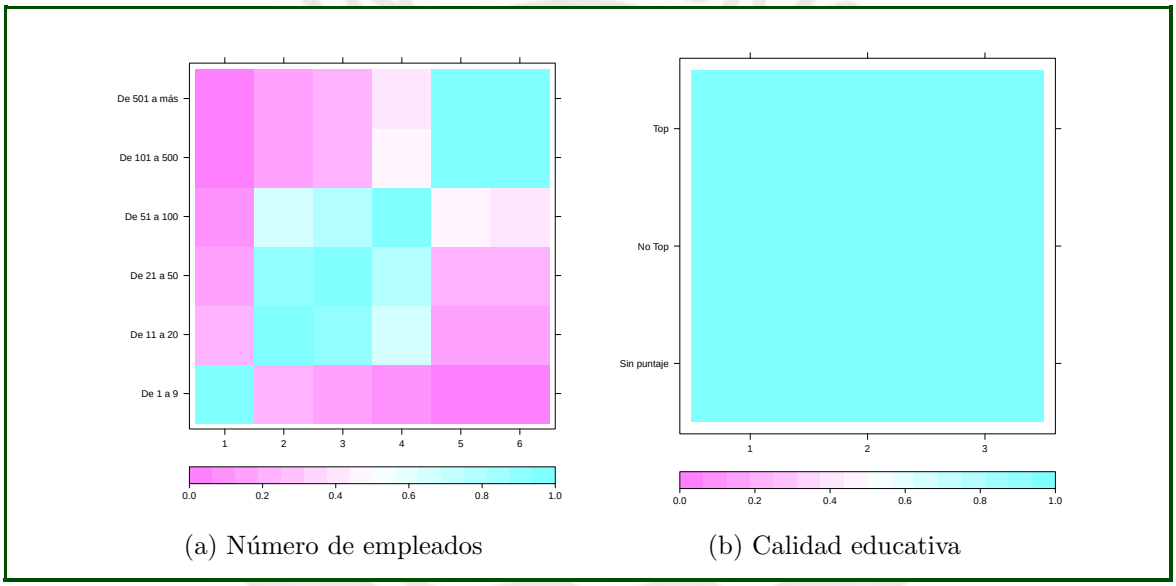
	Modelo completo			Modelo final		
	Media a posteriori	Intervalo HPD (95 %)		Media a posteriori	Intervalo HPD (95 %)	
Intercepto	1.392	1.209	1.589	1.482	1.378	1.583
Número de empleados						
De 10 a 20	0.216	0.004	0.368	0.231	0.128	0.337
De 21 a 50	0.263	0.066	0.431	0.231	0.128	0.337
De 51 a 100	0.316	0.124	0.516	0.231	0.128	0.337
De 101 a 500	0.470	0.352	0.594	0.502	0.428	0.578
De 501 a más	0.482	0.390	0.569	0.502	0.428	0.578
Calidad educativa						
No top	0.035	-0.033	0.102	-	-	-
Top	0.033	-0.044	0.115	-	-	-
Estado civil						
Casado	0.050	-0.006	0.111	-	-	-
Viudo	-0.062	-0.170	0.038	-0.208	-0.270	-0.143
Divorciado o separado	-0.002	-0.079	0.075	-	-	-
Soltero	-0.192	-0.300	-0.095	-0.208	-0.270	-0.143
Grupo étnico						
Aymara	-0.136	-0.241	-0.032	-0.168	-0.245	-0.087
Castellano	0.253	0.188	0.320	0.246	0.184	0.309
Sector						
Secundario	-0.058	-0.216	0.043	-	-	-
Terciario	-0.129	-0.258	-0.002	-0.193	-0.272	-0.111
Ocupación						
Técnicos	0.189	0.039	0.350	0.247	0.131	0.365
Jefes y directivos	-0.052	-0.165	0.059	-	-	-
Servicios y vendedores	-0.011	-0.120	0.094	-	-	-
Agricultura, pecuario, etc.	-1.004	-1.192	-0.828	-1.062	-1.171	-0.953
Construcción	-0.124	-0.285	0.015	-0.223	-0.303	-0.145
Operadores y transportes	0.026	-0.092	0.148	-	-	-
Servicios elementales	-0.123	-0.250	0.001	-0.223	-0.303	-0.145
Tipo de vivienda						
Propia	0.015	-0.047	0.081	-	-	-
Alquilada	-0.015	-0.066	0.034	-	-	-
Sexo						
Mujer	-0.416	-0.475	-0.357	-0.407	-0.463	-0.354
Estrato geográfico						
Urbano	0.274	0.208	0.342	0.268	0.203	0.333
Sindicato						
Sí pertenece	0.205	0.065	0.351	0.209	0.071	0.349
Años de educación	0.329	0.283	0.375	0.330	0.286	0.370
Experiencia en la empresa	0.073	0.040	0.109	0.068	0.033	0.102
Experiencia potencial	-0.054	-0.102	-0.010	-0.032	-0.075	0.008
Precisión ω	1.177	1.138	1.216	1.178	1.139	1.218

Cuadro F.3: Estimación de parámetros: media a posteriori.

Apéndice G

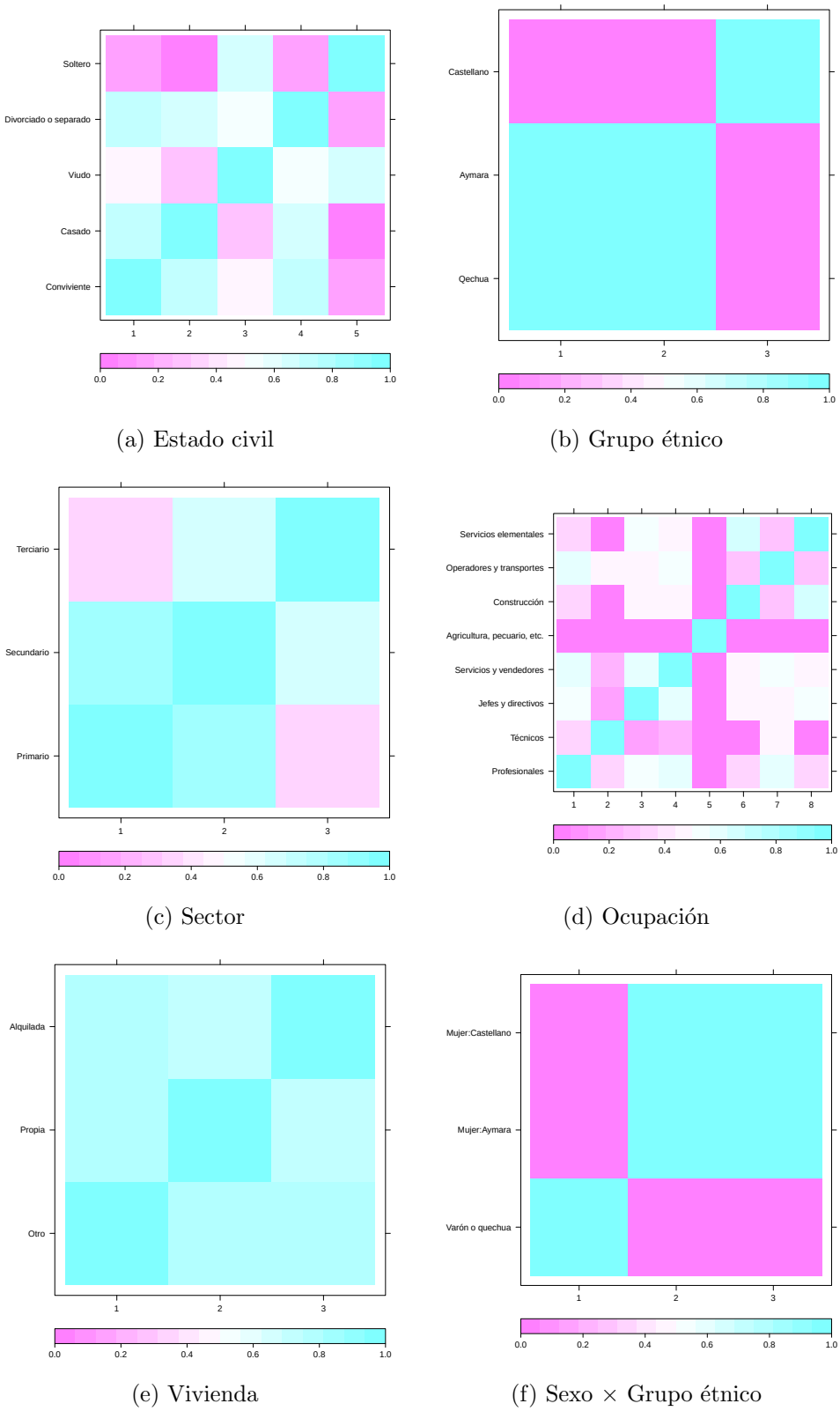
Complementos del caso aplicativo con interacción entre el sexo y el grupo étnico

G.1. Probabilidades de fusión a posteriori del modelo con interacción entre el sexo y el grupo étnico



(a) Número de empleados (b) Calidad educativa

Figura G.1: Probabilidad de fusión a posteriori para covariables ordinales del modelo con interacción entre el sexo y el grupo étnico.



(a) Estado civil

(b) Grupo étnico

(c) Sector

(d) Ocupación

(e) Vivienda

(f) Sexo × Grupo étnico

Figura G.2: Probabilidad de fusión a posteriori para las covariables nominales del modelo con interacción entre el sexo y el grupo étnico.

G.2. Agrupación de niveles en el modelo con interacción entre el sexo y el grupo étnico

	Media a posteriori	Agrupación
Número de empleados		
De 1 a 9 (referencia)	-	1
De 10 a 20	0.201	2
De 21 a 50	0.246	2
De 51 a 100	0.327	2
De 101 a 500	0.471	3
De 501 a más	0.482	3
Calidad educativa		
Sin puntaje (referencia)	-	1
No top	0.032	1
Top	0.028	1
Estado civil		
Conviviente (referencia)	-	1
Casado	0.049	1
Viudo	-0.088	2
Divorciado o separado	-0.004	1
Soltero	-0.186	2
Grupo étnico		
Qechua (referencia)	-	1
Aymara	-0.024	1
Castellano	0.368	2
Sector		
Primario (referencia)	-	1
Secundario	-0.042	1
Terciario	-0.111	1
Ocupación		
Profesionales (referencia)	-	1
Técnicos	0.169	2
Jefes y directivos	-0.047	1
Servicios y vendedores	-0.029	1
Agricultura, pecuario, etc.	-0.988	3
Construcción	-0.131	4
Operadores y transportes	0.022	1
Servicios elementales	-0.122	4
Tipo de vivienda		
Otro (referencia)	-	1
Propia	0.013	1
Alquilada	-0.014	1
Sexo × Grupo étnico		
Varón o Quechua (referencia)	-	1
Mujer:Aymara	-0.275	2
Mujer:Castellano	-0.277	2

Cuadro G.1: Agrupación de efectos bajo el modelo de fusión con interacción entre sexo y grupo étnico.

Apéndice H

Interacción completa entre el número de empleados y el grupo étnico

H.1. Niveles de la interacción entre el número de empleados y el grupo étnico

k	Número de empleados × Grupo étnico
0	De 1 a 9:Quechua (referencia)
1	De 10 a 20:Aymara
2	De 21 a 50:Aymara
3	De 51 a 100:Aymara
4	De 101 a 500:Aymara
5	De 501 a más:Aymara
6	De 10 a 20:Castellano
7	De 21 a 50:Castellano
8	De 51 a 100:Castellano
9	De 101 a 500:Castellano
10	De 501 a más:Castellano

Cuadro H.1: Niveles de la interacción entre el número de empleados y el grupo étnico.

H.2. Matriz estructura para la interacción entre el número de empleados y el grupo étnico

$$\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta}) = \begin{pmatrix} q_{11} & -\kappa_{12} & 0 & 0 & 0 & -\kappa_{16} & -\kappa_{17} & 0 & 0 & 0 \\ -\kappa_{21} & q_{22} & -\kappa_{23} & 0 & 0 & -\kappa_{26} & -\kappa_{27} & -\kappa_{28} & 0 & 0 \\ 0 & -\kappa_{32} & q_{33} & -\kappa_{34} & 0 & 0 & -\kappa_{37} & -\kappa_{38} & -\kappa_{39} & 0 \\ 0 & 0 & -\kappa_{43} & q_{44} & -\kappa_{45} & 0 & 0 & -\kappa_{48} & -\kappa_{49} & -\kappa_{4,10} \\ 0 & 0 & 0 & -\kappa_{54} & q_{55} & 0 & 0 & 0 & -\kappa_{59} & -\kappa_{5,10} \\ -\kappa_{61} & -\kappa_{62} & 0 & 0 & 0 & q_{66} & -\kappa_{67} & 0 & 0 & 0 \\ -\kappa_{71} & -\kappa_{72} & -\kappa_{73} & 0 & 0 & -\kappa_{76} & q_{77} & -\kappa_{78} & 0 & 0 \\ 0 & -\kappa_{82} & -\kappa_{83} & -\kappa_{84} & 0 & 0 & -\kappa_{87} & q_{88} & -\kappa_{89} & 0 \\ 0 & 0 & -\kappa_{93} & -\kappa_{94} & -\kappa_{95} & 0 & 0 & -\kappa_{98} & q_{99} & -\kappa_{9,10} \\ 0 & 0 & 0 & -\kappa_{10,4} & -\kappa_{10,5} & 0 & 0 & 0 & -\kappa_{10,9} & q_{10,10} \end{pmatrix},$$

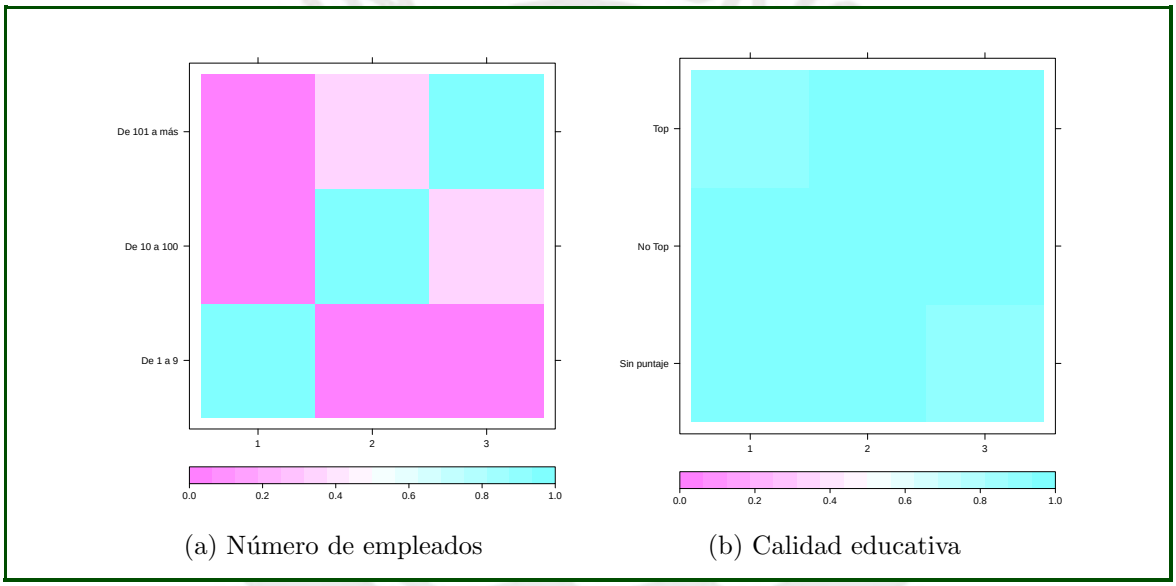
donde los elementos de la diagonal se definen de la siguiente manera:

$$q_{kk} = \begin{cases} \kappa_{10} + \kappa_{60} - \sum_{k \neq j} q_{kj} & , \text{ si } k \in \{1, 6\}, \\ -\sum_{k \neq j} q_{kj} & , \text{ si } k \notin \{1, 6\}. \end{cases}$$

Apéndice I

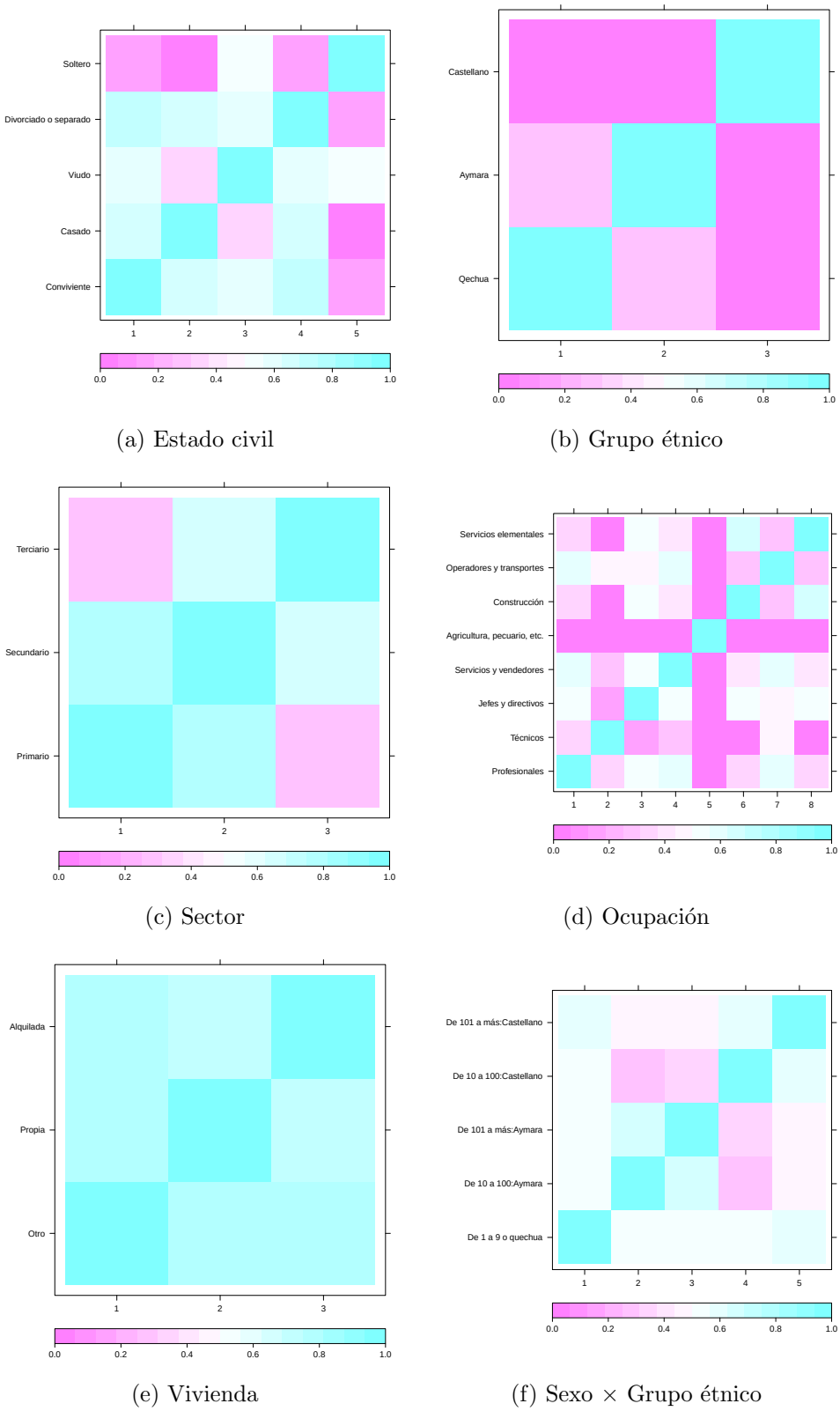
Complementos del caso aplicativo con interacción entre el número de empleados y el grupo étnico

I.1. Probabilidades de fusión a posteriori del modelo con interacción entre el número de empleados y el grupo étnico



(a) Número de empleados (b) Calidad educativa

Figura I.1: Probabilidad de fusión a posteriori para covariables ordinales del modelo con interacción entre el número de empleados y el grupo étnico.



(a) Estado civil

(b) Grupo étnico

(c) Sector

(d) Ocupación

(e) Vivienda

(f) Sexo × Grupo étnico

Figura I.2: Probabilidad de fusión a posteriori para las covariables nominales del modelo con interacción entre el número de empleados y el grupo étnico.

I.2. Agrupación de niveles en el modelo con interacción entre el número de empleados y el grupo étnico

	Media a posteriori	Agrupación
Número de empleados		
De 1 a 9 (referencia)	-	1
De 10 a 100	0.324	2
De 101 a más	0.513	3
Calidad educativa		
Sin puntaje (referencia)	-	1
No top	0.038	1
Top	0.037	1
Estado civil		
Conviviente (referencia)	-	1
Casado	0.051	1
Viudo	-0.066	2
Divorciado o separado	-0.003	1
Soltero	-0.193	2
Grupo étnico		
Qechua (referencia)	-	1
Aymara	-0.154	2
Castellano	0.273	3
Sector		
Primario (referencia)	-	1
Secundario	-0.051	1
Terciario	-0.128	2
Ocupación		
Profesionales (referencia)	-	1
Técnicos	0.182	2
Jefes y directivos	-0.050	1
Servicios y vendedores	-0.011	1
Agricultura, pecuario, etc.	-0.996	3
Construcción	-0.129	4
Operadores y transportes	0.025	1
Servicios elementales	-0.134	4
Tipo de vivienda		
Otro (referencia)	-	1
Propio	0.014	1
Alquilado	-0.016	1
Número de empleados × Grupo étnico		
De 1 a 9 o Quechua (referencia)	-	1
De 10 a 100:Aymara	0.032	2
De 101 a más:Aymara	0.051	2
De 10 a 100:Castellano	-0.107	1
De 101 a más:Castellano	-0.060	1

Cuadro I.1: Agrupación de efectos bajo el modelo de fusión con interacción entre el número de empleados y el grupo étnico.

Bibliografía

Baca, J. E. (2019). *Efecto de la calidad de la educación superior universitaria en las brechas étnicas de ingreso en el perú durante el periodo 2014-2017*, Tesis de maestría, Pontificia Universidad Católica del Perú.

Bondell, H. D. y Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in anova, *Biometrics* **65**(1): 169–177.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2008.01061.x*

Dellaportas, P. y Tarantola, C. (2005). Model determination for categorical data with factor level merging, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 269–283.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00501.x*

Fahrmeir, L., Kneib, T. y Konrath, S. (2010). Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection, *Statistics and Computing* **20**: 203–219.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. y Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Gertheiss, J. y Tutz, G. (2009). Penalized regression with ordinal predictors, *International Statistical Review* **77**(3): 345–365.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2009.00088.x*

Higham, N. J. (2002). *Accuracy and Stability of Numerical Algorithms*, 2nd edn, Society for Industrial and Applied Mathematics, USA.

Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*, Springer Texts in Statistics, Springer New York.

Ishwaran, H. y Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies, *Ann. Statist.* **33**(2): 730–773.
URL: *https://doi.org/10.1214/009053604000001147*

Kyung, M., Gill, J., Ghosh, M. y Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos, *Bayesian Anal.* **5**(2): 369–411.
URL: *https://doi.org/10.1214/10-BA607*

Lau, J. W. y Green, P. J. (2007). Bayesian model-based clustering procedures, *Journal of Computational and Graphical Statistics* **16**(3): 526–558.

Li, Q. y Lin, N. (2010). The bayesian elastic net, *Bayesian Anal.* **5**(1): 151–170.
URL: *https://doi.org/10.1214/10-BA506*

Malsiner-Walli, G., Pauger, D. y Wagner, H. (2018). Effect fusion using model-based clustering, *Statistical Modelling* **18**(2): 175–196.
URL: *https://doi.org/10.1177/1471082X17739058*

Mitchell, T. J. y Beauchamp, J. J. (1988). Bayesian variable selection in linear regression, *Journal of the American Statistical Association* **83**(404): 1023–1032.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694*

Park, T. y Casella, G. (2008). The bayesian lasso, *Journal of the American Statistical Association* **103**(482): 681–686.
URL: *https://doi.org/10.1198/016214508000000337*

Pauger, D. y Wagner, H. (2017). Bayesian effect fusion for categorical predictors.
URL: *https://arxiv.org/pdf/1703.10245.pdf*

Pauger, D. y Wagner, H. (2019). Bayesian effect fusion for categorical predictors, *Bayesian Analysis* **14**(2): 341–369.
URL: *https://doi.org/10.1214/18-BA1096*

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling, *Proceedings of the 3rd international workshop on distributed statistical computing*, Vol. 124, Vienna, Austria, pp. 1–10.

Rue, H. y Held, L. (2005). *Gaussian Markov random fields: theory and applications*, CRC press.

Sal y Rosas, V., Moscoso-Porras, M., Ormeño, R., Artica, F., Bayes, C. y Miranda, J. (2019). Gender income gap among physicians and nurses in Peru: a nationwide assessment, *The Lancet Global Health* **7**(4): e412–e413.
URL: *https://www.sciencedirect.com/science/article/pii/S2214109X19300348*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1): 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. y Knight, K. (2005). Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1): 91–108.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00490.x*

Tutz, G. y Gertheiss, J. (2016). Regularized regression for categorical data, *Statistical Modelling* **16**(3): 161–200.
URL: *https://doi.org/10.1177/1471082X16642560*

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301–320.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x*