PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Método para la Fusión de Categorías usando Técnicas de Agrupamiento

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

Victor Daniel Farro Diaz

Asesor:

Cristian Luis Bayes Rodriguez

Lima, 2021

Dedicatoria

A mis queridos padres, Francisco y Rosa, por su apoyo en todo momento de mi vida y por animarme a seguir estudiando.

Espero llenarlos de orgullo.

Agradecimentos

A Dios, por haberme guiado en cada etapa de mi vida.
A mis padres Francisco y Rosa, por el esfuerzo realizado de brindarme una educación de calidad, y una vida llena de valores y mucho amor. Ambos son un excelente ejemplo de vida, para mí y mis hermanos.
A mi hermana Marivel y mi hermano Javier por siempre confiar en mi.
A mi familia Yamileth y Danna, por apoyarme en conseguir este objetivo y brindarme el aliento para seguir adelante.
Al profesor Cristian Bayes, por su paciencia en sus orientaciones.
A los profesores Wilmer Atoche, Carlos Vera, Carlos Hernández, Luis Valdivieso y Daniel Grados quienes me motivaron a seguir esforzándome.
A mis amigos Claudia Valencia, Yuri Alvarado y Luis Espinoza, por compartir mi entusiasmo con el desarrollo de la tesis.

Resumen

En la actualidad, muchas organizaciones disponen o tienen acceso a una gran cantidad y variedad de datos que les permiten tomar decisiones acordes en temas económicos, sociales, de educación, de salud, entre otros. Con frecuencia, los estudios que se realizan se enfocan en el objetivo de explicar una variable de interés utilizando un conjunto de variables explicativas; y si la relación de dependencia es lineal, se le conoce como modelo de regresión lineal.

Los modelos de regresión lineal presentan su principal reto en la estimación de los parámetros de la regresión, que se consiguen a partir de la información obtenida mediante el análisis de las observaciones de una muestra previamente recogida. La complejidad de los modelos de regresión lineal aumenta con la existencia de covariables que son medidas en una escala nominal u ordinal, y que en muchas ocasiones presentan una gran cantidad de categorías, como por ejemplo: estado civil, grupo sanguíneo, entre otros. Lo habitual para modelar el efecto total de una covariable categórica es definir una categoría (o nivel) como línea base y utilizar variables ficticias para las otras categorías (o niveles).

La presente tesis tiene como principal objetivo el desarrollo del método de fusión de efectos de covariables categóricas usando técnicas de agrupamiento PAM, propuesto por Malsiner-Walli, Pauger y Wagner (2018), y aplicarlo en un conjunto de datos reales relacionados a los ingresos monetarios de la población de Lima Metropolitana y Callao del primer trimestre del 2020.

Palabras-clave: modelos de regresión lineal, covariable categórica, fusión de efectos, mixtura finita, distribución de Dirichlet, MCMC.

Índice general

Lista de abreviaturas

DIC	Criterio de Información del Desvío (*Deviance information criterion*).
MCMC	Cadenas de Markov de Monte Carlo (*Markov Chain Monte Carlo*).
PAM	Partición alrededor del medoide (*Partitioning Around Medoid*).

Lista de símbolos

y_i	respuestas continuas de las i observaciones del modelo de regresión.
β_0	intercepto del modelo de regresión.
β_{jk}	efecto de nivel de la k-ésima categoría de la covariable j.
X_{jk}	variable ficticia de la k-ésima categoría de la covariable j.
ϵ_i	error del modelo de regresión lineal.
σ^2	varianza del término error ϵ_i.
η_{jl}	peso del componente de la mixtura finita de normales del efecto l de la covariabe j.
μ_{jl}	parámetro de localización (media) del componente de la mixtura finita de normales, de efecto l de la covariabe j.
ψ_j	parámetro de escala (varianza) del componente de la mixtura finita de normales, de la covariabe j.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

En la actualidad, muchas organizaciones públicas y privadas disponen de una gran cantidad y variedad de datos que les permiten obtener información relevante sobre temas económicos, sociales, de educación, de salud, entre otros; consiguiendo de esa forma, tomar decisiones acordes para la generación de impactos positivos en sus respectivos objetivos estratégicos. Con frecuencia, los estudios que se realizan se enfocan en el objetivo de explicar una variable de interés utilizando un conjunto de variables explicativas; y si la relación de dependencia que se establece entre dichas variables es lineal, se le conoce como un modelo de regresión lineal.

En ese sentido, se puede afirmar que la estimación de los modelos de regresión lineal tiene como principal objetivo relacionar una variable dependiente y otras variables independientes, que también son conocidas como explicativas, predictivas o covariables. Los modelos de regresión lineal presentan su principal reto en la estimación de los parámetros de la regresión, que se consiguen a partir de la información obtenida mediante el análisis de las observaciones de una muestra previamente recogida. Sin embargo, existen otras complicaciones en la estimación de los modelos regresión lineal, como la existencia de covariables que son medidas en una escala nominal u ordinal, y que en muchas ocasiones presentan una gran cantidad de categorías (por ello el nombre de variables categóricas), como por ejemplo: estado civil, grupo sanguíneo, cargo laboral, nivel socio-económico, entre otros.

Dado que una covariable categórica puede representar un efecto significativo sobre un modelo de regresión lineal, es necesario conocer la metodología para su tratamiento dentro de una estimación de estos modelos. Lo habitual para modelar el efecto total de las covariables categóricas es definir una categoría (o nivel) como línea base y utilizar variables ficticias (usualmente binarias) para las otras categorías (o niveles), permitiéndose de esta forma, la estimación de los efectos de dichas categorías teniendo como referencia a la línea base. Por ejemplo, supongamos que una covariable "nivel de educación básica" tiene 3 categorías: inicial, primaria y secundaria; en este caso, si definimos como línea base al nivel inicial, será necesario crear 2 variables ficticias binarias (P y S), que tomarán el valor de 1 cuando se

cumpla la categoría nivel primaria y secundaria respectivamente, y tomarán el valor de 0 en caso contrario; adicionalmente se puede entender que, la categoría nivel inicial se cumple cuando ambas variables ficticias (P y S) toman el valor de 0. Por lo tanto, el efecto de una covariable categórica con $c+1$ categorías se captura dentro del modelo mediante un conjunto de c coeficientes (o efectos) de regresión.

Por esta razón, Pauger y Wagner (2017) mencionan que el efecto de una covariable categórica no se captura por un solo efecto del modelo de regresión sino por un grupo de efectos (efectos de nivel), y por lo tanto, introducir las variables categóricas como covariables en los modelos de regresión lineal pueden conducir fácilmente a un vector de efectos de regresión de alta dimensión y por ende, ser un problema crucial para la estimación de estos modelos. Por ello, Malsiner-Walli, Pauger y Wagner (2018) mencionan que sería interesante tener un método que permita comenzar con un modelo de regresión grande y posteriormente obtener una representación más reducida de este modelo, que se puede conseguir cuando los efectos de una variable categórica sean representados por menos de c coeficientes (o efectos) de regresión. Existen tres posibilidades de reducción en un modelo de regresión: 1) si todos los efectos de nivel son 0, la covariable completa se puede excluir del modelo; 2) si algunos de los efectos de nivel son 0, las categorías correspondientes se pueden excluir del modelo; y 3) si algunos efectos de nivel tienen el mismo o similar efecto en la variable dependiente, los efectos de estos niveles se pueden fusionar, y con ello se reduciría el modelo. Sin embargo, es importante mencionar que el uso de pocas categorías puede dar como resultado estimaciones imprecisas, mientras que el uso de categorías demasiado generales puede traer como consecuencia estimaciones sesgadas y de bajo rendimiento predictivo.

En Malsiner-Walli, Pauger y Wagner (2018) se realiza una revisión bibliográfica de los métodos usados para obtener una representación más reducida de un modelo de regresión lineal, desde el punto de vista frecuentista y bayesiano. La revisión bibliográfica se puede dividir en dos grandes grupos:

1. **Métodos para covariables cuantitativas:**
   Los métodos que se describen a continuación son usados para la selección de las covariables cuantitativas de un modelo de regresión.

   En el marco frecuentista: Tibshirani (1996) propone un método de selección y reducción de variables llamado lasso, el cual minimiza la suma de los residuos al cuadrado sujeto a que la suma del valor absoluto de los coeficientes sea menor que una constante. Por otro lado, Zou y Hastie (2005) proponen un método de regularización y selección de variables llamada 'elastic net', la cual fomenta un efecto de agrupación de las covariables fuertemente correlacionados, mediante la minimización de la suma de los residuos al cuadrado sujeto a una expresión conocida como 'elastic net' penalizada, que es una combinación convexa de la penalización del lasso. Es importante mencionar que la 'elastic net' es útil cuando el número de covariables p es mucho mayor que el número de observaciones n, a diferencia del lasso que no es un método de selección de variables

adecuado cuando $p \gg n$. Por esta razón, Tibshirani y Saunders (2005) proponen un método llamado lasso fusionado que permite la reducción de los modelos mediante la minimización de la suma de los residuos al cuadrado sujeto a dos restricciones: la primera fomenta la reducción en los coeficientes y la segunda fomenta la reducción en las diferencias de los coeficientes, buscando la uniformidad del modelo.

En el marco bayesiano: Park y Casella (2008) proponen el lasso bayesiano para la selección de variables utilizando una priori jerarquía expandida, con prioris normales conjugadas para los parámetros de regresión y prioris exponenciales independientes para sus varianzas. Por otro lado, Griffin y Brown (2010) proponen usar una distribución a priori normal-gamma sobre los coeficientes del modelo de regresión. Al existir varios métodos de selección de covariables, Kyung, Gill, Ghosh y Casella (2010) deciden realizar la comparación de los métodos lasso, elastic net y lasso fusionado utilizando los métodos de Gibbs y LARS. En el contexto de mixturas, Mitchell y Beauchamp (1988) proponen un método bayesiano de selección de variables, donde a cada coeficiente del modelo de regresión sujeto a eliminación se le define una distribución a priori que es una mixtura de masa puntual en 0 y una distribución uniforme difusa en otros lugares, es decir, una distribución de 'spike and slab', y al componente de error aleatorio del modelo de regresión se le asigna una distribución normal con media 0 y desviación estándar σ. Posteriormente, George y McCulloch (1993) proponen un método llamado SSVS (búsqueda estocástica para la selección de variables), que se basa en incorporar un modelo jerárquico de mixtura de normales en la estimación del modelo de regresión, donde se utilizan variables latentes para identificar las opciones de subconjunto de covariables prometedoras mediante consideraciones probabilísticas. Asimismo, George y McCulloch (1997) comparan varias formulaciones de prioris de mixtura jerárquica para la selección de variables en los modelos de regresión lineal. Por otro lado, Ishwaran y Rao (2005) proponen un método de selección de variables denominado modelo 'spike and slab' reescalado, que ayuda en la reducción de la incertidumbre del modelo de regresión. Después, Malsiner-Walli y Wagner (2011) deciden comparar prioris de 'spike and slab' para la selección de variables bayesianas.

Es preciso mencionar que todos estos métodos presentados no son apropiados para trabajar con covariables categóricas ya que sólo seleccionan o excluyen del modelo de regresión a covariables cuantitativas.

2. **Métodos para covariables categóricas:**
   Los métodos que se describen a continuación se pueden dividir en dos subgrupos:

   a) **Métodos para la exclusión de efectos de covariables categóricas:**
   Estos métodos son usados para la selección (o exclusión) de efectos de las covariables categóricas de un modelo de regresión.

Inicialmente, Chipman (1996) utiliza un enfoque bayesiano para la asignación de grados de confianza a cada modelo generado, que contiene relaciones entre las covariables. Posteriormente, Yuan y Lin (2006) presentan el método lasso de grupo y los compara con otros dos métodos de selección de subconjuntos de factores (o niveles). Asimismo, Raman, Fuchs, Wild, Dahl y Ro (2009) proponen un enfoque bayesiano al lasso de grupo, que es una extensión del lasso bayesiano para variables categóricas. Al existir diversos métodos de selección de variables, Kyung, Gill, Ghosh y Casella (2010) deciden realizar una comparación del desempeño de los lassos frecuentistas y bayesianos, donde prima la precisión de la predicción y la determinación de qué predictores son significativos, resultando un mejor rendimiento del lasso bayesiano frente al lasso frecuentista. Posteriormente, Simon, Friedman, Hastie y Tibshirani (2013) proponen un método conocido como lasso de grupo reducido.

b) **Métodos para la fusión de efectos de covariables categóricas:**
Estos métodos son usados para la selección y fusión de efectos de las covariables categóricas de un modelo de regresión.

En el marco frecuentista: Bondell y Reich (2009) proponen un método llamado CAS-ANOVA que permite determinar los factores que tienen un efecto significativo en la variable de respuesta y detectar diferencias entre los niveles de los factores significativos, de manera simultánea. Por otro lado, Gertheissy Tutz (2009), Gertheiss, Hogger, Oberhauser y Tutz (2010), Gertheiss y Tutz (2010), y Tutz y Gertheiss (2016) trabajan métodos con variables categóricas ordinales y proponen el uso de métodos de penalización sobre los efectos y la diferencia de efectos.

En el marco bayesiano: Pauger y Wagner (2017) proponen una distribución a priori que permite una dependencia casi perfecta como nula entre los efectos de nivel, especificando distribuciones a prioris de 'spike and slab' en todas las diferencias de efecto asociadas a una covariable categórica, reduciendo los efectos no relevantes a 0 y fusionando los efectos casi idénticos.

Todos estos métodos descritos de exclusión y fusión trabajan con covariables categóricas, sin embargo, tienen inconvenientes con covariables que cuentan con un número grande de categorías; ya que para una covariable con $c+1$ categorías se deben considerar $\binom{c+1}{2}$ posibles diferencias de efectos, ocasionando un método excesivamente grande de trabajar. Por tal razón, Dunson, Herring y Engel (2008) proponen un agrupamiento y selección bayesiana usando una mixtura a priori, con una masa puntual para exposiciones sin efecto y un proceso de Dirichlet multinivel para agrupar las exposiciones comunes no nulas. Asimismo, Malsiner-Walli, Frühwirth-Schnatter y Grün (2016) proponen un método para la agrupamiento basada en una mixtura finita reducida de normales multivariadas, usando prioris jerárquicos en los pesos de la mixtura y en las

medias de los componentes de la mixtura, con una normal-gamma como priori para las medias de los componentes de la mixtura y con un proceso de Dirichlet para los pesos de las mixturas.

Por todo lo expuesto, se puede concluir que la reducción de los efectos de nivel de las covariables categóricas es uno de los mayores desafíos en los modelos de regresión de alta dimensión. En ese sentido, Malsiner-Walli, Pauger y Wagner (2018) se interesan en abordar la fusión de efectos de las covariables categóricas y proponen una técnica de agrupamiento de los efectos de nivel β_{jk} (coeficiente de la variable ficticia x_{jk}, correspondiente a la k-ésima categoría de la covariable j), introduciendo una mixtura finita de normales como distribución a priori, con un esquema jerárquico, donde la mixtura finita cuenta con pesos η_{jl} para cada uno de sus componentes, y dichos pesos siguen una distribución de Dirichlet simétrica.

Adicionalmente, en Malsiner-Walli, Pauger y Wagner (2018) se explica que una posible debilidad del método presentado es la especificación a priori del número de componentes, y se plantea como alternativa una mixtura infinita con un proceso Dirichlet DP(α), donde el número de componentes se estimaría a partir de los datos. Sin embargo, en Malsiner-Walli, Frühwirth-Schnatter y Grün (2016) se supera esta debilidad de las mixturas finitas especificando una mixtura finita reducida como priori para los efectos de nivel; de donde se puede concluir que: 1) en las mixturas infinitas, al aumentar el número de categorías también aumenta el número de agrupaciones esperadas, y 2) en las mixturas finitas, el número de agrupaciones es independiente del número de categorías. Por lo tanto, el uso de una mixtura finita a priori es más adecuado para la reducción de un modelo.

El método presentado por Malsiner-Walli, Pauger y Wagner (2018) usa el método MCMC para la inferencia de la distribución a posteriori. El esquema MCMC plantea la iteración de dos grandes pasos: la regresión y el agrupamiento, que permiten analizar todas las posibilidades de agrupamiento para los efectos de las covariables, consiguiendo la evaluación de la incertidumbre y la estimación del modelo promedio. Es importante precisar que algunos investigadores tienen como objetivo la selección de un modelo único y la interpretación de sus resultados, y no se interesan por las estimaciones de un modelo promedio. También es importante mencionar que la selección de un modelo único para covariables categóricas es más complicada que la selección de covariables cuantitativas, ya que en el primer caso, el problema es determinar un agrupamiento apropiada de los efectos de nivel que incluye establecer el número de agrupaciones y la cantidad de miembros en cada agrupación.

Las principales ventajas del método presentado por Malsiner-Walli, Pauger y Wagner (2018), son: 1) trabajar con covariables categóricas con un gran número de categorías, dado que el método de fusión se enfoca sobre los mismos efectos de nivel y no sobre las diferencias de los efectos; 2) especificar los hiperparámetros de la mixtura finita de normales a priori en función de los datos; 3) evitar estimaciones computacionalmente intensas, ya que se puede utilizar el esquema MCMC debido al uso de una mixtura finita de normales y prioris conjugadas condicionalmente.

1.2. Objetivos

El objetivo general de la tesis es presentar el método de fusión de efectos de nivel de covariables categóricas usando la técnica de agrupamiento PAM propuesto por Malsiner-Walli, Pauger y Wagner (2018), y aplicarlo en un conjunto de datos reales relacionados a los ingresos monetarios de la población de Lima Metropolitana y Callao en el primer trimestre del 2020. Asimismo, los objetivos específicos que orientan la tesis son:

- Revisar literatura sobre los métodos de fusión de efectos en modelos de regresión lineal.
- Estudiar el método de fusión de efectos de nivel de covariables categóricas usando una mixtura finita de normales como distribución a priori y la técnica de agrupamiento PAM.
- Revisar la programación del paquete effectFusion del software R usado para la fusión de efectos de nivel.
- Realizar un estudio de simulación para evaluar el desempeño del método en diferentes escenarios.
- Aplicar el método desarrollado en un caso aplicativo con datos reales sobre ingresos monetarios de la población de Lima Metropolitana y Callao.
- Formular las conclusiones y recomendaciones del método presentado y su aplicación.

1.3. Organización del Trabajo

El contenido de la tesis consta de seis capítulos y dos apéndices. En el Capítulo 2 se presenta la mixtura finita de distribuciones normales, la distribución de Dirichlet simétrica y la técnica de agrupamiento PAM, como conceptos preliminares. En el Capítulo 3 se presenta el modelo a priori de mixtura finita de normales, las distribuciones a priori de sus parámetros y la inferencia a posteriori. En el Capítulo 4 se realiza un estudio de simulación con 4 covariables categóricas y 6 escenarios distintos de variabilidad en los componentes de la mixtura, con el fin de evaluar el desempeño del método desarrollado, mediante el uso del paquete effectFusion del software R. En el Capítulo 5 se muestra la aplicación del método de fusión de efectos de nivel de 8 covariables categóricas (recopiladas de la "Encuesta Permanente de Empleo, Trimestre móvil Ene-Feb-Mar 2020", aplicada en Lima Metropolitana y Callao en el primer trimestre del 2020) y los mismos 6 escenarios de variabilidad, con la implementación computacional que usa la técnica de agrupamiento PAM del paquete effectFusion del software R. Finalmente, en el Capítulo 6 se discuten algunas conclusiones obtenidas en este trabajo y se brindan algunas recomendaciones para futuras investigaciones.

Adicionalmente, se incluye un Apéndice A donde se muestra el código computacional para la realización del estudio de simulación de datos usando el paquete effectFusion del software R, y un Apéndice B donde se presenta el código computacional para la aplicación de datos reales usando el paquete effectFusion del software R.

Capítulo 2

Conceptos Preliminares

Para desarrollar el método de la fusión de efectos de nivel de las covariables categóricas haciendo uso de una mixtura finita de normales como priori y la técnica de agrupamiento PAM, propuesto por Malsiner-Walli, Pauger y Wagner (2018), es necesario primero explicar los modelos de mixtura finita de normales, la distribución de Dirichlet simétrica, los modelos de regresión lineales con variables categóricas y las técnicas de agrupamiento.

2.1. Modelos de Mixtura Finita

Según Frühwirth-Schnatter, Celeux y Robert (2018), los modelos de mixtura se representan como una densidad modelada mediante una combinación convexa de componentes, donde cada uno de estos componentes tienen una distribución paramétrica específica, y se define de la siguiente manera:

$$f(w) = \sum_{g=1}^{G} \eta_g f_g(w \mid \theta_g) \tag{2.1}$$

donde $w^\top = (w_1, ..., w_n)$, f_g son las funciones de densidad de las distribuciones paramétricas (componentes), θ_g son los parámetros de la distribución (componente g), y $\boldsymbol{\eta}^\top = (\eta_1, ..., \eta_G)$ son los pesos de los G componentes de la mixtura, cumpliéndose que $\eta_g \geq 0$ y $\sum_{g=1}^{G} \eta_g = 1$.

Cabe precisar que los modelos de mixtura se adoptan mayormente para el modelamiento de variables continuas, sin embargo también pueden usarse para variables discretas; por otro lado, los modelos de mixtura también se pueden trabajar con datos multivariados.

En la inferencia de un modelo de mixtura, las densidades de los componentes f_g son conocidas; y sus parámetros específicos θ_g y sus pesos η_g se consideran desconocidos. En algunos casos, el número de componentes G también se desconoce. Adicionalmente, los modelos de mixtura finita proporcionan un enfoque probabilístico que permite el agrupamiento, donde cada componente de la mixtura corresponde a un grupo de datos que pertenecen a una distribución paramétrica conocida.

La ecuación (2.1) es una representación genérica de los modelos de mixtura, de donde se pueden desprender los siguientes casos:

- Los modelos de mixtura pueden ser finitas e infinitas, siendo esta clasificación dependiente del número de componentes G; es decir, si G es igual a ∞, entonces:

$$f(w) = \sum_{g=1}^{\infty} \eta_g f_g(w \mid \theta_g) \tag{2.2}$$

  Este tipo de modelo evita varias dificultades relacionadas con la elección de G. Sin embargo, ya sea que G sea finito o infinito, en general existirán componentes vacíos, es decir, componentes donde no se tendrán observaciones.

- Si consideramos que tenemos un modelo de mixtura finita con densidades conocidas y fijas, se obtiene:

$$f(w) = \sum_{g=1}^{G} \eta_g f(w \mid \theta_g) \tag{2.3}$$

  donde cada componente de la mixtura es parte de una familia de distribuciones paramétricas conocidas, por ejemplo una mixtura finita de normales se representaría como: $f(\cdot \mid \theta) = f_N(\cdot \mid \mu, \sigma^2)$.

2.1.1. Distribución de Mixtura Finita de Normales

Una distribución de mixtura finita de normales se construye mediante el uso de una familia de distribuciones normales en la ecuación (2.3), obteniéndose:

$$f(w \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} \eta_g f_N(w \mid \mu_g, \sigma_g^2) \tag{2.4}$$

donde $\boldsymbol{\theta} = (\eta_1, ..., \eta_G, \mu_1, ..., \mu_G, \sigma_1^2, ..., \sigma_G^2)$ es el vector de parámetros de la distribución, y $f_N(\cdot \mid \mu_g, \sigma_g^2)$ es la función de densidad de una distribución normal con media μ_g y varianza σ_g^2.

2.1.2. Propiedades de la Distribución de Mixtura Finita de Normales

Teniendo como referencia la ecuación (2.4), la media de la distribución de mixtura finita de normales se calcula de la siguiente forma:

$$\begin{aligned}
\mu_w &= E[w \mid \boldsymbol{\theta}] \\
&= \int_{-\infty}^{\infty} w f(w \mid \boldsymbol{\theta}) dw \\
&= \int_{-\infty}^{\infty} w[\eta_1 f_N(w \mid \mu_1, \sigma_1^2) + ... + \eta_G f_N(w \mid \mu_G, \sigma_G^2)] dw \\
&= \int_{-\infty}^{\infty} w[\eta_1 f_N(w \mid \mu_1, \sigma_1^2)] dw + ... + \int_{-\infty}^{\infty} w[\eta_G f_N(w \mid \mu_G, \sigma_G^2)] dw \\
&= \eta_1 \int_{-\infty}^{\infty} w f_N(w \mid \mu_1, \sigma_1^2) dw + ... + \eta_G \int_{-\infty}^{\infty} w f_N(w \mid \mu_G, \sigma_G^2) dw \\
&= \eta_1 \mu_1 + \eta_2 \mu_2 + ... + \eta_G \mu_G \\
&= \sum_{g=1}^{G} \eta_g \mu_g
\end{aligned} \tag{2.5}$$

Asimismo, la varianza de la distribución de mixtura finita de normales se define de la siguiente forma:

$$\begin{aligned}
\sigma_w^2 &= E[w^2 \mid \boldsymbol{\theta}] - E[w \mid \boldsymbol{\theta}]^2 \\
&= \int_{-\infty}^{\infty} w^2 f(w \mid \boldsymbol{\theta}) dw - \mu_w^2 \\
&= \int_{-\infty}^{\infty} w^2 [\eta_1 f_N(w \mid \mu_1, \sigma_1^2) + ... + \eta_G f_N(w \mid \mu_G, \sigma_G^2)] dw - \mu_w^2 \\
&= \int_{-\infty}^{\infty} w^2 [\eta_1 f_N(w \mid \mu_1, \sigma_1^2)] dw + ... + \int_{-\infty}^{\infty} w^2 [\eta_G f_N(w \mid \mu_G, \sigma_G^2)] dw - \mu_w^2 \\
&= \eta_1 \int_{-\infty}^{\infty} w^2 f_N(w \mid \mu_1, \sigma_1^2) dw + ... + \eta_G \int_{-\infty}^{\infty} w^2 f_N(w \mid \mu_G, \sigma_G^2)) dw - \mu_w^2 \\
&= \eta_1(\sigma_1^2 + \mu_1^2) + ... + \eta_G(\sigma_G^2 + \mu_G^2) - \mu_w^2 \\
&= \sum_{g=1}^{G} \eta_g(\sigma_g^2 + \mu_g^2) - \mu_w^2
\end{aligned} \tag{2.6}$$

2.2. Distribución de Dirichlet

Según Blei, Ng y Jordan (2003), la distribución de Dirichlet es una generalización de la distribución beta[1] para múltiples variables aleatorias. Un vector aleatorio $\boldsymbol{\eta}^\top = (\eta_1, ..., \eta_K)$ con distribución de Dirichlet K-dimensional es un K-vector que se encuentra en el $(K-1)$-simplex, lo cual significa que $\eta_k \in \mathbb{R}$, tal que $\eta_k \geq 0$ y $\sum_{k=1}^{K} \eta_k = 1$. Dado que $\boldsymbol{\eta}$ es un vector cuyos valores son números reales entre $[0, 1]$, el dominio de la distribución de Dirichlet es en sí un conjunto de distribuciones de probabilidad, y que cuentan con las mismas propiedades que las distribuciones de probabilidad. Por esta razón, la distribución de Dirichlet se puede

[1] Sea X una variable aleatoria que sigue una distribución beta con parámetros α y β. La función de densidad de probabilidad de X es: $f(x) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, con media $\frac{\alpha}{\alpha+\beta}$ y varianza $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

considerar como una "distribución de distribuciones".

Adicionalmente, una variable aleatoria $\boldsymbol{\eta}$ sigue una distribución de Dirichlet K-dimensional con parámetro $\boldsymbol{\alpha}$ ($\boldsymbol{\eta} \sim Dir_K(\boldsymbol{\alpha})$), si su función de densidad de probabilidad en el $(K-1)$-simplex se expresa de la siguiente forma:

$$f(\boldsymbol{\eta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + ... + \alpha_K)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \eta_1^{\alpha_1 - 1} ... \eta_K^{\alpha_K - 1} \tag{2.7}$$

donde $\boldsymbol{\alpha}^\top = (\alpha_1, ..., \alpha_K)$ es un vector K-dimensional con componentes $\alpha_k > 0$, y $\Gamma(\cdot)$ es la función gamma. La ecuación (2.7) también se puede expresar como:

$$f(\boldsymbol{\eta} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \eta_k^{\alpha_k - 1} \tag{2.8}$$

donde $B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$ es la función beta multivariada.

Cabe precisar que, las distribuciones de Dirichlet son convenientes para la inferencia, ya que pertenecen a la familia exponencial, tienen estadísticos suficientes de dimensión finita y están conjugadas con la distribución multinomial. Asimismo, las distribuciones de Dirichlet se utilizan mayormente como distribución a priori de variables categóricas o de variables multinomiales en modelos de mixturas bayesianas u otros modelos bayesianos jerárquicos, siendo usado muy a menudo el muestreo de Gibbs para la inferencia.

En Ronning (1989), se muestran las siguientes expresiones para la media y la varianza de la distribución de Dirichlet:

$$\mu_{\eta_k} = E(\eta_k) = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} \tag{2.9}$$

$$\sigma^2_{\eta_k} = V(\eta_k) = \frac{\alpha_k(\sum_{k=1}^{K} \alpha_k - \alpha_k)}{(\sum_{k=1}^{K} \alpha_k)^2(\sum_{k=1}^{K} \alpha_k + 1)} \tag{2.10}$$

2.2.1. Distribución de Dirichlet Simétrica

La distribución de Dirichlet simétrica es un caso especial bastante usual, donde todos los elementos del vector que componen el parámetro $\boldsymbol{\alpha}$ tienen el mismo valor. Dado que todos los elementos del vector tienen el mismo valor, la distribución de Dirichlet simétrica se puede parametrizar mediante un único valor escalar:

$$f(\boldsymbol{\eta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \eta_1^{\alpha - 1} ... \eta_K^{\alpha - 1} = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \eta_k^{\alpha - 1} \tag{2.11}$$

Para cualquier componente del vector $\boldsymbol{\eta}$ que sigue una distribución de Dirichlet simétrica K-dimensional, se tiene como media y varianza a:

$$\mu_\eta = \quad E(\eta_k) \quad = \frac{1}{K} \tag{2.12}$$

$$\sigma_\eta^2 = \quad V(\eta_k) \quad = \frac{K-1}{K^2(K\alpha+1)} \tag{2.13}$$

donde $k = 1, ..., K$.

2.3. Modelos de Regresión Lineal

El análisis de regresión es una técnica estadística usada para identificar la relación entre una variable de respuesta $\mathbf{y}$ y una o más variables explicativas (covariables) $\boldsymbol{x_k}$. Esta relación entre las variables consiste en determinar las k covariables $(x_1, ..., x_k)$ que afectan una variable de respuesta $\mathbf{y}$. Por ello, es necesario construir un modelo que sirva para explicar el comportamiento de $\mathbf{y}$ en función de una combinación lineal de covariables $\boldsymbol{x_k}$, al cual se le conoce como "modelo de regresión lineal".

Para estimar un modelo de regresión lineal, se requiere de una muestra de n observaciones, y cada observación $i = 1, ..., n$ se representa de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i \tag{2.14}$$

donde $\boldsymbol{x_i} = (1, x_{i1}, ..., x_{ik})^\top$ es un vector de covariables, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^\top$ es un vector de coeficientes de regresión, β_0 es un intercepto, y ϵ_i es un término de error que sigue una distribución normal con media 0 y varianza σ^2 .

Otra forma de expresar el modelo de regresión es en la forma matricial.

$$\mathbf{y} = \mathbf{x}\beta + \epsilon \tag{2.15}$$

donde:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad ; \quad \mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad ; \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{2.16}$$

Aquí, $\mathbf{x}$ es una matriz $n \times (k+1)$, $\mathbf{y}$ es un vector de variables aleatorias observables de n componentes, $\boldsymbol{\beta}$ es un vector de parámetros desconocidos, y $\boldsymbol{\epsilon}$ es un vector de errores aleato-

rios no observables de n componentes, que usualmente se considera.

Luego, el modelo dado en la ecuación (2.15) se puede expresar como:

$$\mathbf{y} \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \tag{2.17}$$

2.3.1. Modelos de Regresión Lineal para Variables Categóricas

Una de las complicaciones de la estimación de los modelos regresión lineal es la existencia de covariables que son medidas en una escala nominal u ordinal, y que en muchas ocasiones estas covariables presentan una gran cantidad de categorías (por ello el nombre de variables categóricas), por ejemplo: estado civil, grupo sanguíneo, cargo laboral, nivel socio-económico, entre otros.

Para determinar el efecto que tiene una covariable categórica sobre un modelo de regresión lineal, es necesario definir una categoría (o nivel) como línea base y utilizar variables ficticias (usualmente binarias) para las otras categorías (o niveles), permitiendo la estimación de los efectos de dichas categorías teniendo como referencia a la línea base. Por lo tanto, el efecto de una covariable categórica con $c + 1$ categorías se captura dentro del modelo de regresión lineal mediante un conjunto de c coeficientes (o efectos) de regresión.

Un modelo de regresión lineal que contiene solamente covariables categóricas tiene la siguiente expresión:

$$y_i = \beta_0 + \sum_{j=1}^{J}\sum_{k=1}^{c_j} x_{ijk}\beta_{jk} + \epsilon_i \tag{2.18}$$

donde y_i son las respuestas continuas y J es el número de covariables categóricas, teniendo cada covariable j un total de $c_j + 1$ categorías. Asimismo, para cada covariable, el valor de 0 se define como la categoría de referencia y x_{ijk} denota la variable ficticia correspondiente a la k-ésima categoría de la covariable j, siendo $k = 1, .., c_j$. Por otro lado, $\epsilon_i \sim N(0, \sigma^2)$ es un término de error con distribución normal, β_0 es un intercepto y β_{jk} es un efecto de la k-ésima categoría de la covariable j con respecto a la categoría de referencia. Cabe mencionar que a β_{jk} se le conoce como el "efecto de nivel" de la categoría k de la covariable j.

2.4. Técnicas de Agrupamiento

Estas técnicas permiten realizar grupos de objetivos, analizando todas las posibilidades de agrupamiento entre dichos objetivos. En esta sección se desarrolla la técnica de agrupamiento PAM y los coeficientes de silhouette.

2.4.1. Algoritmo de Agrupamiento PAM (Partitioning Around Medoid)

Según Kaufman y Rousseeuw (1990), PAM es un algoritmo que asigna observaciones a los grupos, se selecciona como centro del grupo a uno de los puntos, y se minimiza la distancia

entre los objetos del grupo y su centro.

El algoritmo PAM consta de dos etapas: 1) la fase de construcción, que consiste en encontrar buenos medoides iniciales; y 2) la fase de intercambio, que busca mejorar el agrupamiento considerando otras observaciones como posibles medoides.

El objetivo del algoritmo PAM es minimizar la suma de diferencias entre las observaciones y sus medoides más cercanos, y la partición final es seleccionada al comparar las particiones con diferentes números de grupos mediante sus coeficientes de silhouette.

2.4.2. Coeficientes de Silhouette

Según Kaufman y Rousseeuw (1990), los coeficientes de silhouette se calculan como:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.19}$$

donde $s(i) \in [-1, 1]$, i es cualquier objeto en el conjunto de datos, A es el grupo al que se le ha asignado, y C es cualquier otro grupo diferente de A. Además, $a(i)$ es la disimilitud promedio de i con todos los demás objetos de A, y $d(i; C)$ es la disimilitud promedio de i con todos los objetos del grupo C; siendo $b(i) = min\{d(i, C)\}$, para todos los grupos $C \neq A$.

Asimismo, las medidas de disimilitud se pueden calcular utilizando cualquier métrica de distancia. Si el grupo A no contiene otros objetos además del objeto i, $s(i) = 0$. Si el valor de $s(i)$ es grande, indica que el objeto i está en promedio más cerca de los objetos que están en el mismo grupo y más lejos de los objetos que pertenecen al grupo C. En cambio, si $s(i)$ es pequeño, el objeto i está más cerca de los objetos de algún otro grupo que de los objetos de su propio grupo.

Cabe mencionar que, la desventaja de usar el coeficiente de silhouette es que no se puede calcular para un único grupo. En otras palabras, considerando el modelo de regresión lineal explicado en la sección 2.3.1, todas las categorías o variables ficticias de una covariable categórica no se podrán fusionar a la categoría de referencia.

Capítulo 3

Inferencia Bayesiana para la Fusión de Efectos

Los modelos de regresión lineal son muy usados en muchas aplicaciones en la actualidad, ya que se dispone de una gran cantidad y variedad de datos referentes a temas económicos, sociales, entre otros. Es por ello, que los modelos de regresión lineal cobran una mayor importancia, debido a que consiguen estimar una variable dependiente en función de una combinación lineal de covariables independientes.

La inclusión de variables categóricas (cualitativas) como covariables dentro de un modelo de regresión lineal es un gran desafío que se ha venido trabajando a lo largo de los años por distintos investigadores, desde el punto de vista frecuentista y bayesiano. El reto es aún mayor cuando se requiere la simplificación de estos modelos de regresión lineal que contienen dichas covariables categóricas, dado que la simplificación no solo se consigue con la selección de las categorías de las covariables relevantes, sino también con la fusión de los efectos de nivel de las covariables categóricas.

En ese sentido, en el presente capítulo se desarrolla el método propuesto por Malsiner-Walli, Pauger y Wagner (2018), que consiste en un enfoque bayesiano para la fusión de efectos de nivel de las covariables categóricas mediante la técnica de agrupamiento PAM. Este método define una mixtura finita de normales como distribución a priori para los coeficientes de regresión, con un esquema jerárquico, y una distribución de Dirichlet simétrica para los pesos de la mixtura.

Para un mejor entendimiento del método de Malsiner-Walli, Pauger y Wagner (2018), primero se trabajará con una adaptación del modelo de regresión lineal en una sola covariable, es decir, la ecuación (2.18) se reduce a:

$$y_i = \beta_0 + \sum_{k=1}^{c} x_{ik}\beta_k + \epsilon_i, \quad i = 1, ..., n \tag{3.1}$$

donde y_i son las respuestas continuas del modelo y la única covariable categórica del modelo cuenta con $c + 1$ categorías. Asimismo, para la covariable categórica, el valor de 0 se define para la categoría de referencia, y x_{ik} es la variable ficticia correspondiente para la k-ésima categoría de la covariable categórica, teniendo esta covariable en total c variables ficticias

$(k = 1, ..., c)$, y donde cada una de ellas sólo toman dos posibles valores: 0 y 1, siendo $x_{jk} = 1$ cuando se cumpla la k-ésima categoría de la covariable categórica y $x_{jk} = 0$ en caso contrario, además, cuando todas las variables ficticias sean igual a 0 es porque se cumple la categoría de referencia de la covariable categórica. Adicionalmente, el intercepto del modelo es β_0, y β_k es el efecto de la k-ésima categoría de la covariable con respecto a la categoría de referencia; por ello, se conoce a β_k como el "efecto de nivel" de la categoría k. Finalmente, el término de error sigue una distribución normal, $\epsilon_i \sim N(0, \sigma^2)$.

3.1. Función de Verosimilitud

Considerando las ecuaciones (2.17) y (3.1), se tiene que:

$$p(y_i \mid \boldsymbol{x_i}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{ \frac{-1}{2\sigma^2}(y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2 \right\} \tag{3.2}$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_c)^\top$ son todos los efectos de la regresión.

Entonces, la función de verosimilitud es dada por:

$$\begin{aligned} L(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} p(y_i \mid \boldsymbol{x_i}, \boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{ \frac{-1}{2\sigma^2}(y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{ \sum_{i=1}^{n} \frac{-1}{2\sigma^2}(y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{ \frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) \right\} \end{aligned} \tag{3.3}$$

3.2. Distribución a Priori

En la distribución a priori, propuesta en el método de Malsiner-Walli, Pauger y Wagner (2018), se define una distribución para el intercepto β_0 y otra distribución para los efectos de regresión β_k, por ello, es conveniente definir $\boldsymbol{\beta^*} = (\beta_1, ..., \beta_c)^\top$.

Con lo anterior, se define la estructura de la distribución a priori con la siguiente expresión:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\beta_0) p(\boldsymbol{\beta^*} \mid \boldsymbol{\xi}) p(\sigma^2) \tag{3.4}$$

donde $\boldsymbol{\xi}$ son los hiperparámetros adicionales para los coeficientes de regresión, que se describirán posteriormente.

En base a la estructura de la distribución a priori, se define que el intercepto β_0 sigue una distribución normal y la varianza del término del error σ^2 sigue una distribución gamma

inversa, de la siguiente manera:

$$\beta_0 \sim N(0, \psi_0) \tag{3.5}$$
$$\sigma^2 \sim G^{-1}(s_0, S_0) \tag{3.6}$$

donde es usual utilizar los valores de $s_0 = 0$ y $S_0 = 0$ en la distribución gamma inversa, lo cual conlleva a una distribución impropia para σ^2.

Además, con lo explicado en la sección 2.1.1 y teniendo como referencia la adaptación para un modelo de regresión lineal con una única covariable categórica, se define la distribución a priori jerárquica de un efecto de nivel β_k descrita en Malsiner-Walli, Pauger y Wagner (2018) como:

$$p(\beta_k) = \sum_{\ell=0}^{L} \eta_\ell f_N(\beta_k \mid \mu_\ell, \psi) \tag{3.7}$$
$$\boldsymbol{\eta} \sim Dir_{L+1}(e) \tag{3.8}$$
$$\mu_0 = 0 \tag{3.9}$$
$$\mu_\ell \sim N(m_0, M_0), \ell = 1, ..., L \tag{3.10}$$

donde la mixtura contiene $L + 1$ componentes: $\ell = 0, ..., L$; y cada componente se define como una distribución normal f_N para modelar el efecto de nivel β_k con los parámetros de localización μ_ℓ y de escala constante ψ. Siendo $\mu_0 = 0$ el primer componente del parámetro de locación, y $\boldsymbol{\mu} = (\mu_1, ..., \mu_L)$ una priori que se define con una distribución normal con parámetros de locación m_0 y de escala M_0. Además, ψ también puede ser un parámetro de escala aleatorio mediante una priori que se define como una gamma inversa: $\psi \sim G^{-1}(g_0, G_0)$. Asimismo, $\boldsymbol{\eta}^\top = (\eta_0, ..., \eta_L)$ son los pesos de los componentes de la mixtura, y es una priori que se asume que sigue una distribución de Dirichlet simétrica con parámetro e, siendo $e > 0$ y cumpliéndose que $\eta_\ell \geq 0$ y $\sum_{\ell=0}^{L} \eta_\ell = 1$.

Según Malsiner-Walli, Pauger y Wagner (2018), la distribución a priori jerárquica definida, permite que los efectos de nivel se fusionen en un mismo componente solo si los efectos son idénticos. Además, los autores mencionan que la especificación de los hiperparámetros se basa en un enfoque empírico y se eligen en función de los datos, en base a una estimación del modelo realizada previamente, considerando una distribución a priori débilmente informativa. A continuación. se explica con mayor detalle la distribución a priori jerárquica:

- Para la covariable en análisis, el primer componente del parámetro de localización μ_0 se fija en 0 para identificar las categorías que tengan el mismo efecto que la categoría de referencia; por ello, si todos los efectos de nivel de una covariable se asignan al primer componente, la covariable es excluida completamente del modelo. Los componentes restantes del parámetro de localización $\boldsymbol{\mu} = (\mu_1, ..., \mu_L)$, conocidos como medias de los componentes de la mixtura, se asume que son condicionalmente independientes y que siguen una priori normal con parámetros de localización y escala: m_0 y M_0,

respectivamente; estableciéndose la media y la varianza como:

$$m_0 = media(\hat{\beta}^*) = \bar{\beta}^* \tag{3.11}$$

$$M_0 = (max_k \hat{\beta}_k^* - min_k \hat{\beta}_k^*)^2 \tag{3.12}$$

donde $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_1^*, ..., \hat{\beta}_c^*)^\top$ es el vector de coeficientes estimados previamente de la covariable en análisis bajo una priori débilmente informativa, y $\bar{\beta}^* = \frac{1}{c}\sum_{k=1}^{c} \hat{\beta}_k^*$. Cabe mencionar que, los autores sugieren que los valores iniciales sean estimados mediante un proceso previo del modelo considerando una priori poca informativa, como por ejemplo: β_k con una distribución normal de media igual a 0 y varianza grande, y σ^2 con una distribución gamma inversa de parámetros iguales a 0 (distribución impropia).

- Para la covariable en análisis, el parámetro de escala ψ, conocido como varianza del componente, es el mismo para todos los componentes de la mixtura con el fin de garantizar que al fusionar los efectos, se mantenga la misma dispersión. Por ello, ψ es constante y se define como la varianza de los efectos de nivel estimados $\hat{\boldsymbol{\beta}}^*$ de la covariable en análisis bajo la priori entre una constante v:

$$\psi = \left(\tfrac{1}{v}\right)\left(\tfrac{1}{c-1}\right)\sum_{k=1}^{c}(\hat{\beta}_k^* - \bar{\beta}^*)^2 = \left(\frac{V}{v}\right) \tag{3.13}$$

  Cabe mencionar que las formas de los componentes de la mixtura varían conforme cambia la constante v, es decir, las formas serán más "puntiaguda" si se aumenta v. En ese sentido, se puede afirmar que ψ controla implícitamente la precisión de la partición de los efectos de nivel, dicho de otra forma: la reducción del modelo de regresión lineal depende de ψ.
  Por otro lado, para obtener clústers o agrupamientos más robustos, se requiere que ψ sea aleatorio con la siguiente priori:

$$\psi \sim G^{-1}(g_0, G_0) \tag{3.14}$$

  Dado que $E(\psi) = \frac{G_0}{g_0 - 1}$, y teniendo en consideración que $E(\psi) \approx \frac{V}{v}$, se establece que $G_0 = \left(\frac{V}{v}\right)(g_0 - 1)$. Asimismo, la varianza se define como $V(\psi) = \frac{E(\psi)^2}{g_0 - 2}$. Por ende, para obtener desviaciones pequeñas en los componentes, se requieren que g_0 tome valores grandes.

- Para el número de componentes de la mixtura $L + 1$, primero se establece que todos los efectos son diferentes entre sí, es decir $L = c$. Por ende, la priori define un modelo de mixtura sobreajustada, donde la distribución de la mixtura tiene más componentes que los efectos de nivel a estimar. Adicionalmente, los pesos de los componentes de la mixtura $\boldsymbol{\eta}^\top = (\eta_0, ..., \eta_L)$ se asumen que siguen una distribución de Dirichlet simétrica $Dir_{L+1}(e)$, con parámetro e, siendo $e > 0$ y es quien controla concentración de observaciones dentro de los componentes de la mixtura. Por ello, cuando $e < 1$, las observaciones se concentran en sólo unos pocos componentes grandes, resultando muchos componentes pequeños y quedando algunos vacíos; y si se define $e \ll 1$, se está

induciendo a una mayor concentración de los componentes en unos pocos grupos.

Según lo explicado, es necesario simplificar la forma de la distribución a posteriori, por ello, se introducen variables latentes de asignación $\mathbf{S} = (S_1, ..., S_c)$ al método planteado por Malsiner-Walli, Pauger y Wagner (2018), las cuales tienen como finalidad indicar el componente de la mixtura al cual es asignado un efecto de regresión β_k, donde S_k toma valores en $0, 1, ..., L$. En ese sentido, considerando la ecuación (3.7), la distribución a priori para los efectos de nivel β_k condicionada por S_k se reduce a la siguiente expresión jerárquica:

$$\begin{aligned} \beta_k \mid S_k &\sim N(\mu_{S_k}, \psi) && (3.15)\\ S_k &\sim Cat(\boldsymbol{\eta}), \boldsymbol{\eta}^\top = (\eta_0, ..., \eta_L), S_k \in \{0, 1, ..., L\} && (3.16) \end{aligned}$$

donde S_k sigue una distribución categórica $Cat(\boldsymbol{\eta})$ con parámetro $\boldsymbol{\eta}^\top = (\eta_0, ..., \eta_L)$, siendo su función de probabilidad:

$$f(S_k) = \eta_0^{I(S_k=0)} \eta_1^{I(S_k=1)} ... \eta_L^{I(S_k=L)} = \prod_{\ell=0}^{L} \eta_l^{I(S_k=\ell)} \tag{3.17}$$

Por lo tanto, si $S_k = \ell$, la ecuación (3.15) se reduce a:

$$\beta_k \mid S_k = \ell \sim N(\mu_\ell, \psi) \tag{3.18}$$

Adicionalmente, si se define $\boldsymbol{b_0(S)} = (0, \mu_{S_1}, ..., \mu_{S_L})$ como el vector de medias de tamaño $(L+1) \times 1$ condicionada a sus indicadores S_k, y $\boldsymbol{B_0} = (\psi_0, \psi, ..., \psi)$ como la matriz diagonal de covarianza del vector de todos los efectos de regresión de tamaño $(L+1) \times (L+1)$, se obtiene la siguiente expresión:

$$\boldsymbol{\beta} \mid \mathbf{S} \sim N(\boldsymbol{b_0(S)}, \boldsymbol{B_0}) \tag{3.19}$$

3.3. Distribución a Posteriori

La distribución a posteriori es proporcional a la multiplicación de la función de verosimilitud (sección 3.1) por la distribución a priori (sección 3.2) y se define como:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto L(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto L(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) && (3.20) \end{aligned}$$

donde $\boldsymbol{\theta}$ son todos los parámetros $(\boldsymbol{\beta}, \sigma^2, \mathbf{S}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi)$. Asimismo se puede observar que la última expresión de la ecuación (3.20) coincide con la ecuación (3.4), y reemplazando se obtiene:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto L(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) p(\beta_0) p(\boldsymbol{\beta}^* \mid \boldsymbol{\xi}) p(\sigma^2) \\ p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto L(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) p(\beta_0) p(\boldsymbol{\beta}^* \mid \mathbf{S}, \boldsymbol{\mu}, \psi) p(\mathbf{S} \mid \boldsymbol{\eta}) p(\boldsymbol{\mu}) p(\psi) p(\boldsymbol{\eta}) p(\sigma^2) && (3.21) \end{aligned}$$

Entonces, considerando la ecuación (3.19), se obtiene:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \quad \propto \quad L(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{b_0(S)}, \boldsymbol{B_0}) p(\mathbf{S} \mid \boldsymbol{\eta}) p(\boldsymbol{\mu}) p(\psi) p(\boldsymbol{\eta}) p(\sigma^2) \qquad (3.22)$$

Cabe mencionar que para realizar la inferencia sobre la distribución a posteriori se utilizará el algoritmo de Gibbs, con el cual se simulará sobre la distribución a posteriori, permitiendo posteriormente la evaluación del modelo y la estimación del modelo promedio. A continuación se presentan las distribuciones condicionales completas, con las cuales se realizará la simulación para la implementación del algoritmo de Gibbs. Las distribuciones condicionales completas se presentarán en dos etapas: 1) los parámetros asociados a la regresión y 2) los parámetros asociados al agrupamiento.

1. **Etapa de regresión:** aquí se desarrollarán las distribuciones condicionales de los efectos de nivel ($\boldsymbol{\beta}$) y la varianza del error (σ^2), quienes se encuentran condicionadas a las componentes de la mixtura a las cuales se han asignado los efectos.

   a) Muestreo de los coeficientes de regresión $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_c)^\top$ condicionados a la variable latente de asignación $\mathbf{S}$:
   Para encontrar la distribución condicional de $\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi, \mathbf{y}$, se inicia con la simplificación de la ecuación:

$$p(\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi, \mathbf{y}) \quad \propto \quad L(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{b_0(S)}, \boldsymbol{B_0}) \qquad (3.23)$$

   Con lo cual, se obtiene:

$$\begin{aligned}
p(\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi, \mathbf{y}) \quad &\propto \quad \left(\frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{ \frac{-1}{2\sigma^2} (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) \right\} \right) \\
&\quad \left(\frac{1}{((2\pi)^{c+1} |\boldsymbol{B_0}|)^{1/2}} exp\left\{ \frac{-1}{2} (\boldsymbol{\beta} - \boldsymbol{b_0(S)})^\top \boldsymbol{B_0}^{-1} (\boldsymbol{\beta} - \boldsymbol{b_0(S)}) \right\} \right) \\
&\propto \quad \frac{1}{(2\pi\sigma^2)^{n/2} ((2\pi)^{c+1} |\boldsymbol{B_0}|)^{1/2}} \\
&\quad exp\left\{ \frac{-1}{2\sigma^2} (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) - \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{b_0(S)})^\top \boldsymbol{B_0}^{-1} (\boldsymbol{\beta} - \boldsymbol{b_0(S)}) \right\} \\
&\propto \quad exp\left\{ -\frac{1}{2\sigma^2} \left(\mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{x}^\top \mathbf{x}\boldsymbol{\beta} \right) \right. \\
&\quad \left. -\frac{1}{2} \left(\boldsymbol{\beta}^\top \boldsymbol{B_0}^{-1} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \boldsymbol{B_0}^{-1} \boldsymbol{b_0(S)} + \boldsymbol{b_0(S)}^\top \boldsymbol{B_0}^{-1} \boldsymbol{b_0(S)} \right) \right\} \\
&\propto \quad exp\left\{ \frac{\mathbf{y}^\top \mathbf{x}\boldsymbol{\beta}}{\sigma^2} - \frac{\boldsymbol{\beta}^\top \mathbf{x}^\top \mathbf{x}\boldsymbol{\beta}}{2\sigma^2} - \frac{\boldsymbol{\beta}^\top \boldsymbol{B_0}^{-1} \boldsymbol{\beta}}{2} + \boldsymbol{\beta}^\top \boldsymbol{B_0}^{-1} \boldsymbol{b_0(S)} \right\} \\
&\propto \quad exp\left\{ -\frac{1}{2} \boldsymbol{\beta}^\top \left(\frac{\mathbf{x}^\top \mathbf{x}}{\sigma^2} + \boldsymbol{B_0}^{-1} \right) \boldsymbol{\beta} + \boldsymbol{\beta}^\top \left(\frac{\mathbf{x}^\top \mathbf{y}}{\sigma^2} + \boldsymbol{B_0}^{-1} \boldsymbol{b_0(S)} \right) \right\}
\end{aligned}$$

Ahora, considerando:

$$\boldsymbol{B_N} = \sigma^2(\mathbf{x}^\top\mathbf{x} + \sigma^2\boldsymbol{B_0}^{-1})^{-1} \tag{3.24}$$

$$\boldsymbol{b_N} = \boldsymbol{B_N}\left(\frac{\mathbf{x}^\top\mathbf{y}}{\sigma^2} + \boldsymbol{B_0}^{-1}\boldsymbol{b_0(S)}\right) \tag{3.25}$$

La expresión se reduce a:

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi, \mathbf{y}) &\propto exp\Big\{ -\frac{1}{2}\boldsymbol{\beta}^\top\left(\boldsymbol{B_N}^{-1}\right)\boldsymbol{\beta} + \boldsymbol{\beta}^\top\left(\boldsymbol{B_N}^{-1}\boldsymbol{b_N}\right)\Big\} \\ &\propto exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta}^\top\boldsymbol{B_N}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{B_N}^{-1}\boldsymbol{b_N})\Big\} \\ &\propto exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta}^\top\boldsymbol{B_N}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}\boldsymbol{B_N}^{-1}\boldsymbol{b_N} + \boldsymbol{b_N}^\top\boldsymbol{B_N}^{-1}\boldsymbol{b_N})\Big\} \\ &\propto exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{b_N})^\top\boldsymbol{B_N}^{-1}(\boldsymbol{\beta} - \boldsymbol{b_N})\Big\} \end{aligned} \tag{3.26}$$

Resultando finalmente la siguiente distribución condicional para $\boldsymbol{\beta} \mid \mathbf{S}$:

$$\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi, \mathbf{y} \sim N(\boldsymbol{b_N}, \boldsymbol{B_N}) \tag{3.27}$$

b) Muestreo de la varianza del error σ^2:
Para encontrar la distribución condicional de σ^2, se inicia con la simplificación de la ecuación:

$$p(\sigma^2 \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi) \propto L(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2)p(\sigma^2) \tag{3.28}$$

Con lo cual, se obtiene:

$$\begin{aligned} p(\sigma^2 \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi) &\propto \left(\frac{1}{(2\pi\sigma^2)^{n/2}}exp\Big\{\frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})\Big\}\right) \\ &\quad \left(\frac{S_0^{s_0}}{\Gamma(s_0)}\left(\frac{1}{\sigma^2}\right)^{s_0+1}exp\Big\{\frac{-S_0}{\sigma^2}\Big\}\right) \\ &\propto \frac{S_0^{s_0}}{(2\pi\sigma^2)^{n/2}\Gamma(s_0)(\sigma^2)^{s_0+1}}exp\Big\{\frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) - \frac{S_0}{\sigma^2}\Big\} \\ &\propto \frac{1}{(\sigma^2)^{n/2+s_0+1}}exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) - \frac{S_0}{\sigma^2}\Big\} \\ &\propto \left(\frac{1}{\sigma^2}\right)^{(\frac{n}{2}+s_0)+1}exp\Big\{\frac{-1}{\sigma^2}\left(\frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) + S_0\right)\Big\} \end{aligned}$$

Ahora, considerando:

$$s_N = s_0 + \frac{n}{2} \tag{3.29}$$

$$S_N = S_0 + \frac{1}{2}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) \tag{3.30}$$

La expresión se reduce a:

$$p(\sigma^2 \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi) \quad \propto \quad \left(\frac{1}{\sigma^2}\right)^{s_N+1} exp\left\{\frac{-S_N}{\sigma^2}\right\} \tag{3.31}$$

Resultando finalmente la siguiente distribución condicional para σ^2:

$$\sigma^2 \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \psi \sim G^{-1}(s_N, S_N) \tag{3.32}$$

2. **Etapa de agrupamiento:** aquí se encontrarán las distribuciones condicionales de los parámetros de la mixtura ($\boldsymbol{\eta}$, $\boldsymbol{\mu}$, ψ) y las variables latentes de asignación ($\mathbf{S}$).

   a) Muestreo de los pesos de los componentes de la mixtura $\boldsymbol{\eta}$:
   Para encontrar la distribución condicional de $\boldsymbol{\eta}$, se inicia con la simplificación de la ecuación:

$$p(\boldsymbol{\eta} \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\mu}, \psi) \quad \propto \quad p(\mathbf{S} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \tag{3.33}$$

   Con lo cual, se obtiene:

$$\begin{aligned} p(\boldsymbol{\eta} \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\mu}, \psi) &\propto \left(\prod_{k=1}^{c} \left(\eta_0^{I(S_k=0)} ... \eta_L^{I(S_k=L)}\right)\right) \left(\frac{\Gamma(Le)}{(\Gamma(e)^L} \eta_0^{e-1} ... \eta_L^{e-1}\right) \\ &\propto \left(\eta_0^{\sum_{k=1}^{c} I(S_k=0)} ... \eta_L^{\sum_{k=1}^{c} I(S_k=L)}\right) \left(\eta_0^{e-1} ... \eta_L^{e-1}\right) \\ &\propto \left(\eta_0^{e+\sum_{k=1}^{c} I(S_k=0)-1} ... \eta_L^{e+\sum_{k=1}^{c} I(S_k=L)-1}\right) \end{aligned}$$

   Ahora, considerando:

$$e_\ell \quad = \quad e + N_\ell, \ell = 0, ..., L \tag{3.34}$$

$$N_\ell \quad = \quad \sum_{k=1}^{c} I(S_k = \ell) \tag{3.35}$$

   donde N_ℓ es el número de coeficientes de regresión β_k asignados al componente de la mixtura ℓ. Entonces, la expresión se reduce a:

$$p(\boldsymbol{\eta} \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\mu}, \psi) \quad \propto \quad \left(\eta_0^{(e+N_0)-1} ... \eta_L^{(e+N_L)-1}\right) \tag{3.36}$$

   Resultando finalmente la siguiente distribución condicional para $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\mu}, \psi \sim Dir_{L+1}(e_0, e_1, ..., e_L) \tag{3.37}$$

   b) Muestreo de las medias de los componentes de la mixtura $\boldsymbol{\mu}$:
   Para encontrar la distribución condicional de $\boldsymbol{\mu}$, se inicia con la simplificación de la ecuación:

$$p(\boldsymbol{\mu} \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \psi) \quad \propto \quad p(\boldsymbol{\beta^*} \mid \mathbf{S}, \boldsymbol{\mu}, \psi) p(\boldsymbol{\mu}) \tag{3.38}$$

Teniendo en cuenta que, se realizará el muestreo para cada μ_ℓ, siendo $\ell = 1, ...L$, se obtiene:

$$\begin{aligned} p(\mu_\ell \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \psi) &\propto \left(\prod_{\substack{k=1 \\ S_k=\ell}}^{c} p(\beta_k \mid S_k, \mu_{S_k}, \psi) \right) p(\mu_\ell) \\ &\propto \left(\prod_{\substack{k=1 \\ S_k=\ell}}^{c} \left(\frac{1}{(2\pi\psi)^{1/2}} exp\left\{ \frac{-1}{2\psi} (\beta_k - \mu_{S_k})^2 \right\} \right) \right) \\ &\quad \left(\frac{1}{(2\pi M_0)^{1/2}} exp\left\{ \frac{-1}{2M_0} (\mu_\ell - m_0)^2 \right\} \right) \end{aligned}$$

De la expresión anterior, sólo se consideran los componentes del producto donde $S_k = \ell$, y el resto de componentes donde $(S_k \neq \ell)$ se considerarán como una constante. Entonces:

$$\begin{aligned} p(\mu_\ell \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \psi) &\propto \left(\frac{1}{(2\pi\psi)^{c/2}} exp\left\{ \sum_{\substack{k=1 \\ S_k=l}}^{c} \left(\frac{-1}{2\psi} (\beta_k - \mu_\ell^2) \right) \right\} \right) \\ &\quad \left(\frac{1}{(2\pi M_0)^{1/2}} exp\left\{ \frac{-1}{2M_0} (\mu_\ell - m_0)^2 \right\} \right) \\ &\propto \frac{1}{(2\pi\psi)^{c/2}(2\pi M_0)^{1/2}} exp\left\{ \frac{-1}{2\psi} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} (\beta_k - \mu_\ell)^2 - \frac{1}{2M_0} (\mu_\ell - m_0)^2 \right\} \\ &\propto exp\left\{ -\frac{1}{2\psi} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} (\beta_k^2 - 2\beta_k\mu_\ell + \mu_\ell^2) - \frac{1}{2M_0} (\mu_\ell^2 - 2\mu_\ell m_0 + m_0^2) \right\} \\ &\propto exp\left\{ \frac{1}{\psi} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} \beta_k\mu_\ell - \frac{1}{2\psi} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} \mu_\ell^2 - \frac{\mu_\ell^2}{2M_0} + \frac{\mu_\ell m_0}{M_0} \right\} \end{aligned}$$

Además, tomando como:

$$\bar{\beta}_\ell = \frac{1}{N_\ell} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} \beta_k \tag{3.39}$$

que representa a la media de los elementos de β_k asignados al componente de la mixtura ℓ, y N_ℓ es el número de coeficientes de regresión β_k asignados al componente de la mixtura ℓ; se obtiene la siguiente expresión:

$$\begin{aligned} p(\mu_\ell \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \psi) &\propto exp\left\{ \frac{\mu_\ell N_\ell \bar{\beta}_\ell}{\psi} - \frac{N_\ell \mu_\ell^2}{2\psi} - \frac{\mu_\ell^2}{2M_0} + \frac{\mu_\ell m_0}{M_0} \right\} \\ &\propto exp\left\{ -\frac{1}{2}\mu_\ell^2 \left(\frac{N_\ell}{\psi} + \frac{1}{M_0} \right) + \mu_\ell \left(\frac{N_\ell \bar{\beta}_\ell}{\psi} + \frac{m_0}{M_0} \right) \right\} \end{aligned}$$

Ahora, considerando:

$$M_\ell = (N_\ell/\psi + 1/M_0)^{-1} \tag{3.40}$$
$$m_\ell = M_\ell(N_\ell\bar{\beta}_\ell/\psi + m_0/M_0) \tag{3.41}$$

La expresión se reduce a:

$$\begin{aligned} p(\mu_\ell \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \psi) &\propto exp\left\{ -\frac{1}{2}\mu_\ell^2\left(\frac{1}{M_\ell}\right) + \mu_\ell\left(\frac{m_\ell}{M_\ell}\right)\right\} \\ &\propto exp\left\{ -\frac{1}{2M_\ell}(\mu_\ell^2 - 2m_\ell\mu_\ell)\right\} \\ &\propto exp\left\{ -\frac{1}{2M_\ell}(\mu_\ell^2 - 2m_\ell\mu_\ell + m_\ell^2)\right\} \\ &\propto exp\left\{ -\frac{1}{2M_\ell}(\mu_\ell - m_\ell)^2\right\} \end{aligned} \tag{3.42}$$

Resultando finalmente la siguiente distribución condicional para μ_ℓ:

$$\mu_\ell \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \psi \sim N(m_\ell, M_\ell) \tag{3.43}$$

c) Muestreo de las varianzas de los componentes de la mixtura ψ (es necesario indicar que este paso sólo se desarrolla si se especifica una priori en ψ, caso contrario, se omite este paso):
Para encontrar la distribución condicional de ψ, se inicia con la simplificación de la ecuación:

$$p(\psi \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}) \propto p(\boldsymbol{\beta^*} \mid \mathbf{S}, \boldsymbol{\mu}, \psi)p(\psi) \tag{3.44}$$

Con lo cual, se obtiene:

$$\begin{aligned} p(\psi \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}) &\propto \left(\prod_{\substack{k=1 \\ S_k=\ell}}^{c} p(\beta_k \mid S_k, \mu_{S_k}, \psi)\right) p(\psi) \\ &\propto \left(\prod_{\substack{k=1 \\ S_k=\ell}}^{c} \left(\frac{1}{(2\pi\psi)^{1/2}} exp\left\{\frac{-1}{2\psi}(\beta_k - \mu_{S_k})^2\right\}\right)\right) \\ &\quad \left(\frac{G_0^{g_0}}{\Gamma(g_0)}\left(\frac{1}{\psi}\right)^{g_0+1} exp\left\{\frac{-G_0}{\psi}\right\}\right) \end{aligned}$$

De la expresión anterior, sólo se consideran los componentes del producto donde $S_k = \ell$, y el resto de componentes donde $(S_k \neq \ell)$ se considerarán como una

constante. Entonces:

$$\begin{aligned} p(\psi \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}) &\propto \left(\frac{1}{(2\pi\psi)^{c/2}} exp\left\{ \sum_{\substack{k=1 \\ S_k=\ell}}^{c} \left(\frac{-1}{2\psi}(\beta_k - \mu_\ell)^2 \right) \right\} \right) \\ &\quad \left(\frac{G_0^{g_0}}{\Gamma(g_0)} \left(\frac{1}{\psi} \right)^{g_0+1} exp\left\{ \frac{-G_0}{\psi} \right\} \right) \\ &\propto \frac{G_0^{g_0}}{(2\pi\psi)^{c/2}\Gamma(g_0)(\psi)^{g_0+1}} exp\left\{ \frac{-1}{2\psi} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} (\beta_k - \mu_\ell)^2 - \frac{G_0}{\psi} \right\} \\ &\propto \frac{1}{(\psi)^{c/2+g_0+1}} exp\left\{ -\frac{1}{2\psi} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} (\beta_k - \mu_\ell)^2 - \frac{G_0}{\psi} \right\} \\ &\propto \left(\frac{1}{\psi} \right)^{(c/2+g_0)+1} exp\left\{ \frac{-1}{\psi} \left(\frac{1}{2} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} (\beta_k - \mu_\ell)^2 + G_0 \right) \right\} \end{aligned}$$

Ahora, considerando:

$$g_N = \frac{c}{2} + g_0 \tag{3.45}$$

$$G_N = \frac{1}{2} \sum_{\substack{k=1 \\ S_k=\ell}}^{c} (\beta_k - \mu_\ell)^2 + G_0 \tag{3.46}$$

La expresión se reduce a:

$$p(\psi \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu}) \propto \left(\frac{1}{\psi} \right)^{g_N+1} exp\left\{ \frac{-G_N}{\psi} \right\} \tag{3.47}$$

Resultando finalmente la siguiente distribución condicional para ψ:

$$\psi \mid \mathbf{y}, \mathbf{S}, \boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}, \boldsymbol{\mu} \sim G^{-1}(g_N, G_N) \tag{3.48}$$

d) Muestreo de las variables latentes de asignación **S**:
Para encontrar la distribución condicional de **S**, se inicia con la simplificación de la ecuación:

$$p(\mathbf{S} \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \psi, \mathbf{y}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}) \propto p(\boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{\mu}, \psi) p(\mathbf{S} \mid \boldsymbol{\eta}) \tag{3.49}$$

Teniendo en cuenta sólo una variable de asignación S_h, la expresión se reduce a:

$$\begin{aligned} p(S_h \mid \beta_h, \boldsymbol{\mu}, \psi, \mathbf{y}, \boldsymbol{\sigma^2}) &\propto p(\beta_h \mid S_h, \mu_{S_h}, \psi) p(S_h \mid \boldsymbol{\eta}) \qquad (3.50) \\ &\propto f_N(\beta_h \mid \mu_{S_h}, \psi) \left(\prod_{l=0}^{L} \eta_l^{I(S_h=l)} \right) \\ &\propto f_N(\beta_h \mid \mu_{S_h}, \psi) \left(\eta_0^{I(S_h=0)} ... \eta_l^{I(S_h=l)} ... \eta_L^{I(S_h=L)} \right) \end{aligned}$$

Ahora, considerando $S_h = l$, se obtiene que la probabilidad para cada variable latente es proporcional a:

$$P(S_h = l \mid \beta_h, \boldsymbol{\mu}, \psi, \mathbf{y}, \boldsymbol{\sigma^2}) \quad \propto \quad f_N(\beta_h \mid \mu_l, \psi)(\eta_l) \tag{3.51}$$

donde $h = 1, ..., L$. Resultando finalmente que la distribución condicional para S_h es una distribución categórica.

3.4. Selección del Modelo

En esta sección, se desarrollan conceptos relacionados a la estimación, selección y comparación de los modelos de regresión.

3.4.1. Estimaciones del Modelo Promedio

El algoritmo de Gibbs realiza una aproximación de la distribución a posteriori completa de los efectos de regresión β_k, teniendo en cuenta la incertidumbre del modelo. Es decir, para la estimación del modelo promedio de un efecto de regresión β_k, la distribución a posteriori se representa mediante la siguiente mixtura:

$$p(\beta_k \mid \mathbf{y}) = \sum_m p(\beta_k \mid \mathbf{y}, \mathcal{M}^{(m)}) p(\mathcal{M}^{(m)} \mid \mathbf{y}) \tag{3.52}$$

donde m son las iteraciones del algoritmo de Gibbs, $p(\beta_k \mid \mathbf{y}, \mathcal{M}^{(m)})$ son las distribuciones a posteriori del modelo específico o componentes de la mixtura, y $p(\mathcal{M}^{(m)} \mid \mathbf{y})$ son las probabilidades del modelo a posteriori o pesos de las mixtura. Malsiner-Walli, Pauger y Wagner (2018) indica que la media de todas las salidas del algoritmo de Gibbs para β_k debería ser un estimador robusto del modelo promedio.

3.4.2. Métodos para la Selección del Modelo

Durante el desarrollo del algoritmo de Gibbs se pueden visualizar cambios en las etiquetas asociadas con los componentes de la mixtura, por ello, es usual realizar la estimación del modelo después de las salidas del algoritmo de Gibbs, con el fin de obtener un etiquetado único.

La selección del modelo se basa en la información de la agrupación de un par de efectos de nivel β_g y β_h, es decir, si β_g y β_h se asignan al mismo grupo o a diferentes grupos. Para cada iteración m del algoritmo de Gibbs y considerando una covariable, se construye una matriz $M^{(m)}$ de orden $(L+1) \times (L+1)$; con una función indicadora, con entrada de 1, si los dos niveles g y h pertenecen al mismo grupo, y de 0 en caso contrario, es decir:

$$M_{gh}^{(m)} = I(S_g^{(m)}, S_h^{(m)}) \tag{3.53}$$

Esta matriz $M^{(m)}$ es independiente del etiquetado de los componentes y por ende, invariable al cambio de etiquetas. La matriz contiene la información de la agrupación para la covariable,

es decir, toda la información con respecto al número de grupos de efectos y elementos dentro de los grupos.

Después de la aplicación del algoritmo de Gibbs, el agrupamiento de la distribución a posteriori y la selección de la partición final de los efectos de nivel β_k, puede darse mediante:

1. Seleccionar la partición final mediante la mayor probabilidad de la matriz $M^{(m)}$, considerando todas las iteraciones N_m del modelo muestreado durante el algoritmo de Gibbs.
   Dado que el parámetro e de la distribución de Dirichlet se especifica muy pequeño, entonces, los grupos "verdaderos" no deberán dividirse, la distribución a posteriori se concentrará en particiones parsimoniosas de los efectos y el número de agrupaciones dependerá únicamente del tamaño de la varianza especificada.

2. Seleccionar la partición final mediante el promedio de la matriz $M^{(m)}$, considerando todas las iteraciones N_m del modelo muestreado durante el algoritmo de Gibbs, construyéndose la matriz C para una covariable:

$$C = \frac{1}{N_m} \sum_{m=1}^{N_m} M^{(m)} \tag{3.54}$$

   donde la matriz C contiene entradas C_{gh} que corresponden a la frecuencia relativa con la que los efectos de dos niveles β_g y β_h se asignan a un mismo grupo, y con ello, se aproxima a la probabilidad a posteriori de que β_g y β_h sean miembros de un mismo grupo. Por tanto, la matriz C puede interpretarse como una matriz de "similitud", es decir, un valor de C_{gh} cercano a 1 indica que los efectos de dos niveles son casi idénticos y deberían agruparse.
   Para el agrupamiento de los efectos de nivel considerando la matriz de similitud, se puede utilizar el algoritmo PAM (Partitioning Around Medoid), que usa la técnica de agrupamiento por k-medoides. Esta técnica se adapta fácilmente al modelo descrito, ya que la matriz de similitud se puede transformar fácilmente en una matriz de distancia considerando: $D = 1 - C$, donde 1 es una matriz con elementos 1.

3.4.3. Criterios de Comparación para Selección del Modelo

Spiegelhalter, Best, Carlin y Van der Linde (2002) definen el criterio de información del desvío (DIC) como una medida bayesiana de comparación de modelos, que combina el ajuste y la complejidad de un modelo.

$$DIC = 2\overline{D(\theta)} - D(\overline{\theta}) \tag{3.55}$$

donde $\overline{D(\theta)} = \frac{1}{M}\sum_{m=1}^{M} D(\theta^{(m)})$, $D(\overline{\theta})$ es el desvío evaluado en $\frac{1}{M}\sum_{m=1}^{M}\theta^{(m)}$, $\theta^{(m)}$ son las muestras de parámetros a posteriori, y M es el número de iteraciones de MCMC.

Asimismo, el desvío se define como:

$$D(\theta) = -2(log(p(y \mid \theta)) - log(p(y \mid \theta_s))) \tag{3.56}$$

donde θ_s son las estimaciones del modelo saturado para la comparación de modelos, y el segundo término $log(p(y \mid \theta_s))$ es una constante, por lo que se puede omitir.

Finalmente, cuando se comparan modelos de regresión, desde el enfoque bayesiano se prefieren aquellos modelos con un menor valor de DIC.

Capítulo 4

Estudio de Simulación

En el estudio de simulación se aplica el método de fusión de efectos de nivel, el cual se realiza sobre un modelo de regresión lineal de cuatro covariables categóricas, las cuales se han generado de manera aleatoria con distintas cantidades de categorías. Para el estudio de simulación se hace uso del paquete effectFusion del software R; y se trabaja con distintos escenarios que se diferencian en los valores usados para la variabilidad de los componentes de la mixtura de la distribución a priori. Todo ello, con la finalidad de comparar los resultados de dichos escenarios, y analizar si existe alguna relación entre los resultados del criterio de información del desvío (DIC) para cada escenario y los distintos valores de la variabilidad de los componentes de la mixtura de la distribución a priori. Los códigos computacionales del estudio de simulación se encuentran en el Apéndice A.

Cabe mencionar que el paquete effectFusion del software R permite la aplicación del método de fusión de efectos de nivel mediante el método de inferencia bayesiana propuesto por Malsiner-Walli, Pauger y Wagner (2018). El paquete effectFusion ha sido desarrollado en Pauger D., Leitner M., Wagner H. y Malsiner-Walli G. (2019), véase en: `https://CRAN.R-project.org/package=effectFusion`.

4.1. Modelo en Estudio

A diferencia del Capítulo 3, en el cual se explicó el método de la fusión de efectos de nivel adaptado para una única covariable categórica; en el presente capítulo, se desarrolla un estudio de simulación para un modelo de regresión lineal con cuatro covariables categóricas, por ello, es necesario definir el método de fusión de efectos de nivel para varias covariables categóricas.

En ese sentido, teniendo como referencia la ecuación (2.18), que representa un modelo de regresión con varias variables categóricas, se define la estructura de la priori para el desarrollo del método de fusión de efectos de covariables categóricas con la siguiente expresión:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\beta_0) \prod_{j=1}^{J} p(\beta_j^* \mid \xi_j) p(\sigma^2) \tag{4.1}$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_{j1}, ..., \beta_{jc})$ son los efectos de la regresión, siendo $\boldsymbol{\beta}_j^* = (\beta_{j1}, ..., \beta_{jc})$; y ξ_j son los hiperparámetros adicionales de la covariable j, los cuales son: $\boldsymbol{\eta_j}, \boldsymbol{\mu_j}, \psi_j$. Además, se define que:

$$\begin{aligned} \beta_0 &\sim N(0, \psi_0) && (4.2)\\ \sigma^2 &\sim G^{-1}(s_0, S_0) && (4.3) \end{aligned}$$

Asimismo, es importante mencionar que la distribución a priori de un efecto de nivel, que se especifica jerárquicamente, tiene como objetivo clasificar a cada covariable categórica en un número fijo de grupos, siendo su expresión de la siguiente forma:

$$\begin{aligned} p(\beta_{jk}) &= \sum_{\ell=0}^{L_j} \eta_{j\ell} f_N(\beta_{jk} \mid \mu_{j\ell}, \psi_j) && (4.4)\\ \eta_j &\sim Dir_{L_j+1}(e_0) && (4.5)\\ \mu_{j0} &= 0 && (4.6)\\ \mu_{j\ell} &\sim N(m_{j0}, M_{j0}), \ell = 1, ..., L_j && (4.7)\\ \psi_j &\sim G^{-1}(g_0, G_{j0}) && (4.8) \end{aligned}$$

donde la mixtura contiene $L_j + 1$ componentes: $\ell = 0, ..., L_j$, y cada componente se define como una distribución normal f_N para modelar el efecto de nivel β_{jk} con los parámetros de localización $\mu_{j\ell}$ y de escala ψ_j. Siendo $\mu_{j0} = 0$ el primer componente del parámetro de locación, y $\boldsymbol{\mu_{j\ell}} = (\mu_{j1}, ..., \mu_{jL})$ una priori que se define con una distribución normal con parámetros de locación m_{j0} y de escala M_{j0}. Además, ψ_j es el parámetro de escala (aleatorio) mediante una priori que se define como una gamma inversa: $\psi \sim G^{-1}(g_0, G_0)$. Asimismo, $\boldsymbol{\eta_j}^\top = (\eta_{j0}, ..., \eta_{jL})$ son los pesos de los componentes de la mixtura, y es una priori que se asume que sigue una distribución de Dirichlet simétrica con parámetro e_0, siendo $e_0 > 0$, y cumpliéndose que $\eta_{j\ell} \geq 0$ y $\sum_{\ell=0}^{L_j} \eta_{j\ell} = 1$.

Cabe mencionar que los hiperparámetros se escogen en base a una estimación previa usando los datos, bajo una priori poco informativa, y se realiza de la siguiente manera:

$$\begin{aligned} m_{j0} &= media(\hat{\beta}_j^*) && (4.9)\\ M_{j0} &= (max_k \hat{\beta}_{jk}^* - min_k \hat{\beta}_{jk}^*)^2 && (4.10)\\ G_{j0} &= \frac{V_j}{v}(g_0 - 1) && (4.11) \end{aligned}$$

donde $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^*$ es el vector de coeficientes estimados de la covariable j bajo la priori descrita. Además, $V_j = \frac{1}{c_j - 1} \sum_{k=1}^{c_j} (\hat{\beta}_{jk}^* - \bar{\beta}_j^*)^2$ y $\bar{\beta}_j^* = \frac{1}{c_j} \sum_{k=1}^{c_j} \hat{\beta}_{jk}^*$.

Por último, en la sección 4.3 se definen los valores de los hiperparámetros: ψ_0, s_0, S_0, e_0, g_0 y v.

4.2. Simulación de Datos

Para iniciar con el estudio de simulación, primero se generan los datos aleatorios de las 4 covariables categóricas independientes y de la variable respuesta, mediante los siguientes pasos:

- Se generan 1,000 observaciones de manera aleatoria para cada unas de las covariables categóricas independientes, definiendo: 5 categorías para la covariable $\boldsymbol{x_1}$, 6 categorías para la covariable $\boldsymbol{x_2}$, 7 categorías para la covariable $\boldsymbol{x_3}$ y 8 categorías para la covariable $\boldsymbol{x_4}$. Asimismo, todas las covariables categóricas se asumen como nominales.
- Se define que los efectos de las categorías para la covariable $\boldsymbol{x_1}$ sean $\boldsymbol{\beta_1} = (0, 0, 0, 0, 0{,}5)$, para la covariable $\boldsymbol{x_2}$ sean $\boldsymbol{\beta_2} = (0, 0, 0{,}5, 0{,}5, 1, 1)$, para la covariable $\boldsymbol{x_3}$ sean $\boldsymbol{\beta_3} = (0, 0, 0{,}5, 0{,}5, 0{,}75, 0{,}75, 1)$, y para la covariable $\boldsymbol{x_4}$ sean $\boldsymbol{\beta_4} = (0, 0, 0{,}5, 0{,}5, 0{,}75, 0{,}75, 1, 1)$. Además, el intercepto β_0 se establece en 0; es preciso mencionar que, se puede asumir cualquier otro valor para el intercepto, sin que esto conlleve a tener variaciones en los resultados de las fusiones.
- Se generan 1,000 respuestas de manera aleatoria para la variable dependiente $\boldsymbol{y}$, mediante una regresión lineal simple con un error e que sigue una distribución normal de media 0 y varianza 0.5.

De la generación de datos se puede resumir que existen 4 covariables categóricas nominales y un total de 26 categorías (incluyendo el nivel de la línea base de cada covariable). En la Figura 4.1 se muestra el histograma de la variable respuesta $\boldsymbol{y}$, y los diagramas de cajas de la variable respuesta $\boldsymbol{y}$ con respecto a cada una de las 26 categorías pertenecientes a las 4 covariables $\boldsymbol{x_j}$.

En los diagramas de cajas de la Figura 4.1, se reflejan valores similares de la variable respuesta $\boldsymbol{y}$ para algunas categorías dentro de una misma covariable $\boldsymbol{x_j}$. Esto se debe a que algunos efectos de nivel β_{jk} se definieron con valores iguales dentro de cada covariable, durante la simulación de los datos. En ese sentido, considerando los niveles de línea base, la covariable $\boldsymbol{x_1}$ cuenta con 2 grupos de efectos de nivel, la covariable $\boldsymbol{x_2}$ cuenta con 3 grupos de efectos de nivel, y las covariables $\boldsymbol{x_3}$ y $\boldsymbol{x_4}$ cuentan con 4 grupos de efectos de nivel cada una.

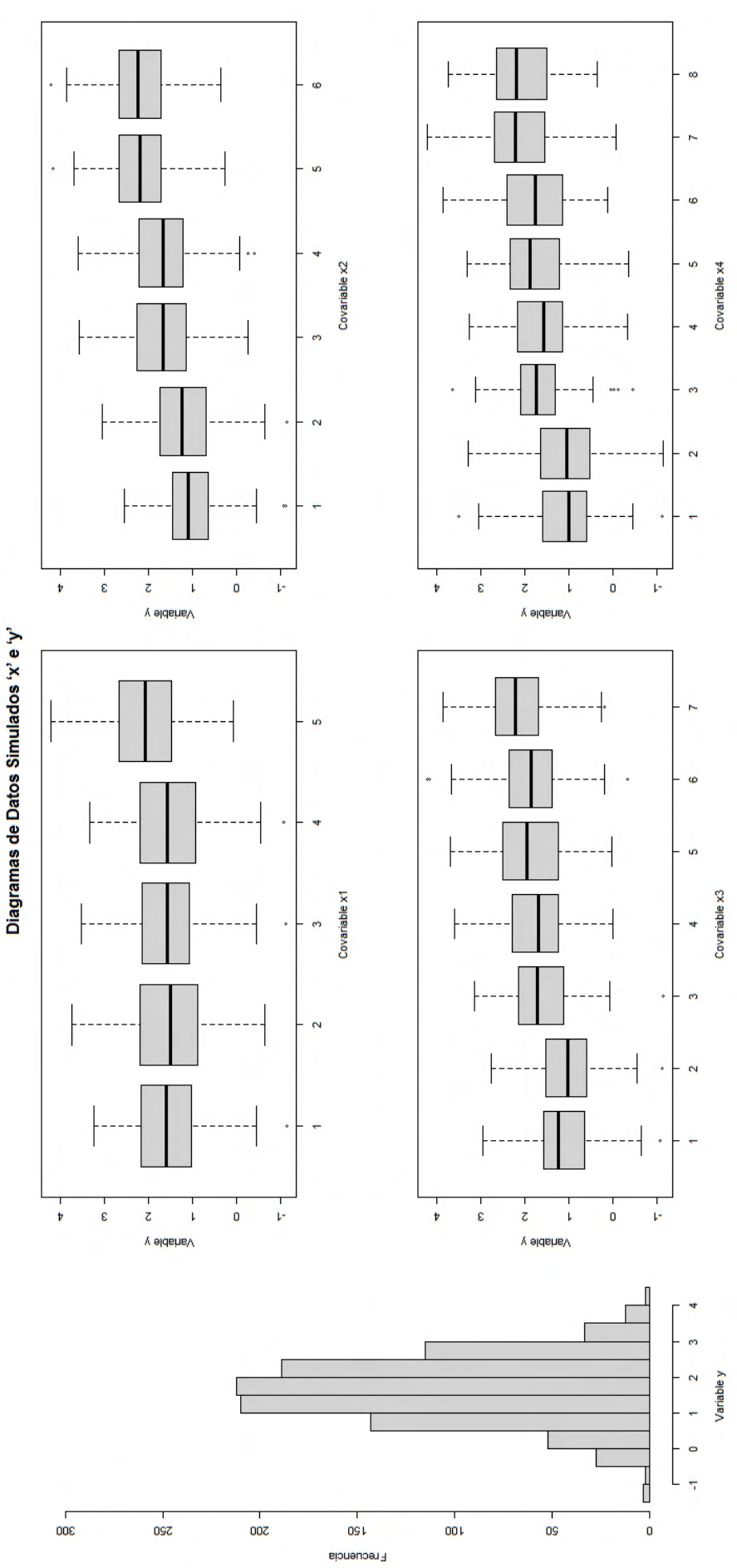

Figura 4.1: Diagramas de datos simulados: covariable categórica $\boldsymbol{x_j}$ y variable respuesta $\boldsymbol{y}$

4.3. Consideraciones para el Estudio de Simulación

En esta sección se describen las consideraciones para la estimación del modelo de regresión lineal con el uso del paquete effectFusion del software R, y también se definen los hiperparámetros de la distribución a priori.

- **method = 'FinMix':**
  Con este argumento, el método de fusión de efectos de nivel se realiza mediante una mixtura finita de normales como distribución a priori. Es importante mencionar que, este método no recoge la información de ordenamiento de las variables ordinales y trata a todas como variables nominales.

- **modelSelection = 'pam':**
  Con este argumento, la técnica de agrupamiento se realiza mediante la técnica 'pam', que usa los coeficientes de silhouette. Es importante mencionar que, esta técnica de agrupamiento no permite soluciones de un solo grupo, es decir, no es posible excluir covariables completas.

- **mcmc = c(M=30000, burnin=15000):**
  Con este argumento, la estimación del modelo de regresión por MCMC, considera 30,000 iteraciones, habiéndose descartado las primeras 15,000 iteraciones.

- Los hiperparámetros descritos en la sección 4.1, asumen los siguientes valores:
  Para el intercepto β_0: $\psi_0 = 0{,}01$.
  Para la varianza σ^2: $s_0 = 0$, $S_0 = 0$.
  Para los pesos de la mixtura $\boldsymbol{\eta_j}$: $e_0 = 0{,}01$.
  Para la varianza de los componentes de la mixtura $\boldsymbol{\psi_j}$: $g_0 = 100$ y v tomará 6 distintos valores en G_{j0}: 10, 10^2, 10^3, 10^4, 10^5, 10^6.

Al asumir distintos valores de v, se obtendrán diferentes variabilidades en los componentes de la mixtura de la distribución a priori. Con lo cual, se conseguirá distintas estimaciones del modelo de regresión lineal, y se podrá conocer el modelo que mejor se ajusta a los datos simulados. Cabe recordar que, a mayor valor de v se obtiene menor variabilidad en los componentes de la mixtura.

Con lo mencionado anteriormente, se definen los siguientes 7 modelos de regresión lineal (escenarios) para la etapa de simulación:

1. **Modelo General:** Modelo de regresión lineal general. No se usa un método para la fusión de efectos, y ni una técnica de agrupamiento.

2. **Modelo MF01:** Modelo de regresión lineal usando el método de mixtura finita de normales, la técnica de agrupamiento 'pam', y una variabilidad de los componentes de la mixtura con $v = 10$.

3. **Modelo MF02:** Modelo de regresión lineal usando el método de mixtura finita de normales, la técnica de agrupamiento 'pam', y una variabilidad de los componentes de la mixtura con $v = 10^2$.

4. **Modelo MF03:** Modelo de regresión lineal usando el método de mixtura finita de normales, la técnica de agrupamiento 'pam', y una variabilidad de los componentes de la mixtura con $v = 10^3$.

5. **Modelo MF04:** Modelo de regresión lineal usando el método de mixtura finita de normales, la técnica de agrupamiento 'pam', y una variabilidad de los componentes de la mixtura con $v = 10^4$.

6. **Modelo MF05:** Modelo de regresión lineal usando el método de mixtura finita de normales, la técnica de agrupamiento 'pam', y una variabilidad de los componentes de la mixtura con $v = 10^5$.

7. **Modelo MF06:** Modelo de regresión lineal usando el método de mixtura finita de normales, la técnica de agrupamiento 'pam', y una variabilidad de los componentes de la mixtura con $v = 10^6$.

4.4. Resultados de Simulación

Los resultados de las estimaciones de los 7 modelos de regresión lineal (escenarios) se muestran en la Tabla 4.1, donde se presenta el valor promedio de las estimaciones de los β_{jk} de cada modelo (escenario), y los valores mínimos y máximos de los intervalos de credibilidad al 95 %.

Adicionalmente, en la Tabla 4.2, se muestra el valor promedio de las re-estimaciones de los β_{jk}, considerando el modelo seleccionado en cada escenario, y los valores mínimos y máximos de los intervalos de credibilidad al 95 %. En dicha tabla, los efectos de nivel fusionados presentan un único valor, y los efectos de nivel con valor de 0 indican que han sido fusionados con el nivel de la línea base de cada covariable. Cabe mencionar que el paquete effectFusion, realiza la re-estimación del modelo seleccionado con 3,000 iteraciones, descartando las 1,000 iteraciones iniciales.

Con la información anterior, se diseña la Tabla 4.3, donde se muestran los efectos de nivel fusionados en los modelos de regresión lineal (escenarios). Asimismo, desde la Figura 4.2 hasta la Figura 4.8, se muestran los diagramas de cajas de las estimaciones y re-estimaciones de los β_{jk} en cada escenario.

			Valores estimados de los efectos de nivel																			
			Modelo General			Modelo 01			Modelo 02			Modelo 03			Modelo 04			Modelo 05			Modelo 06		
Covariable	Efecto	Valor inicial	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)
	β_0	0.000	-0.31115	-0.16320	-0.01588	-0.25410	-0.13757	-0.02228	-0.15567	-0.06770	0.01825	-0.09822	-0.02914	0.04040	-0.10341	-0.03755	0.02712	-0.15145	-0.09291	-0.02836	-0.15467	-0.11885	-0.08434
x_1	$\beta_{1,2}$	0.000	-0.08455	0.01124	0.10616	-0.07024	0.00729	0.08624	-0.04230	0.00142	0.04517	-0.01634	0.00008	0.01649	-0.00517	0.00003	0.00519	-0.00168	0.00001	0.00162	-0.00051	0.00000	0.00053
	$\beta_{1,3}$	0.000	-0.06511	0.02884	0.12558	-0.05918	0.02024	0.09981	-0.03619	0.00723	0.05075	-0.01536	0.00094	0.01770	-0.00521	0.00007	0.00518	-0.00167	0.00000	0.00163	-0.00054	0.00000	0.00051
	$\beta_{1,4}$	0.000	-0.08369	0.01041	0.10716	-0.06951	0.00627	0.08899	-0.04322	0.00016	0.04442	-0.01619	0.00019	0.01691	-0.00515	0.00003	0.00524	-0.00165	0.00000	0.00168	-0.00052	0.00000	0.00052
	$\beta_{1,5}$	0.500	0.45198	0.54850	0.64525	0.46744	0.54650	0.62982	0.47827	0.53930	0.60104	0.47785	0.53620	0.59228	0.46241	0.53350	0.59684	0.46099	0.52970	0.58720	0.49879	0.51900	0.53683
x_2	$\beta_{2,2}$	0.000	0.01916	0.12157	0.22983	0.01722	0.12352	0.22592	-0.02142	0.05529	0.14882	-0.01721	0.00648	0.02993	-0.00705	0.00065	0.00809	-0.00327	0.07296	0.16731	0.05430	0.09162	0.14960
	$\beta_{2,3}$	0.500	0.50175	0.60590	0.71302	0.50506	0.60380	0.70344	0.48153	0.55770	0.63768	0.46947	0.52650	0.58464	0.46955	0.52740	0.58658	0.50929	0.56440	0.61898	0.53438	0.55430	0.57228
	$\beta_{2,4}$	0.500	0.46300	0.56200	0.66361	0.47070	0.56260	0.65659	0.46024	0.53500	0.61052	0.46578	0.52250	0.57963	0.46958	0.52700	0.58631	0.50948	0.56430	0.61917	0.53414	0.55430	0.57198
	$\beta_{2,5}$	1.000	0.94907	1.05330	1.16286	0.93746	1.03980	1.13877	0.94944	1.02610	1.10541	0.94982	1.00990	1.06606	0.94354	1.00280	1.06247	0.98118	1.04080	1.09528	1.04592	1.05900	1.07119
	$\beta_{2,6}$	1.000	0.96257	1.06950	1.17119	0.96047	1.05520	1.15660	0.95831	1.03700	1.11278	0.95517	1.01230	1.07062	0.94307	1.00310	1.06165	0.98095	1.04100	1.09492	1.04574	1.05900	1.07096
x_3	$\beta_{3,2}$	0.000	-0.07075	0.04729	0.16125	-0.06343	0.03127	0.12781	-0.04527	0.00985	0.06541	-0.01884	0.00102	0.02005	-0.00617	0.00007	0.00623	-0.00190	0.00003	0.00209	-0.00063	0.00000	0.00061
	$\beta_{3,3}$	0.500	0.37383	0.48850	0.60689	0.41186	0.50820	0.60770	0.44317	0.53530	0.61758	0.42473	0.52330	0.62482	0.41466	0.46780	0.51983	0.46115	0.51720	0.55664	0.48999	0.50330	0.51752
	$\beta_{3,4}$	0.500	0.41107	0.52490	0.63939	0.44544	0.54080	0.63877	0.45897	0.54750	0.63308	0.42878	0.52500	0.62656	0.41394	0.46800	0.51922	0.46314	0.51720	0.55845	0.49005	0.50330	0.51764
	$\beta_{3,5}$	0.750	0.61734	0.72890	0.84631	0.62167	0.71800	0.81742	0.56253	0.65450	0.77869	0.55017	0.67600	0.81492	0.64184	0.69030	0.74383	0.68732	0.72410	0.76598	0.67037	0.70090	0.72533
	$\beta_{3,6}$	0.750	0.59469	0.70350	0.81708	0.59924	0.69280	0.78657	0.55222	0.64220	0.74359	0.54367	0.66830	0.81435	0.64161	0.69010	0.74464	0.68717	0.72410	0.76603	0.67054	0.70090	0.72548
	$\beta_{3,7}$	1.000	0.89436	1.00710	1.11847	0.86311	0.95870	1.05438	0.87264	0.97920	1.07101	0.76778	0.93860	1.04605	0.90781	0.97140	1.03704	0.98759	1.03370	1.08537	0.95823	0.98980	1.01143
x_4	$\beta_{4,2}$	0.000	-0.04167	0.07902	0.19519	-0.03652	0.06103	0.15826	-0.03738	0.02023	0.07996	-0.01855	0.00244	0.02246	-0.00598	0.00033	0.00709	-0.00205	0.00002	0.00211	0.06284	0.08161	0.10602
	$\beta_{4,3}$	0.500	0.52303	0.64610	0.77094	0.55667	0.66600	0.76596	0.57523	0.65270	0.73064	0.60649	0.66760	0.73140	0.64911	0.69530	0.74360	0.57049	0.60100	0.62736	0.64264	0.65750	0.67329
	$\beta_{4,4}$	0.500	0.51744	0.63200	0.75495	0.54746	0.64740	0.74794	0.56989	0.64820	0.72179	0.60647	0.66610	0.72835	0.64835	0.69520	0.74288	0.57059	0.60100	0.62762	0.64251	0.65750	0.67316
	$\beta_{4,5}$	0.750	0.65825	0.78030	0.90626	0.67348	0.77820	0.88141	0.60789	0.70990	0.80552	0.61531	0.67720	0.74192	0.64914	0.69600	0.74373	0.72592	0.75970	0.79712	0.77741	0.81270	0.83528
	$\beta_{4,6}$	0.750	0.69573	0.82020	0.94054	0.70731	0.81150	0.91421	0.64190	0.74460	0.91600	0.61513	0.68030	0.74598	0.64914	0.69620	0.74368	0.72571	0.75970	0.79701	0.77743	0.81270	0.83532
	$\beta_{4,7}$	1.000	0.95884	1.08050	1.20626	0.93075	1.03480	1.13756	0.93165	1.02160	1.10213	0.95013	1.01350	1.07611	0.97776	1.03700	1.10620	0.98065	1.01890	1.05937	1.00437	1.02940	1.04939
	$\beta_{4,8}$	1.000	0.91761	1.03950	1.15738	0.89667	0.99850	1.09919	0.92155	1.00460	1.09205	0.94719	1.01100	1.07188	0.97811	1.03680	1.10643	0.98027	1.01890	1.05916	1.00461	1.02900	1.04961

Tabla 4.1: Resultados de los valores estimados de los efectos de nivel del modelo general y los 6 modelos de mixtura finita, en la etapa de simulación

			Valores re-estimados de los efectos de nivel																	
			Modelo MF01			Modelo MF02			Modelo MF03			Modelo MF04			Modelo MF05			Modelo MF06		
Covariable	Efecto	Valor inicial	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)
	β_0	0.000	-0.11715	-0.02553	0.06795	-0.10718	-0.02622	0.06471	-0.10981	-0.02497	0.06164	-0.11336	-0.02742	0.05995	-0.11493	-0.02874	0.05553	-0.24066	-0.12324	-0.00246
x_1	$\beta_{1,2}$	0.000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
	$\beta_{1,3}$	0.000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
	$\beta_{1,4}$	0.000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
	$\beta_{1,5}$	0.500	0.46863	0.55440	0.63157	0.45962	0.53400	0.61308	0.45779	0.53490	0.61076	0.45616	0.53580	0.60959	0.44833	0.52970	0.60134	0.45879	0.53430	0.60985
x_2	$\beta_{2,2}$	0.000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01048	0.11730	0.21935
	$\beta_{2,3}$	0.500	0.42932	0.51090	0.58902	0.45198	0.52580	0.59858	0.44619	0.52350	0.59477	0.45513	0.52580	0.59875	0.45626	0.52770	0.59662	0.49788	0.58250	0.67663
	$\beta_{2,4}$	0.500	0.42932	0.51090	0.58902	0.45198	0.52580	0.59858	0.44619	0.52350	0.59477	0.45513	0.52580	0.59875	0.45626	0.52770	0.59662	0.49788	0.58250	0.67663
	$\beta_{2,5}$	1.000	0.92883	1.01220	1.09877	0.93052	1.00750	1.07952	0.93686	1.00680	1.08402	0.93378	1.00900	1.08490	0.93387	1.00510	1.07878	0.97611	1.06090	1.15651
	$\beta_{2,6}$	1.000	0.92883	1.01220	1.09877	0.93052	1.00750	1.07952	0.93686	1.00680	1.08402	0.93378	1.00900	1.08490	0.93387	1.00510	1.07878	0.97611	1.06090	1.15651
x_3	$\beta_{3,2}$	0.000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
	$\beta_{3,3}$	0.500	0.59466	0.67130	0.74370	0.39258	0.47480	0.55776	0.38969	0.47500	0.55332	0.39550	0.47550	0.55775	0.39511	0.47890	0.55679	0.39879	0.48400	0.56481
	$\beta_{3,4}$	0.500	0.59466	0.67130	0.74370	0.39258	0.47480	0.55776	0.38969	0.47500	0.55332	0.39550	0.47550	0.55775	0.39511	0.47890	0.55679	0.39879	0.48400	0.56481
	$\beta_{3,5}$	0.750	0.59466	0.67130	0.74370	0.61790	0.69550	0.78139	0.61681	0.69560	0.77336	0.62005	0.69540	0.77726	0.62353	0.69900	0.78257	0.61598	0.69550	0.77933
	$\beta_{3,6}$	0.750	0.59466	0.67130	0.74370	0.61790	0.69550	0.78139	0.61681	0.69560	0.77336	0.62005	0.69540	0.77726	0.62353	0.69900	0.78257	0.61598	0.69550	0.77933
	$\beta_{3,7}$	1.000	0.59466	0.67130	0.74370	0.88241	0.98010	1.07624	0.88161	0.97980	1.07761	0.88368	0.97980	1.07498	0.89216	0.98630	1.08431	0.88664	0.98480	1.08136
x_4	$\beta_{4,2}$	0.000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.05066	0.07276	0.18721
	$\beta_{4,3}$	0.500	0.71207	0.78510	0.86357	0.60273	0.67300	0.74829	0.59958	0.67210	0.74702	0.60252	0.67380	0.74983	0.50740	0.59560	0.67601	0.52565	0.63400	0.73437
	$\beta_{4,4}$	0.500	0.71207	0.78510	0.86357	0.60273	0.67300	0.74829	0.59958	0.67210	0.74702	0.60252	0.67380	0.74983	0.50740	0.59560	0.67601	0.52565	0.63400	0.73437
	$\beta_{4,5}$	0.750	0.71207	0.78510	0.86357	0.60273	0.67300	0.74829	0.59958	0.67210	0.74702	0.60252	0.67380	0.74983	0.66757	0.75590	0.83946	0.68793	0.79590	0.90009
	$\beta_{4,6}$	0.750	0.71207	0.78510	0.86357	0.60273	0.67300	0.74829	0.59958	0.67210	0.74702	0.60252	0.67380	0.74983	0.66757	0.75590	0.83946	0.68793	0.79590	0.90009
	$\beta_{4,7}$	1.000	0.71207	0.78510	0.86357	0.92954	1.01400	1.09856	0.93447	1.01390	1.10516	0.92836	1.01450	1.10114	0.93272	1.01470	1.09733	0.95252	1.05270	1.16408
	$\beta_{4,8}$	1.000	0.71207	0.78510	0.86357	0.92954	1.01400	1.09856	0.93447	1.01390	1.10516	0.92836	1.01450	1.10114	0.93272	1.01470	1.09733	0.95252	1.05270	1.16408

Tabla 4.2: Resultados de los valores re-estimados de los efectos de nivel de los 6 modelos de mixtura finita, en la etapa de simulación

		Fusión de efectos de nivel						
Covariable	Nro de grupos	Fusión inicial	Modelo MF01	Modelo MF02	Modelo MF03	Modelo MF04	Modelo MF05	Modelo MF06
Intercepto		β_0	β_0	β_0	β_0	β_0	β_0	β_0
$\boldsymbol{x_1}$	1	$\beta_{1,1}$, $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{1,4}$	$\beta_{1,1}$, $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{1,4}$	$\beta_{1,1}$, $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{1,4}$	$\beta_{1,1}$, $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{1,4}$	$\beta_{1,1}$, $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{1,4}$	$\beta_{1,1}$, $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{1,4}$	$\beta_{1,1}$, $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{1,4}$
	2	$\beta_{1,5}$	$\beta_{1,5}$	$\beta_{1,5}$	$\beta_{1,5}$	$\beta_{1,5}$	$\beta_{1,5}$	$\beta_{1,5}$
$\boldsymbol{x_2}$	1	$\beta_{2,1}$, $\beta_{2,2}$	$\beta_{2,1}$, $\beta_{2,2}$	$\beta_{2,1}$, $\beta_{2,2}$	$\beta_{2,1}$, $\beta_{2,2}$	$\beta_{2,1}$, $\beta_{2,2}$	$\beta_{2,1}$, $\beta_{2,2}$	$\beta_{2,1}$
	2	$\beta_{2,3}$, $\beta_{2,4}$	$\beta_{2,3}$, $\beta_{2,4}$	$\beta_{2,3}$, $\beta_{2,4}$	$\beta_{2,3}$, $\beta_{2,4}$	$\beta_{2,3}$, $\beta_{2,4}$	$\beta_{2,3}$, $\beta_{2,4}$	$\beta_{2,2}$
	3	$\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,3}$, $\beta_{2,4}$
	4							$\beta_{2,5}$, $\beta_{2,6}$
$\boldsymbol{x_3}$	1	$\beta_{3,1}$, $\beta_{3,2}$	$\beta_{3,1}$, $\beta_{3,2}$	$\beta_{3,1}$, $\beta_{3,2}$	$\beta_{3,1}$, $\beta_{3,2}$	$\beta_{3,1}$, $\beta_{3,2}$	$\beta_{3,1}$, $\beta_{3,2}$	$\beta_{3,1}$, $\beta_{3,2}$
	2	$\beta_{3,3}$, $\beta_{3,4}$	$\beta_{3,3}$, $\beta_{3,4}$, $\beta_{3,5}$, $\beta_{3,6}$, $\beta_{3,7}$	$\beta_{3,3}$, $\beta_{3,4}$	$\beta_{3,3}$, $\beta_{3,4}$	$\beta_{3,3}$, $\beta_{3,4}$	$\beta_{3,3}$, $\beta_{3,4}$	$\beta_{3,3}$, $\beta_{3,4}$
	3	$\beta_{3,5}$, $\beta_{3,6}$		$\beta_{3,5}$, $\beta_{3,6}$	$\beta_{3,5}$, $\beta_{3,6}$	$\beta_{3,5}$, $\beta_{3,6}$	$\beta_{3,5}$, $\beta_{3,6}$	$\beta_{3,5}$, $\beta_{3,6}$
	4	$\beta_{3,7}$		$\beta_{3,7}$	$\beta_{3,7}$	$\beta_{3,7}$	$\beta_{3,7}$	$\beta_{3,7}$
$\boldsymbol{x_4}$	1	$\beta_{4,1}$, $\beta_{4,2}$	$\beta_{4,1}$, $\beta_{4,2}$	$\beta_{4,1}$, $\beta_{4,2}$	$\beta_{4,1}$, $\beta_{4,2}$	$\beta_{4,1}$, $\beta_{4,2}$	$\beta_{4,1}$, $\beta_{4,2}$	$\beta_{4,1}$
	2	$\beta_{4,3}$, $\beta_{4,4}$	$\beta_{4,3}$, $\beta_{4,4}$, $\beta_{4,5}$, $\beta_{4,6}$, $\beta_{4,7}$, $\beta_{4,8}$	$\beta_{4,3}$, $\beta_{4,4}$, $\beta_{4,5}$, $\beta_{4,6}$	$\beta_{4,3}$, $\beta_{4,4}$, $\beta_{4,5}$, $\beta_{4,6}$	$\beta_{4,3}$, $\beta_{4,4}$, $\beta_{4,5}$, $\beta_{4,6}$	$\beta_{4,3}$, $\beta_{4,4}$	$\beta_{4,2}$
	3	$\beta_{4,5}$, $\beta_{4,6}$		$\beta_{4,7}$, $\beta_{4,8}$	$\beta_{4,7}$, $\beta_{4,8}$	$\beta_{4,7}$, $\beta_{4,8}$	$\beta_{4,5}$, $\beta_{4,6}$	$\beta_{4,3}$, $\beta_{4,4}$
	4	$\beta_{4,7}$, $\beta_{4,8}$					$\beta_{4,7}$, $\beta_{4,8}$	$\beta_{4,5}$, $\beta_{4,6}$
	5							$\beta_{4,7}$, $\beta_{4,8}$

Tabla 4.3: Comparación de la fusión de los efectos de nivel de los 6 modelos de mixtura finita, en la etapa de simulación

A continuación, se describen los resultados de las estimaciones de los 7 modelos de regresión lineal (escenarios); esto permitirá la comparación de las fusiones de los efectos de nivel resultantes con los efectos de nivel definidos al inicio de la simulación de los datos.

1. **Modelo General:**
   El modelo general presenta valores distintos para cada efecto de nivel de las covariables $\boldsymbol{x_j}$, es decir, no presenta fusiones, y por ende, es el modelo más extenso con:
   - 5 grupos de efectos de nivel para la covariable $\boldsymbol{x_1}$,
   - 6 grupos de efectos de nivel para la covariable $\boldsymbol{x_2}$,
   - 7 grupos de efectos de nivel para la covariable $\boldsymbol{x_3}$,
   - 8 grupos de efectos de nivel para la covariable $\boldsymbol{x_4}$.

   Sin embargo, es importante recalcar que los valores estimados de β_{jk} se asemejan a los valores simulados en un inicio, véase Figura 4.2.

2. **Modelo MF01 (pam ; $v = 10$):**
   El modelo de mixtura finita 01 es el más reducido (considerando sólo los modelos de mixtura finita) con:
   - 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_1}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_2}$,
   - 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_3}$,
   - 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_4}$.

   Las fusiones de los efectos de nivel de las covariables $\boldsymbol{x_1}$ y $\boldsymbol{x_2}$ coinciden con los valores simulados, sin embargo, las covariables $\boldsymbol{x_3}$ y $\boldsymbol{x_4}$ presentan una mayor fusión de efectos; véase Figura 4.3. Adicionalmente, en la Tabla 4.3, se puede evidenciar que en total

se obtienen 9 grupos de efectos de nivel, incluyendo los niveles de línea base de cada covariable.

3. **Modelo MF02 (pam ; $v = 10^2$):**
   El modelo de mixtura finita 02 cuenta con:
   - 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_1}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_2}$,
   - 4 grupos de efectos de nivel para la covariable $\boldsymbol{x_3}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_4}$.

   Las fusiones de los efectos de nivel de las covariables $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ y $\boldsymbol{x_3}$ coinciden con los valores simulados, sin embargo, la covariable $\boldsymbol{x_4}$ presenta una mayor fusión de efectos; véase Figura 4.4. Adicionalmente, en la Tabla 4.3, se puede evidenciar que en total se obtienen 12 grupos de efectos de nivel, incluyendo los niveles de línea base de cada covariable.

4. **Modelo MF03 (pam ; $v = 10^3$):**
   El modelo de mixtura finita 03 cuenta con:
   - 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_1}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_2}$,
   - 4 grupos de efectos de nivel para la covariable $\boldsymbol{x_3}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_4}$.

   Las fusiones de los efectos de nivel de las covariables $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ y $\boldsymbol{x_3}$ coinciden con los valores simulados, sin embargo, la covariable $\boldsymbol{x_4}$ presenta una mayor fusión de efectos; véase Figura 4.5. Adicionalmente, en la Tabla 4.3, se puede evidenciar que en total se obtienen 12 grupos de efectos de nivel, incluyendo los niveles de línea base de cada covariable.

5. **Modelo MF04 (pam ; $v = 10^4$):**
   El modelo de mixtura finita 04 cuenta con con:
   - 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_1}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_2}$,
   - 4 grupos de efectos de nivel para la covariable $\boldsymbol{x_3}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_4}$.

   Las fusiones de los efectos de nivel de las covariables $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ y $\boldsymbol{x_3}$ coinciden con los valores simulados, sin embargo, la covariable $\boldsymbol{x_4}$ presenta una mayor fusión de efectos; véase Figura 4.6. Adicionalmente, en la Tabla 4.3, se puede evidenciar que en total se obtienen 12 grupos de efectos de nivel, incluyendo los niveles de línea base de cada covariable.

6. **Modelo MF05 (pam ; $v = 10^5$):**
   El modelo de mixtura finita 05 cuenta con:
   - 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_1}$,
   - 3 grupos de efectos de nivel para la covariable $\boldsymbol{x_2}$,
   - 4 grupos de efectos de nivel para la covariable $\boldsymbol{x_3}$,

- 4 grupos de efectos de nivel para la covariable $\boldsymbol{x_4}$.

Las fusiones de los efectos de nivel de las 4 covariables ($\boldsymbol{x_1}$, $\boldsymbol{x_2}$ y $\boldsymbol{x_3}$ y $\boldsymbol{x_4}$) coinciden con los valores simulados; véase Figura 4.7. Adicionalmente, en la Tabla 4.3, se puede evidenciar que en total se obtienen 13 grupos de efectos de nivel, incluyendo los niveles de línea base de cada covariable.

7. **Modelo MF06 (pam ; $v = 10^6$):**
El modelo de mixtura finita 06 es el más extenso (considerando sólo los modelos de mixtura finita) con:
- 2 grupos de efectos de nivel para la covariable $\boldsymbol{x_1}$,
- 4 grupos de efectos de nivel para la covariable $\boldsymbol{x_2}$,
- 4 grupos de efectos de nivel para la covariable $\boldsymbol{x_3}$,
- 5 grupos de efectos de nivel para la covariable $\boldsymbol{x_4}$.

Las fusiones de los efectos de nivel de las covariables $\boldsymbol{x_1}$ y $\boldsymbol{x_3}$ coinciden con los valores simulados, sin embargo, las covariables $\boldsymbol{x_2}$ y $\boldsymbol{x_4}$ presentan una menor fusión de efectos; véase Figura 4.8. Adicionalmente, en la Tabla 4.3, se puede evidenciar que en total se obtienen 15 grupos de efectos de nivel, incluyendo los niveles de línea base de cada covariable.

De los resultados obtenidos y descritos anteriormente, y teniendo como referencia la Tabla 4.3 se resume lo siguiente:

- Con el modelo MF01 se obtiene el modelo de regresión lineal más reducido que el original, dado que las covariables $\boldsymbol{x_3}$ y $\boldsymbol{x_4}$ presentan menos grupos de efectos de nivel.
- Con el modelo MF06 se obtiene el modelo de regresión lineal más extenso que el original, dado que las covariables $\boldsymbol{x_2}$ y $\boldsymbol{x_4}$ presentan más grupos de efectos de nivel.
- Con el modelo MF05 se obtiene un modelo de regresión lineal idéntico al original, dado que las covariables $\boldsymbol{x_j}$ presentan la misma cantidad de grupos de efectos de nivel.
- Con los modelos MF02, MF03 y MF04 se obtienen modelos de regresión lineal similares pero un poco más reducidos que el original, dado que la covariable $\boldsymbol{x_4}$ presenta menos grupos de efectos de nivel.

En ese sentido, se puede afirmar que al trabajar con una mayor variabilidad (menores valores de v) en los componentes de la mixtura de la distribución a priori, se consiguen modelos de regresión lineal más reducidos. Adicionalmente, se puede observar que en los 6 modelos donde se usa la mixtura finita como distribución a priori, ninguna covariable se excluye del modelo de regresión; esto se debe a que la técnica de agrupamiento 'pam' no permite soluciones de un solo grupo en una covariable.

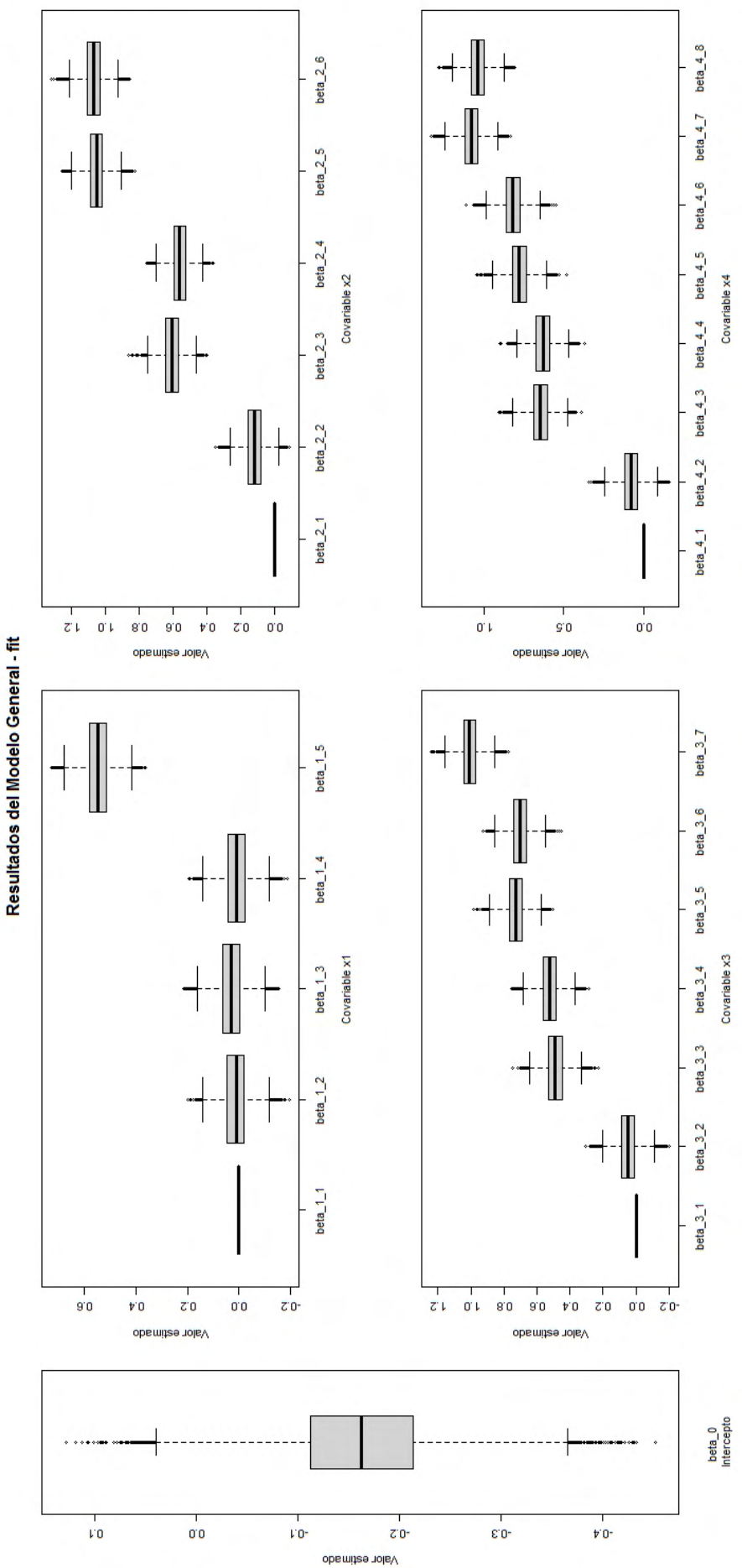

Figura 4.2: Resultados de la estimación de los efectos de nivel en el modelo general

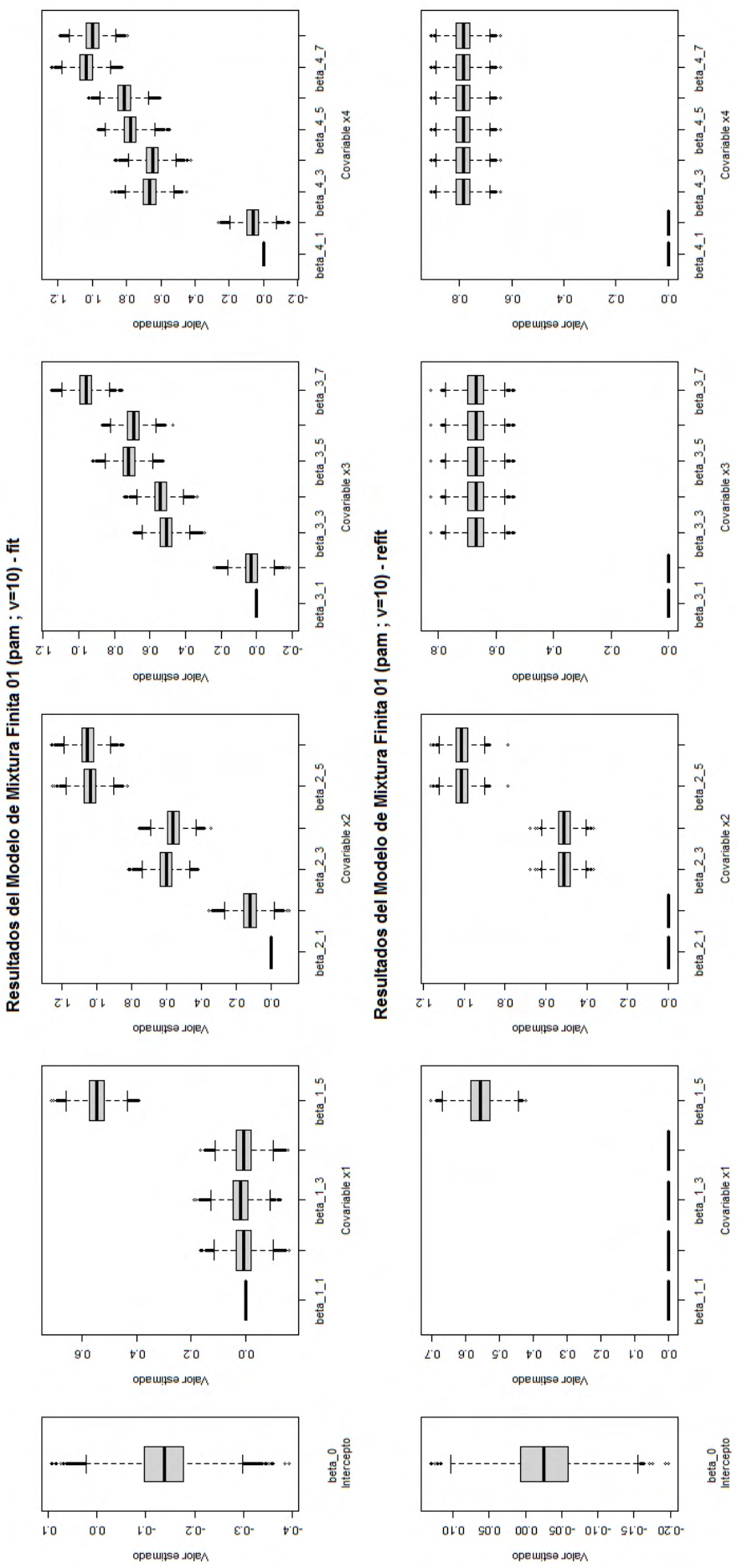

Figura 4.3: Resultados de la estimación y re-estimación de los efectos de nivel en el modelo de mixtura finita 01 (pam ; $v = 10$)

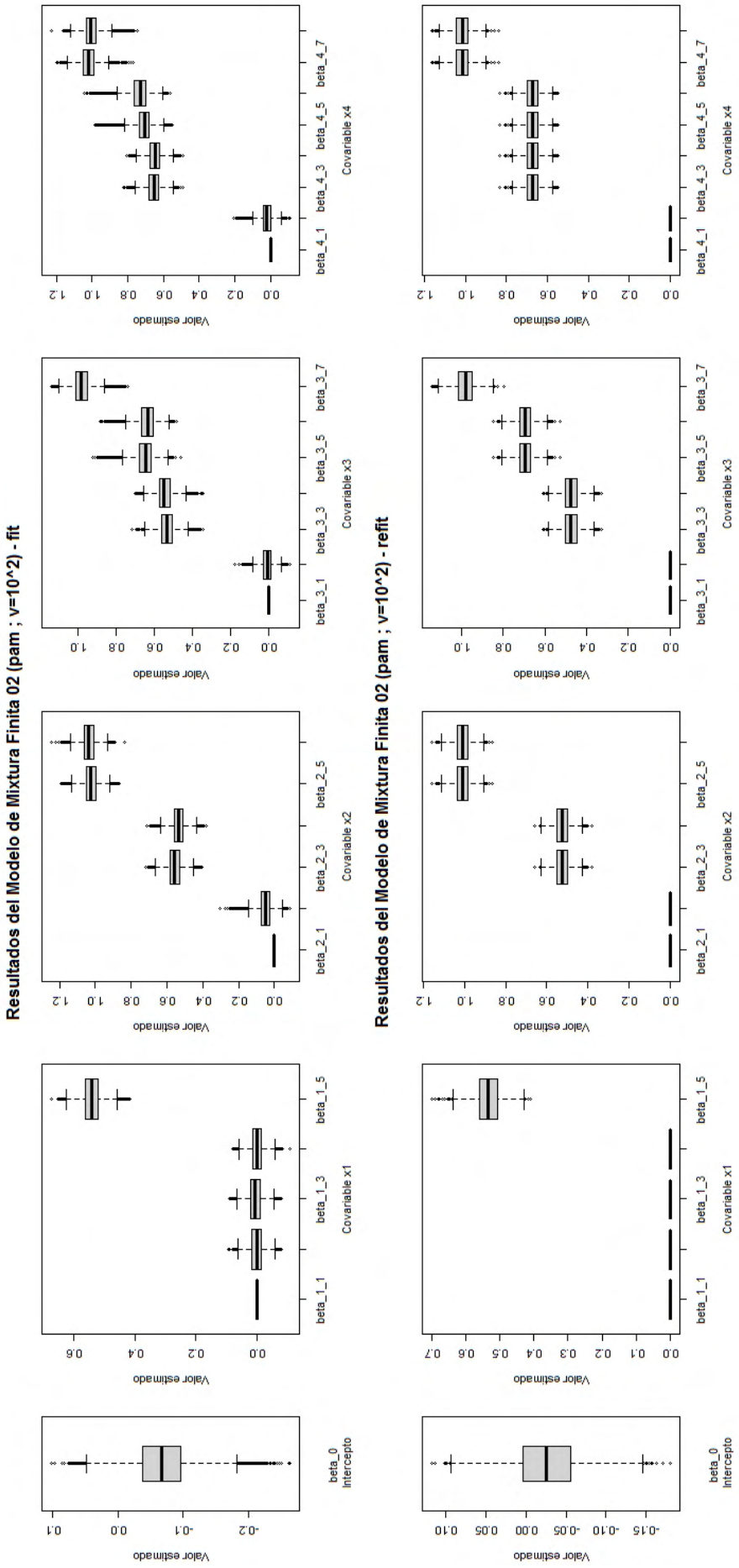

Figura 4.4: Resultados de la estimación y re-estimación de los efectos de nivel en el modelo de mixtura finita 02 (pam ; $v = 10^2$)

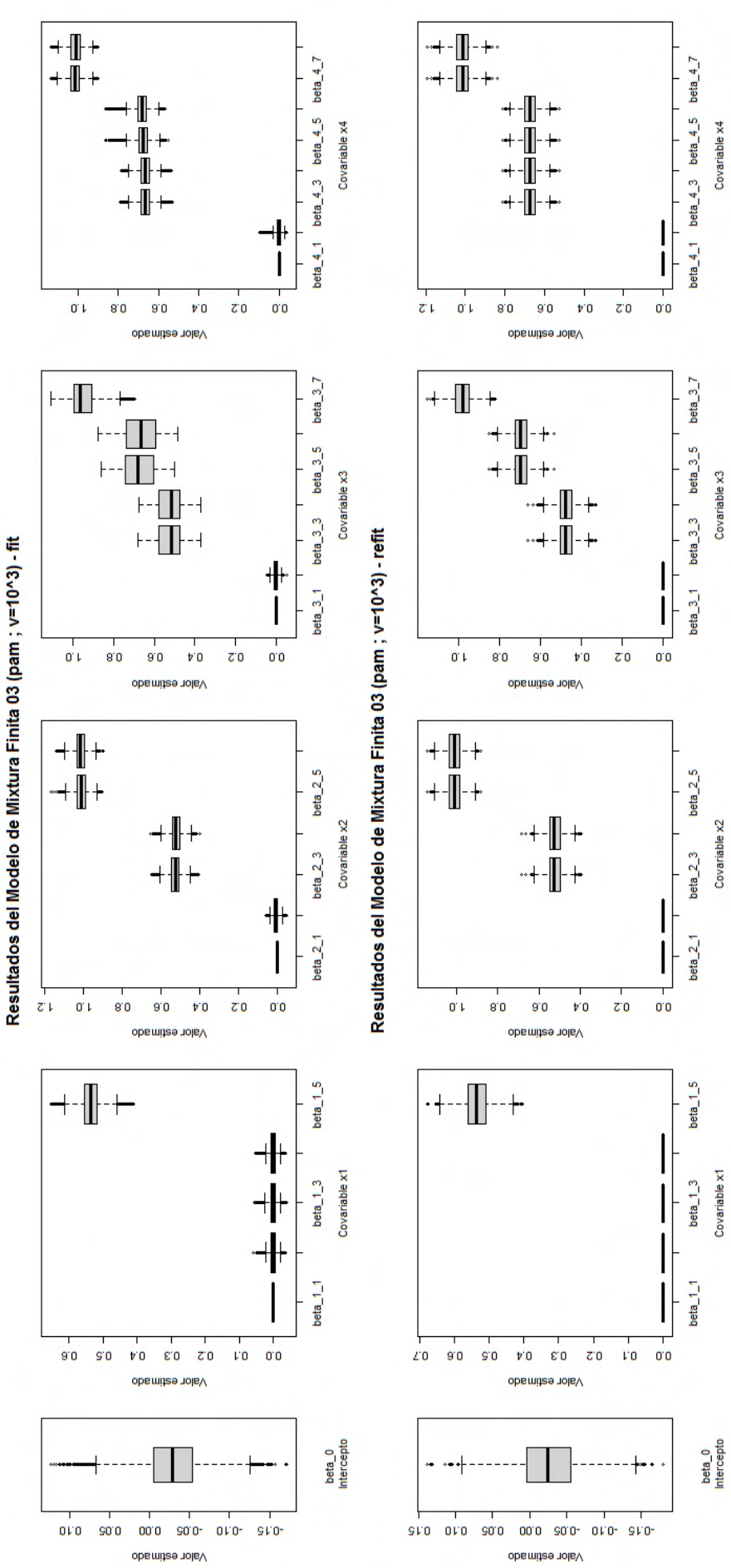

Figura 4.5: Resultados de la estimación y re-estimación de los efectos de nivel en el modelo de mixtura finita 03 (pam ; $v = 10^3$)

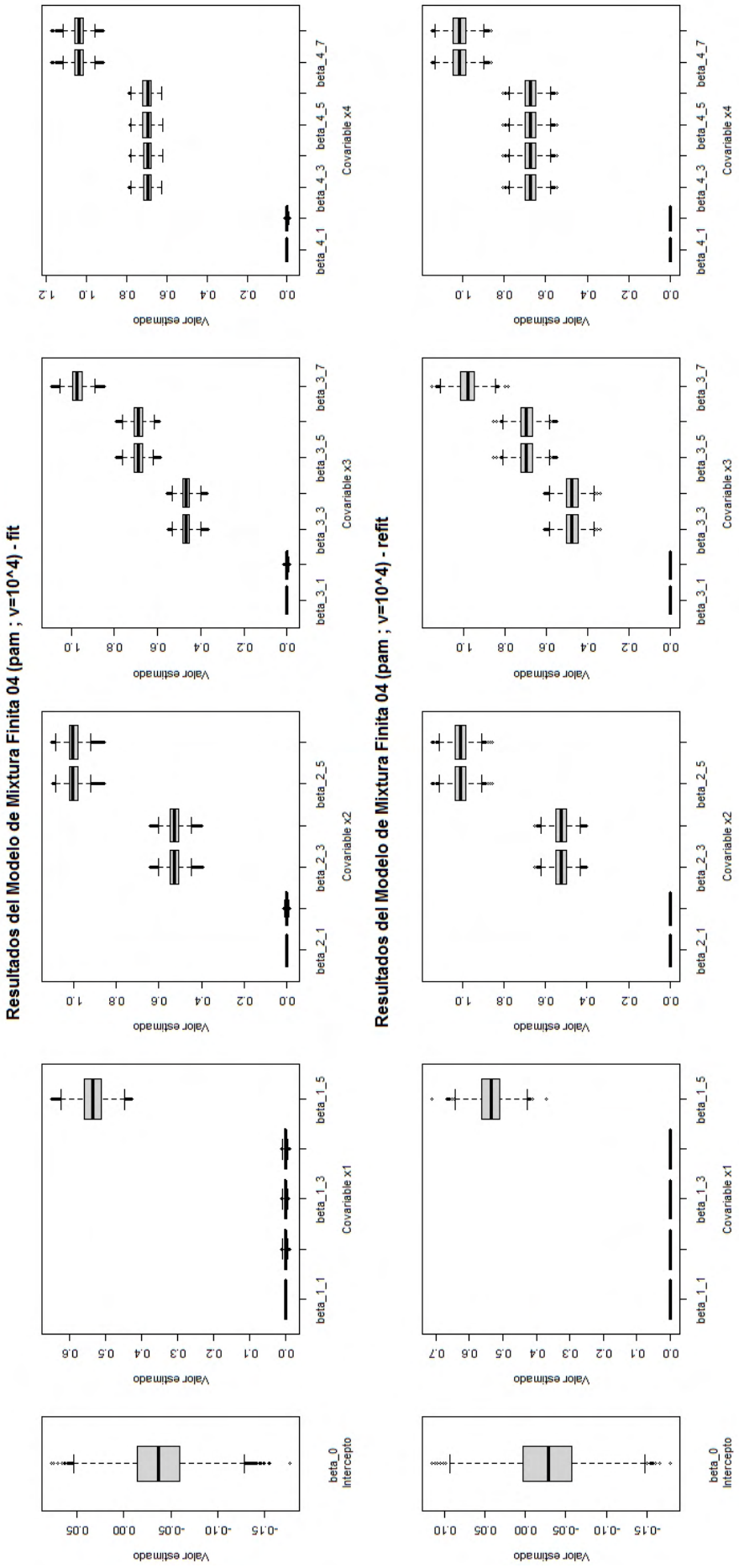

Figura 4.6: Resultados de la estimación y re-estimación de los efectos de nivel en el modelo de mixtura finita 04 (pam ; $v = 10^4$)

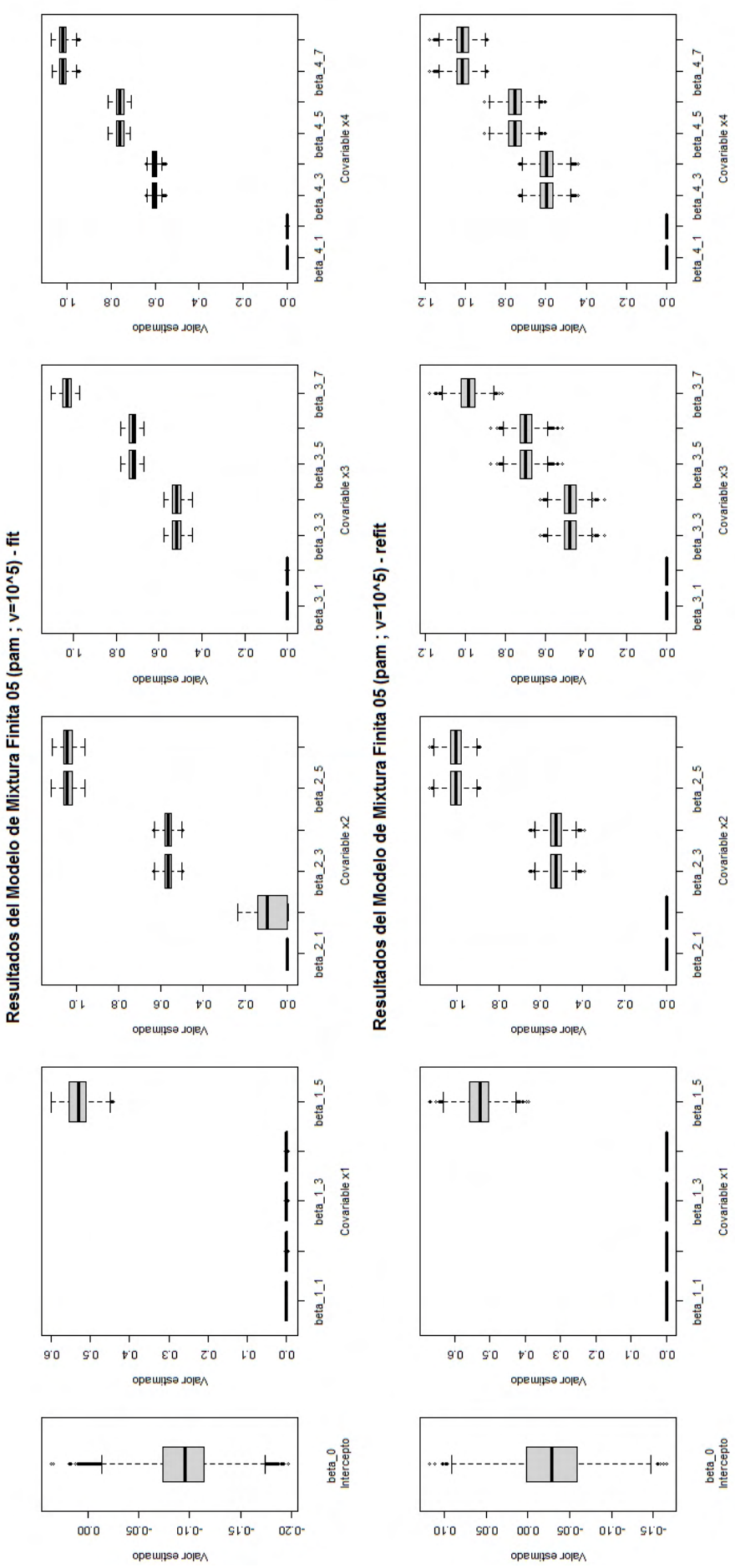

Figura 4.7: Resultados de la estimación y re-estimación de los efectos de nivel en el modelo de mixtura finita 05 (pam ; $v = 10^5$)

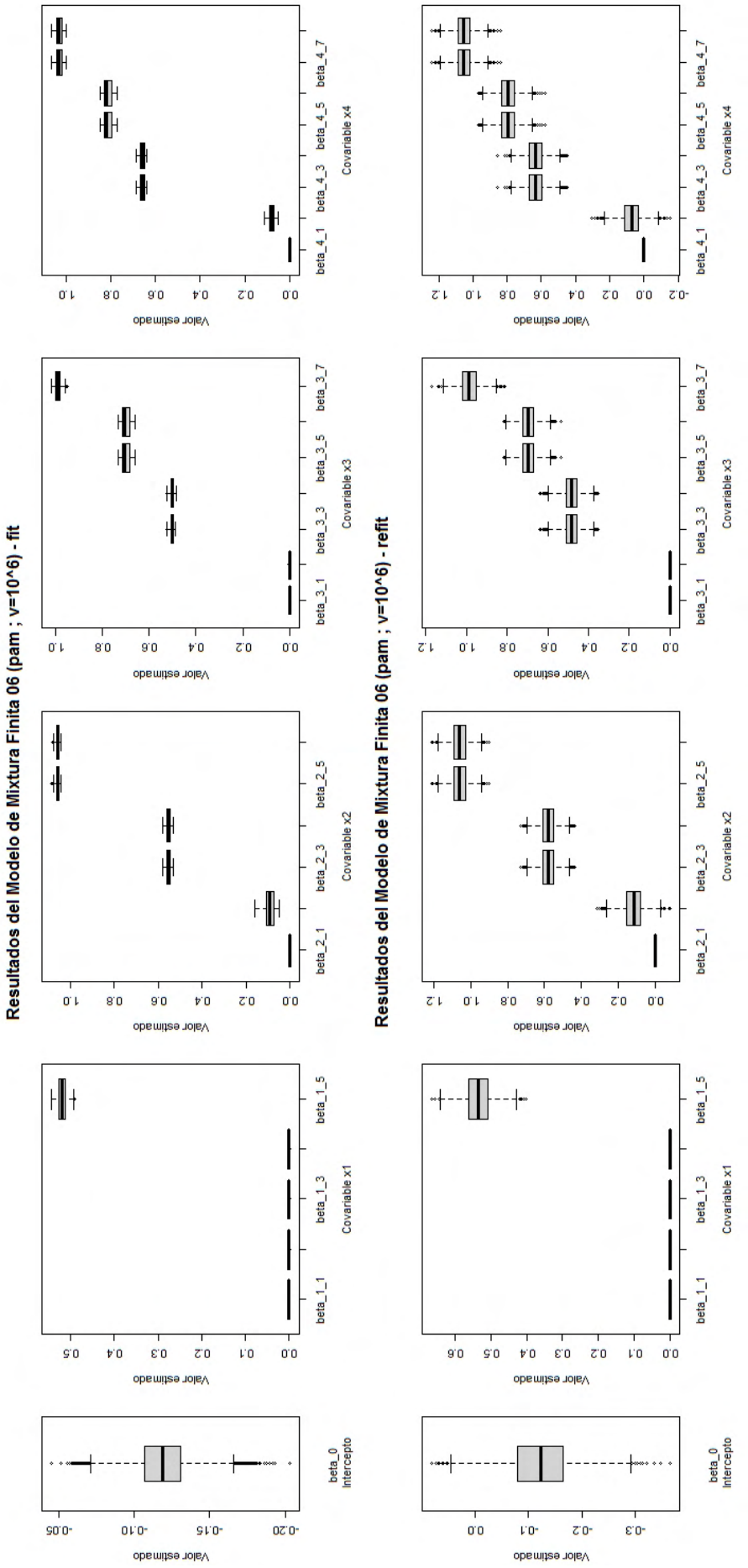

Figura 4.8: Resultados de la estimación y re-estimación de los efectos de nivel en el modelo de mixtura finita 06 (pam ; $v = 10^6$)

4.5. Criterios de Comparación

En esta sección, se evaluará si mediante los resultados del criterio de información de desvío (DIC) se obtiene que el mejor ajuste lo obtiene el modelo MF05 (pam ; $v = 10^5$), el cual presenta la misma cantidad de grupos de efectos de nivel. Además, se analizará si existe alguna relación entre los resultados del criterio de información del desvío (DIC) para cada escenario y los distintos valores de la variabilidad de los componentes de la mixtura de la distribución a priori.

En la Tabla 4.4, se aprecian los resultados del criterio de información de desvío (DIC) para cada escenario, resultando que el modelo M06 (pam ; $v = 10^6$) proporciona el mejor ajuste del modelo de regresión lineal.

Criterio	Modelo MF01	Modelo MF02	Modelo MF03	Modelo MF04	Modelo MF05	Modelo MF06
DIC	1,583.529	1,420.943	1,420.758	1,420.664	1,409.390	1,407.130

Tabla 4.4: Comparación del DIC de los modelos de regresión lineal, en la etapa de simulación

Finalmente, de los resultados obtenidos en el estudio de simulación, se puede concluir los siguientes puntos:

- El modelo MF01 (pam ; $v = 10$) presenta la forma más reducida del modelo de regresión lineal, sin embargo, el valor del DIC es mayor que el resto de modelos.
- El modelo MF06 (pam ; $v = 10^6$) presenta un mejor ajuste que los otros modelos de regresión lineal, debido a su menor valor en el DIC. Sin embargo, presenta la forma más extensa del modelo de regresión lineal, incluso mayor que el modelo simulado.
- El modelo MF05 (pam ; $v = 10^5$) presenta la fusión de efectos de nivel igual que el modelo simulado, y el DIC es el segundo menor valor de todos los modelos.
- Los modelos de regresión lineal más reducidos se obtienen cuando se trabaja con mayor variabilidad (menores valores de v) en los componentes de la mixtura de la distribución a priori.
- Los modelos de regresión lineal con mejor ajuste a los datos se obtienen cuando se trabaja con menor variabilidad (mayores valores de v) en los componentes de la mixtura de la distribución a priori.

Capítulo 5

Aplicación

El método de fusión de efectos de categorías se aplicará en un modelo de regresión lineal, que ha sido obtenido de un conjunto de datos recogidos de la "Encuesta Permanente de Empleo, Trimestre móvil Ene-Feb-Mar 2020". Esta encuesta es realizada por el Instituto Nacional de Estadística e Informática (INEI) en Lima Metropolitana y Callao. La base de datos de la encuesta ha sido publicada en julio del 2020 mediante la siguiente dirección electrónica: `http://iinei.inei.gob.pe/microdatos/consulta.asp?cmbencuesta=Encuesta+Permanente+de+Empleo+-+EPE&cmbanno=2020&cmbTrimestre=36`.

Asimismo, los códigos computacionales de la aplicación del método de fusión de efectos de categorías se encuentran en el Apéndice B.

5.1. Estudio de Caso

En el año 2001, el Instituto Nacional de Estadística e Informática (INEI) inició con la aplicación de la Encuesta Permanente de Empleo (EPE) en las áreas metropolitanas de Lima y Callao, con el objetivo de generar y monitorear indicadores referentes al empleo e ingresos en estas áreas metropolitanas de manera trimestral.

En la actualidad, la Encuesta Permanente de Empleo (EPE) se considera una investigación estadística continua, dado que se recopila información trimestral y de manera móvil. Es importante mencionar que debido a la pandemia causada por el COVID-19, la última Encuesta Permanente de Empleo (EPE) fue aplicada en los meses de enero, febrero y marzo del año 2020, recopilándose información de este mismo periodo de tiempo.

Por ese motivo, la presente investigación trabaja con la última Encuesta Permanente de Empleo (EPE), que fue realizada en el primer trimestre (enero, febrero, marzo) del 2020, en los 43 distritos de la provincia de Lima y los 6 distritos de la provincia del Callao.

5.1.1. Descripción del Caso

El diseño de la Encuesta Permanente de Empleo (EPE) de trimestre móvil permite analizar sobre los siguientes temas de investigación:

- Características de los miembros del hogar.
- Empleo e ingreso.
- Seguro de salud.
- Discapacidad.
- Etnicidad.
- Idioma o lengua materna que aprendió en su niñez.

En resumen, con la Encuesta Permanente de Empleo (EPE) se puede recopilar información sobre las categorías de trabajadores, la dinámica del mercado laboral, la evolución del empleo en el tiempo y los ingresos de los hogares, entre otros aspectos relevantes en el ámbito socio-económico.

Por lo mencionado, la Encuesta Permanente de Empleo (EPE) cobra una mayor importancia en el ámbito socio-económico, ya que permite investigar aspectos relevantes de la oferta laboral, como son las tasas de desempleo y subempleo existentes en Lima Metropolitana y Callao.

Con respecto al diseño muestral, para el año 2020, el marco muestral para la selección de la muestra de la Encuesta Permanente de Empleo (EPE), tiene como fuente principal la Cartografía, el Empadronamiento de Población y Vivienda 2012 - 2013 (SISFHO) y el Censo de Población y Vivienda 2017 (CPV-2017). Asimismo, el muestreo fue probabilístico y bietápico: sistemático con probabilidad proporcional al tamaño (PPT) en la primera etapa y sistemático simple de una muestra compacta en la segunda etapa; además, la muestra es estratificada de manera implícita, porque previamente a la selección, la población se ha dividido en estratos socio-económicos, con el objeto de mejorar su representatividad.

En ese sentido, el tamaño de muestra que se obtuvo para la Encuesta Permanente de Empleo (EPE) fue de 19,200 viviendas particulares en el año; y el tamaño de muestra para el primer trimestre móvil (enero, febrero, marzo) del 2020 fue de 4,800 viviendas, que serán visitadas en los 43 distritos de la provincia de Lima y los 6 distritos de la provincia del Callao.

Finalmente, la Encuesta Permanente de Empleo (EPE) se realiza mediante la visita mensual a 1,600 viviendas, distribuidas en 400 conglomerados, seleccionando 4 viviendas por conglomerado; y se entrevistan a las personas de 14 y más años de edad que viven en el hogar. Sin embargo, de las 4,800 viviendas programadas para el trimestre móvil (enero, febrero, marzo) del 2020, se registraron 3,916 viviendas como entrevistadas (completas e incompletas), obteniéndose una tasa de no respuesta del 9 %.

5.1.2. Descripción de los Datos

Para la formulación del modelo de regresión lineal se debe seleccionar una variable que será considerada como dependiente y otras variables que serán consideradas como independientes, las cuales se usarán para la estimación de la variable respuesta. Todas estas variables se obtendrán de la base de datos de la Encuesta Permanente de Empleo del Trimestre móvil Ene-Feb-Mar 2020, que cuenta con un total de 102 variables (columnas) y 14,806 observaciones (filas).

Luego de analizar la base de datos de la Encuesta Permanente de Empleo del Trimestre móvil Ene-Feb-Mar 2020, se seleccionaron 9 variables: 8 de ellas serán independientes y 1 será la dependiente. Cabe mencionar que, para la selección de las variables se ha tenido como referencia la aplicación realizada en Malsiner-Walli, Pauger y Wagner (2018), es por ello que la variable de respuesta es continua y todas las covariables seleccionadas son categóricas; sin embargo, es preciso mencionar que, el método de fusión de categorías, propuesto por Malsiner-Walli, Pauger y Wagner (2018), también admite que las covariables sean continuas. En la Figura 5.1 y Figura 5.2, se muestra el comportamiento de cada una de las variables seleccionadas.

Después de la selección de variables, ha sido necesario realizar la limpieza de los datos, mediante el reemplazo de los valores atípicos y la eliminación de los datos no válidos en todas las variables seleccionadas, resultando sólo 3,660 observaciones. En la Figura 5.3 y Figura 5.4, se muestra el comportamiento de cada una de las variables seleccionadas, después de la depuración.

A continuación, se describen con mayor detalle la variable respuesta y cada una de las 8 covariables, con sus respectivas categorías:

1. **Variable respuesta y: Ingreso Total Mensual (ingtot)**
   El "ingreso total mensual" (en soles) es una variable continua calculada con los datos recopilados de la encuesta. Dicha variable se encuentra en la columna "ingtot" de la base de datos, y se calcula mediante la suma de los ingresos de la actividad principal y las secundarias de las personas encuestadas, incluyendo sus remuneraciones monetarias y en especies.
   Después de la limpieza de los datos, la media del ingreso total mensual de una persona es de 1,893 soles y la mediana es de 1,413 soles; además, el cuartil 1 es de 1,000 soles y el cuartil 3 es de 2,165 soles.
   Como se puede apreciar en los histogramas de la Figura 5.1 y Figura 5.3, la distribución de los datos de la variable "ingreso total mensual" es asimétrica. Por tal motivo, para la aplicación del método de fusión de efectos, al igual que se realizó en la aplicación desarrollada en Malsiner-Walli, Pauger y Wagner (2018), se utilizará una transformación no lineal mediante la aplicación del logaritmo a todos los datos de la variable "ingreso total mensual", con el objetivo de obtener una distribución más simétrica.

Después de la transformación de la variable "ingreso total mensual", la media es 7.313, la mediana es 7.253, el cuartil 1 es 6.908 y el cuartil 3 es 7.680. Los resultados de la transformación se pueden evidenciar en la Figura 5.5, donde se muestra el histograma con una distribución más simétrica de los datos y la variación de los diagramas de cajas para cada variable; es preciso mencionar que, las frecuencias absolutas de las categorías de las covariables no varían después de la transformación, por ello, los diagramas de barras de las covariables que se muestran en la Figura 5.3 siguen siendo las mismas después de la transformación.

2. **Covariable x_1: Parentesco con el Jefe del Hogar (p103)**
   El "parentesco con el jefe del hogar" es una variable categórica nominal, que se encuentra en la columna "p103" de la base de datos y pertenece al tema de investigación "características de los miembros del hogar".
   El objetivo de la pregunta es conocer si la persona encuestada es el jefe(a) del hogar, o caso contrario, conocer el parentesco que tiene la persona con el jefe(a) del hogar. La variable cuenta originalmente con 11 categorías, las cuales se describen a continuación:

   - 1: Jefe(a)
   - 2: Esposo(a)/Compañero(a)
   - 3: Hijo(a)/Hijastro(a)
   - 4: Yerno/Nuera
   - 5: Padres/Suegros
   - 6: Otros parientes
   - 7: Trabajador(a) del hogar
   - 8: Pensionista
   - 9: Otros no parientes
   - 10: Nieto(a)
   - 11: Hermano(a)

   En la base de datos, la categoría 8 (pensionista) no cuenta con observaciones, teniendo un total de 10 categorías. Después de la limpieza de datos, sólo se cuenta con 9 categorías: 1, 2, 3, 4, 5, 6, 9, 10, 11; siendo las categorías 3 (hijo(a)/hijastro(a)) y 1 (jefe/jefa), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

3. **Covariable x_2: Rango de Edades (p108)**
   El "rango de edades" se ha considerado como una variable categórica ordinal, al igual que en Pauger, D. y Wagner, H. (2017), la cual ha sido trabajada a partir de los datos que se encuentran en la columna "p108" de la base de datos y que pertenecen al tema de investigación "características de los miembros del hogar".
   El objetivo de la pregunta es conocer la edad de la persona encuestada, y los rangos de edades permitirán analizar los datos de una manera más simplificada. Para un mejor análisis de la covariable, se ha dividido en 7 categorías, las cuales se describen a continuación:

- 1: Edades en el rango de [14,20[años
- 2: Edades en el rango de [20,30[años
- 3: Edades en el rango de [30,40[años
- 4: Edades en el rango de [40,50[años
- 5: Edades en el rango de [50,60[años
- 6: Edades en el rango de [60,70[años
- 7: Edades en el rango de [70,100] años

Después de la limpieza de datos, no se descartó ninguna categoría; siendo las categorías 2 (edades en el rango de [20,30[años) y 3 (edades en el rango de [30,40[años), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

4. **Covariable x_3: Nivel Educativo (p109)**
   El "nivel educativo" es una variable categórica ordinal, que se encuentra en la columna "p109" de la base de datos y pertenece al tema de investigación "características de los miembros del hogar".
   El objetivo de la pregunta es conocer el nivel educativo máximo alcanzado de la persona encuestada. La variable cuenta originalmente con 10 categorías, las cuales se describen a continuación:

   - 1: Sin nivel
   - 2: Inicial
   - 3: Primaria incompleta
   - 4: Primaria completa
   - 5: Secundaria incompleta
   - 6: Secundaria completa
   - 7: Superior no universitaria incompleta
   - 8: Superior no universitaria completa
   - 9: Superior universitaria incompleta
   - 10: Superior universitaria completa

   Después de la limpieza de datos, sólo se cuenta con 9 categorías: 1, 3, 4, 5, 6, 7, 8 9, 10; siendo las categorías 6 (secundaria completa) y 10 (superior universitaria completa), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

5. **Covariable x_4: Empleador (P206AA)**
   El "empleador" es una variable categórica nominal, que se encuentra en la columna "P206AA" de la base de datos y pertenece al tema de investigación "empleo e ingreso".
   El objetivo de la pregunta es conocer el tipo de empleador de la persona encuestada, referente a su actividad principal. La variable cuenta originalmente con 6 categorías, las cuales se describen a continuación:

- 1: Fuerzas Armadas, Policía Nacional del Perú (militares)
- 2: Administración pública
- 3: Empresa pública
- 4: Empresas especiales de servicios (service)
- 5: Empresa o patrono privado
- 6: Otro

En la base de datos, la categoría 6 (otro) no cuenta con observaciones, teniendo un total de 5 categorías. Después de la limpieza de datos, no se descartó ninguna categoría; siendo las categorías 5 (empresa o patrono privado) y 2 (administración pública), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

6. **Covariable x_5: Frecuencia de Pago (p210)**
   La "frecuencia de pago" es una variable categórica nominal, que se encuentra en la columna "p210" de la base de datos y pertenece al tema de investigación "empleo e ingreso".
   El objetivo de la pregunta es conocer la frecuencia de pago del empleador a la persona encuestada, referente a su actividad principal. La variable cuenta originalmente con 5 categorías, las cuales se describen a continuación:

   - 0: Practicante
   - 1: Diaria
   - 2: Semanal
   - 3: Quincenal
   - 4: Mensual

   Después de la limpieza de datos, sólo se cuenta con 4 categorías: 1, 2, 3, 4; siendo las categorías 4 (mensual) y 2 (semanal), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

7. **Covariable x_6: Seguro de Salud Afiliado (p222)**
   El "seguro de salud afiliado" es una variable categórica nominal, que se encuentra en la columna "p222" de la base de datos y pertenece al tema de investigación "seguro de salud".
   El objetivo de la pregunta es conocer si la persona encuestada se encuentra o no afiliada a algún tipo de seguro. La variable cuenta originalmente con 5 categorías, las cuales se describen a continuación:

   - 1: EsSalud (antes IPSS)
   - 2: Seguro privado de salud
   - 3: Ambos seguros
   - 4: Otro

- 5: No está afiliado

Después de la limpieza de datos, no se descartó ninguna categoría; siendo las categorías 1 (EsSalud) y 5 (no está afiliado), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

8. **Covariable x_7: Costumbres (P224)**
   Las "costumbres" es una variable categórica nominal, que se encuentra en la columna "P224" de la base de datos y pertenece al tema de investigación "etnicidad".
   El objetivo de la pregunta es conocer cómo se considera la persona encuestada, en base a sus costumbres y sus antepasados. La variable cuenta originalmente con 8 categorías, las cuales se describen a continuación:

   - 1: Quechua
   - 2: Aymara
   - 3: Nativo o indígena de la Amazonía
   - 4: Negro/Mulato/Zambo/Afroperuano
   - 5: Blanco
   - 6: Mestizo
   - 7: Otro
   - 8: No sabe

   Después de la limpieza de datos, no se descartó ninguna categoría; siendo las categorías 6 (mestizo) y 1 (quechua), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

9. **Covariable x_8: Idioma o Lengua Materna (P225)**
   El "idioma o lengua materna" es una variable categórica nominal, que se encuentra en la columna "P225" de la base de datos y pertenece al tema de investigación "idioma o lengua materna que aprendió en su niñez".
   El objetivo de la pregunta es conocer el idioma o la lengua materna que aprendió en su niñez la persona encuestada. La variable cuenta originalmente con 8 categorías, las cuales se describen a continuación:

   - 1: Quechua
   - 2: Aymara
   - 3: Otra lengua nativa
   - 4: Castellano
   - 5: Portugués
   - 6: Otra lengua extranjera
   - 7: Es sordo-mudo(a)/mudo(a)
   - 8: Lengua de señas peruanas

En la base de datos, la categoría 8 (lengua de señas peruanas) no cuenta con observaciones, teniendo un total de 7 categorías. Después de la limpieza de datos, no se descartó ninguna categoría; siendo las categorías 4 (castellano) y 1 (quechua), las que cuentan con mayor frecuencia absoluta; véase la Figura 5.3.

En base a algunos valores similares que se obtienen de la variable respuesta $\boldsymbol{y}$ con respecto a las categorías dentro de una misma covariable $\boldsymbol{x_j}$ (véase diagramas de cajas de la Figura 5.5), se puede tener una idea inicial de las posibles fusiones de los efectos de nivel resultantes de la aplicación del método propuesto por Malsiner-Walli,Pauger y Wagner (2018).

5.2. Estimación por Inferencia Bayesiana

La fusión de efectos de nivel y la selección de covariables categóricas tienen como objetivo una representación reducida de un modelo de regresión lineal. Una manera práctica de conseguirlo es usando el paquete effectFusion del software R, que se basa en un enfoque bayesiano usando una mixtura finita de normales como distribución a priori, que se desarrolló en el Capítulo 3 y se corroboró en el estudio de simulación desarrollado en el Capítulo 4.

5.2.1. Formulación del Modelo de Regresión

Con lo explicado en las secciones anteriores, se puede definir un modelo de regresión lineal con las 8 covariables categóricas (2 ordinales y 6 nominales) y un total de 54 categorías, incluyendo el nivel de la línea base de cada covariable. Tomando como referencia la ecuación (2.18), el intercepto β_0, un error $e \sim N(0;\sigma^2)$ y las categorías existentes en cada covariable seleccionada de la Encuesta Permanente de Empleo, se formula el siguiente modelo de regresión lineal:

$$
\begin{aligned}
y_i \;=\; & \beta_0 + \\
& (x_{i1,1}\beta_{1,1} + x_{i1,2}\beta_{1,2} + x_{i1,3}\beta_{1,3} + x_{i1,4}\beta_{1,4} + x_{i1,5}\beta_{1,5} + x_{i1,6}\beta_{1,6} + x_{i1,9}\beta_{1,9} + \\
& x_{i1,10}\beta_{1,10} + x_{i1,11}\beta_{1,11}) + \\
& (x_{i2,1}\beta_{2,1} + x_{i2,2}\beta_{2,2} + x_{i2,3}\beta_{2,3} + x_{i2,4}\beta_{2,4} + x_{i2,5}\beta_{2,5} + x_{i2,6}\beta_{2,6} + x_{i2,7}\beta_{2,7}) + \\
& (x_{i3,1}\beta_{3,1} + x_{i3,3}\beta_{3,3} + x_{i3,4}\beta_{3,4} + x_{i3,5}\beta_{3,5} + x_{i3,6}\beta_{3,6} + x_{i3,7}\beta_{3,7} + x_{i3,8}\beta_{3,8} + \\
& x_{i3,9}\beta_{3,9} + x_{i3,10}\beta_{3,10}) + \\
& (x_{i4,1}\beta_{4,1} + x_{i4,2}\beta_{4,2} + x_{i4,3}\beta_{4,3} + x_{i4,4}\beta_{4,4} + x_{i4,5}\beta_{4,5}) + \\
& (x_{i5,1}\beta_{5,1} + x_{i5,2}\beta_{5,2} + x_{i5,3}\beta_{5,3} + x_{i5,4}\beta_{5,4}) + \\
& (x_{i6,1}\beta_{6,1} + x_{i6,2}\beta_{6,2} + x_{i6,3}\beta_{6,3} + x_{i6,4}\beta_{6,4} + x_{i6,5}\beta_{6,5}) + \\
& (x_{i7,1}\beta_{7,1} + x_{i7,2}\beta_{7,2} + x_{i7,3}\beta_{7,3} + x_{i7,4}\beta_{7,4} + x_{i7,5}\beta_{7,5} + x_{i7,6}\beta_{7,6} + x_{i7,7}\beta_{7,7} + x_{i7,8}\beta_{7,8}) + \\
& (x_{i8,1}\beta_{8,1} + x_{i8,2}\beta_{8,2} + x_{i8,3}\beta_{8,3} + x_{i8,4}\beta_{8,4} + x_{i8,5}\beta_{8,5} + x_{i8,6}\beta_{8,6} + x_{i8,7}\beta_{8,7}) + \\
& \epsilon_i
\end{aligned}
\tag{5.1}
$$

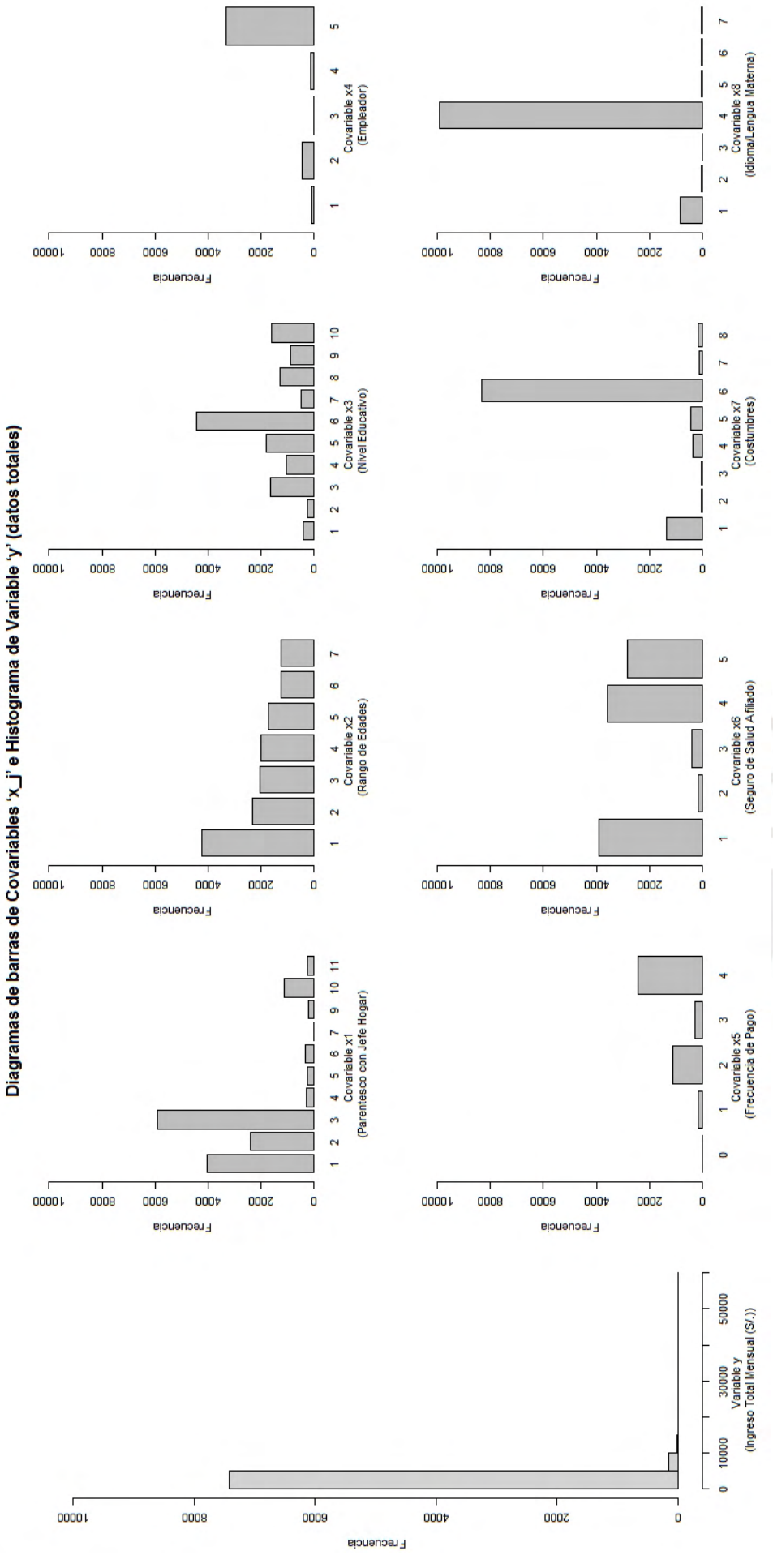

Figura 5.1: Diagramas de barras de de covariables e histograma de variable respuesta (observaciones totales: 14,806)

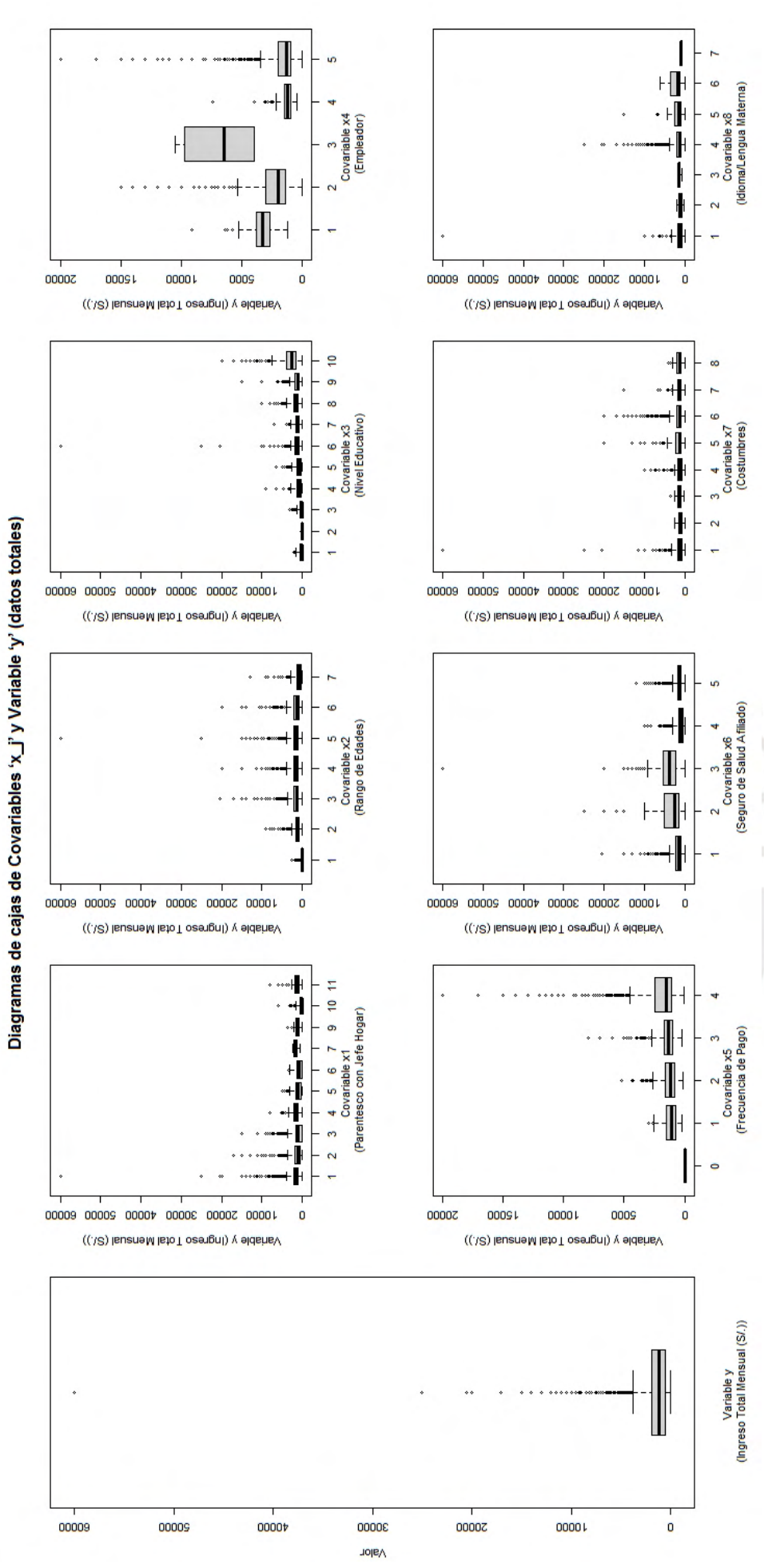

Figura 5.2: Diagramas de cajas de covariables y variable respuesta (observaciones totales: 14,806)

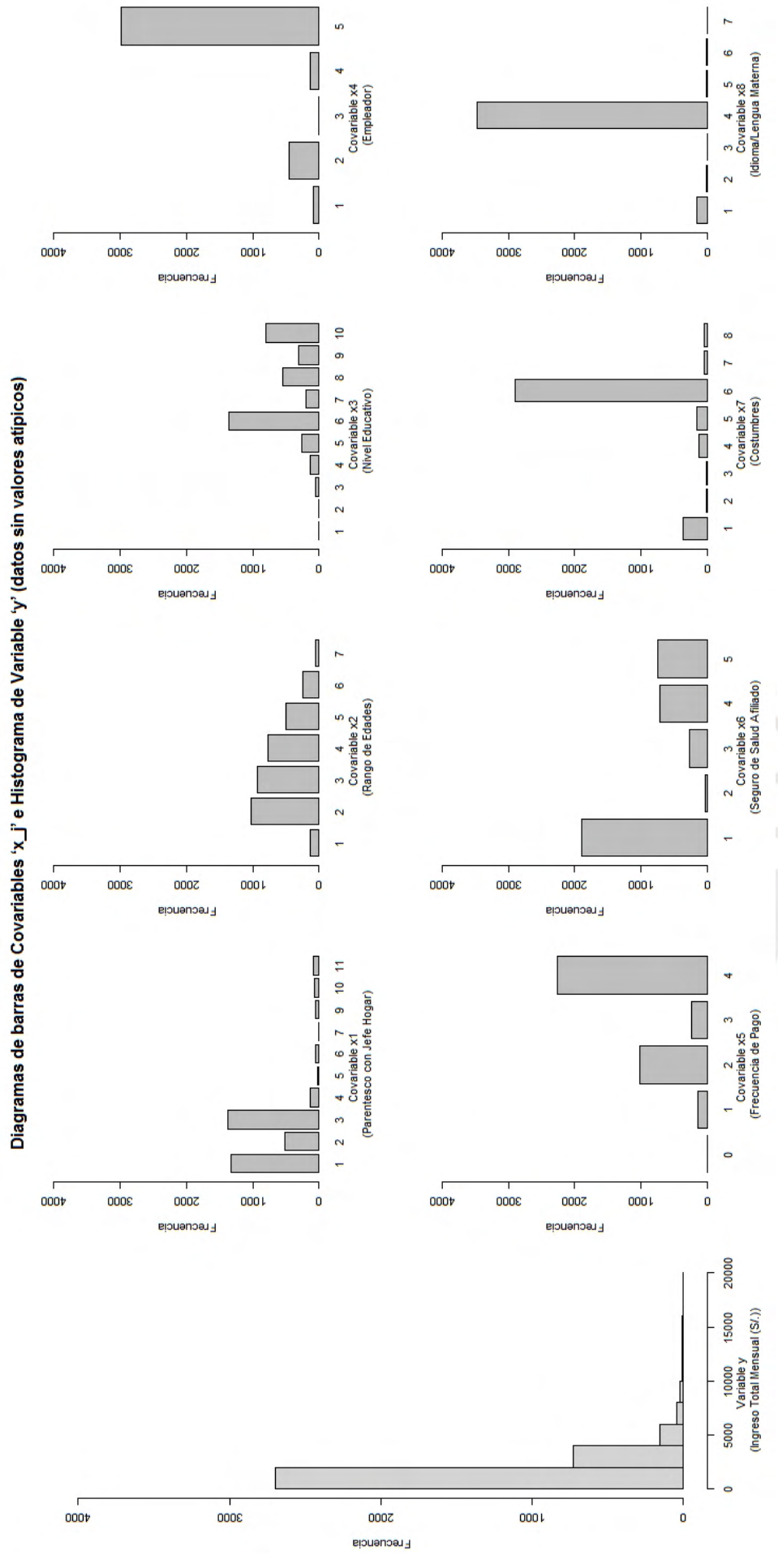

Figura 5.3: Diagramas de barras de covariables e histograma de variable respuesta (observaciones válidas: 3,660 observaciones)

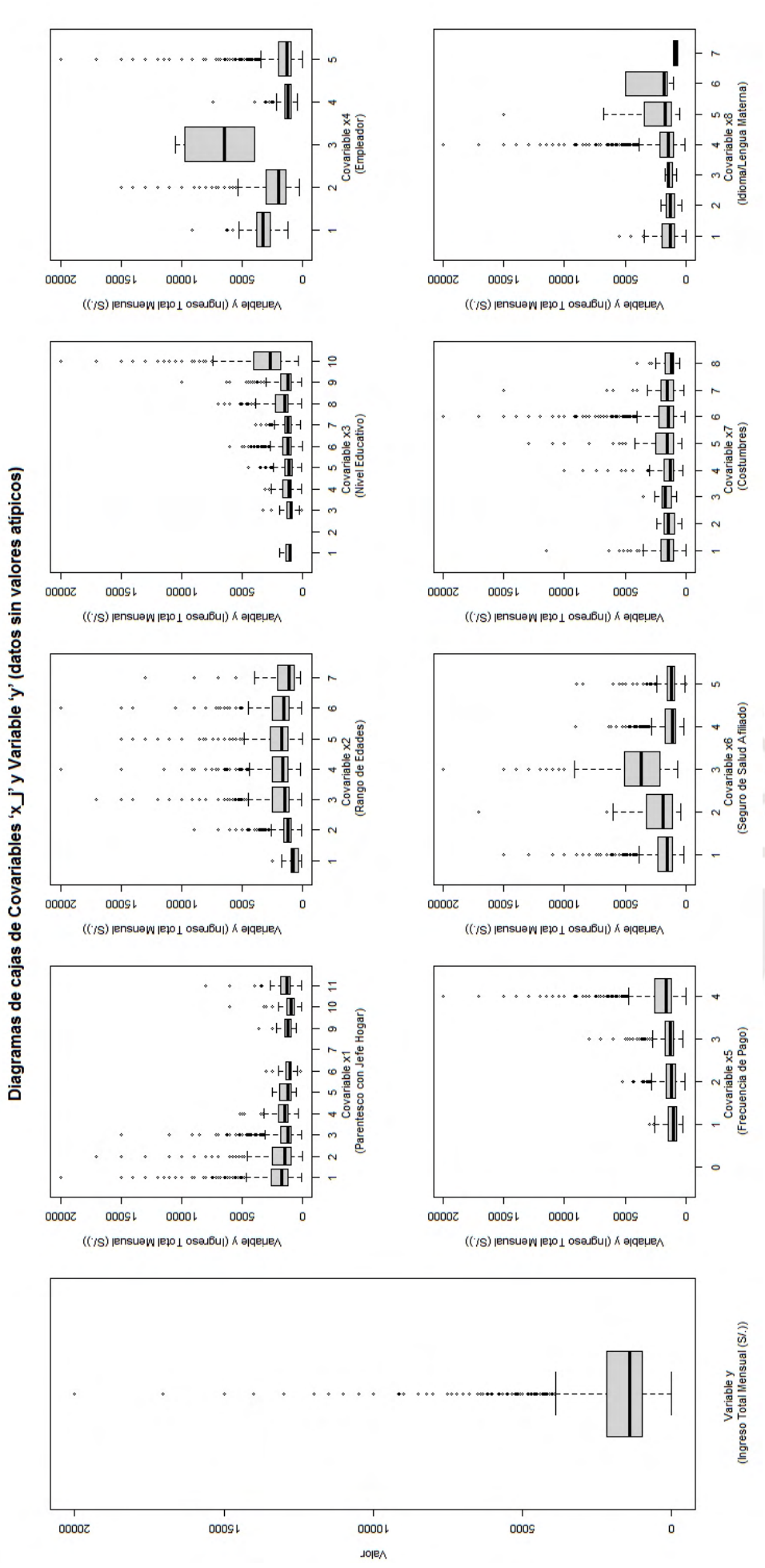

Figura 5.4: Diagramas de cajas de covariables y variable respuesta (observaciones válidas: 3,660 observaciones)

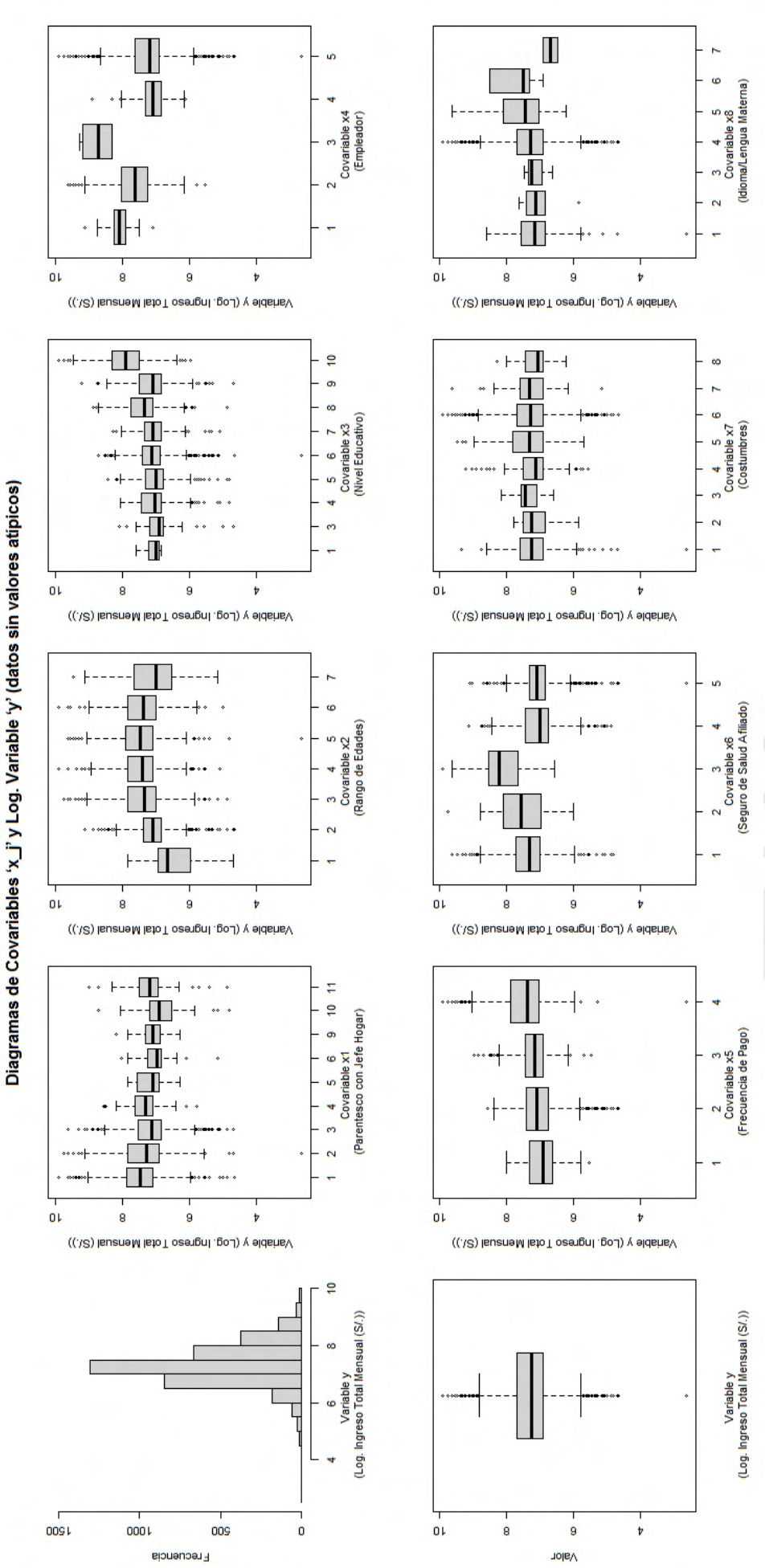

Figura 5.5: Diagramas de covariables y variable respuesta transformada (observaciones válidas: 3,660 observaciones)

5.2.2. Consideraciones de la Aplicación

Para la aplicación del método de fusión de efectos de nivel mediante la distribución a priori descrita en la sección 4.1, se definen los valores de los hiperparámetros similares a los usados en el estudio de simulación:

- En la ecuación (4.2), se define al intercepto β_0 con una distribución normal con media 0 y varianza ψ_0. Para la aplicación, se define la varianza del intercepto, como: $\psi_0 = 0{,}01$.
- En la ecuación (4.3), se define a la varianza σ^2 con una distribución gamma inversa con los parámetros s_0 y S_0. Para la aplicación, se define los parámetros, como: $s_0 = 0$ y $S_0 = 0$. Con lo cual, se obtiene una distribución impropia para σ^2.
- En la ecuación (4.5), se define a los pesos de las distribuciones de la mixtura $\boldsymbol{\eta_j}$ con una distribución de Dirichlet con parámetro e_0. Para la aplicación, se define el parámetro, como: $e_0 = 0{,}01$. Con lo cual, se está induciendo a que los componentes de la mixtura se concentren en unos pocos grupos.
- En la ecuación (4.8), se define a la varianza de los componentes de la mixtura $\boldsymbol{\psi_j}$ con una distribución gamma inversa con los parámetros g_0 y G_0. Para la aplicación, se define los parámetros, como: $g_0 = 100$ y G_{j0} será evaluado en 6 escenarios distintos, ya que v tomará los valores: 10, 10^2, 10^3, 10^4, 10^5, 10^6. Con $g_0 = 100$, se obtiene desviaciones pequeñas en los componentes de la mixtura. Con respecto al parámetro G_{j0}, se puede afirmar del estudio de simulación que, los modelos de regresión lineal son más reducidos cuando v toma valores pequeños.

Cabe recordar que, del estudio de simulación también se puede afirmar que ψ controla la precisión de las particiones de los efectos de nivel. En otras palabras, la reducción del modelo de regresión depende de ψ, y de ahí la importancia de evaluar los 6 escenarios antes mencionados.

En el presente capítulo, al igual que en el estudio de simulación, se usarán los mismos argumentos del paquete effectFusion del software R: method = 'FinMix', modelSelection = 'pam', mcmc = c(M=30000,burnin=15000). De la misma forma, se estimarán los 7 modelos de regresión lineal (escenarios) que se describieron en la sección 4.1:

1. Modelo General.
2. Modelo MF01 (pam ; $v = 10$).
3. Modelo MF02 (pam ; $v = 10^2$).
4. Modelo MF03 (pam ; $v = 10^3$).
5. Modelo MF04 (pam ; $v = 10^4$).
6. Modelo MF05 (pam ; $v = 10^5$).

7. Modelo MF06 (pam ; $v = 10^6$).

Es preciso mencionar que el paquete effectFusion, considera al primer efecto de nivel de cada covariable como línea base, es decir: $\beta_{1,1}$ (jefe(a)), $\beta_{2,1}$ (edades en el rango de [14;20[años), $\beta_{3,1}$ (sin nivel), $\beta_{4,1}$ (Fuerzas Armadas, Policía Nacional del Perú (militares)), $\beta_{5,1}$ (diaria), $\beta_{6,1}$ (EsSalud), $\beta_{7,1}$ (quechua) y $\beta_{8,1}$ (quechua) serían los niveles de línea base.

Asimismo, es importante mencionar que, el método de fusión de efectos propuesto Malsiner-Walli,Pauger y Wagner (2018), considera a todas las covariables como nominales. En otras palabras, para el método de fusión de efectos, no existe un orden en las categorías de las covariables, lo cual representa una desventaja del método.

5.3. Resultados de Aplicación

Después de describir el modelo de regresión lineal con las 8 covariables categóricas y definir los valores de las prioris en el método de fusión de efectos; en la presente sección, se muestran los resultados de la aplicación del paquete effectFusion en los datos de la Encuesta Permanente de Empleo del Trimestre móvil Ene-Feb-Mar 2020.

El valor promedio de las estimaciones y re-estimaciones de los β_{jk} de los modelos de regresión lineal (escenarios), así como los intervalos de credibilidad al 95 %, se muestran en la Tabla 5.1 y Tabla 5.2, respectivamente. Al igual que se explicó en el estudio de simulación, la re-estimación de los β_{jk} considera el modelo seleccionado en cada escenario. Asimismo, en la re-estimación de los modelos de regresión lineal, los efectos de nivel fusionados presentan un único valor, y los efectos de nivel con valor de 0 indican que han sido fusionados con el nivel de la línea base de cada covariable. A modo de resumen, se diseña la Tabla 5.3, donde se muestran los efectos de nivel fusionados en los modelos de regresión lineal (escenarios).

Adicionalmente, en las Figuras 5.6, 5.7, 5.9, 5.11, 5.13, 5.15, 5.17 se muestran de manera gráfica las 30,000 iteraciones realizadas en el MCMC, para la estimación de los β_{jk} en cada modelo (escenario). Asimismo, en las Figuras 5.8, 5.10, 5.12, 5.14, 5.16, 5.18 se observan las re-estimaciones de los β_{jk} en base al modelo seleccionado, y se confirma lo anteriormente mencionado, ya que los efectos fusionados presentan una misma distribución y los efectos fusionados a la línea base toman el valor de 0.

Los resultados de la aplicación, permiten la comparación de los 7 modelos de regresión lineal (escenarios) en base a las fusiones de los efectos de nivel resultantes. A continuación, una mayor explicación de los resultados de cada modelo generado:

1. **Modelo General:**
   El modelo general presenta valores distintos para cada efecto de nivel de las covariables $\boldsymbol{x_j}$, es decir, no presenta fusiones de efectos, y por ende, es el modelo más extenso. A pesar de ello, en base a los valores estimados de β_{jk} del modelo que se muestran en la

Figura 5.6, se puede mencionar lo siguiente:

- La covariable $\boldsymbol{x_1}$, referente al parentesco con el jefe del hogar, presenta 9 efectos de nivel. Los efectos de las categorías 3 (hijo(a)/hijastro(a), 4 (yerno/nuera), 5 (padres/suegros) y 6 (otros parientes) aparentan estar en un mismo grupo; al igual que los efectos de las categorías 2 (esposo(a)/compañero(a)) y 9 (otros no parientes) que también podrían estar en un mismo grupo; y el resto de efectos de nivel se estarían fusionando con el nivel de la línea base (jefe(a)).
- La covariable $\boldsymbol{x_2}$, referente al rango de edades, presenta 7 efectos de nivel. El efecto de la categoría 7 (edades en el rango de [70,100] años) aparenta ser el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se estarían fusionando con el nivel de la línea base (edades en el rango de [14;20[años).
- La covariable $\boldsymbol{x_3}$, referente al nivel educativo, presenta 9 efectos de nivel. Los efectos de las categorías 6 (secundaria completa) y 7 (superior no universitaria incompleta) aparentan estar en un mismo grupos; el resto de efectos de nivel estarían formando un mismo grupo; y ningún efecto de nivel se estaría fusionando con el nivel de la línea base (sin nivel).
- La covariable $\boldsymbol{x_4}$, referente al empleador, presenta 5 efectos de nivel. El efecto de la categoría 3 (empresa pública) aparenta ser el único elemento de un grupo; el resto de efectos de nivel estarían formando un mismo grupo; y ningún efecto de nivel se estaría fusionando con el nivel de la línea base (Fuerzas Armadas, Policía Nacional del Perú (militares)).
- La covariable $\boldsymbol{x_5}$, referente a la frecuencia de pago, presenta 4 efectos de nivel. El efecto de la categoría 2 (semanal) aparenta ser el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se estarían fusionando con el nivel de la línea base (diaria).
- La covariable $\boldsymbol{x_6}$, referente al seguro de salud afiliado, presenta 5 efectos de nivel. El efecto de la categoría 3 (ambos seguros) aparenta ser el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se estarían fusionando con el nivel de la línea base (EsSalud).
- La covariable $\boldsymbol{x_7}$, referente a las costumbres, presenta 8 efectos de nivel. El efecto de la categoría 2 (aymara) aparenta ser el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se estarían fusionando con el nivel de la línea base (quechua).
- La covariable $\boldsymbol{x_8}$, referente al idioma o lengua materna, presenta 7 efectos de nivel. Los efectos de las categorías 2 (aymara), 6 (otra lengua extranjera) y 7 (es sordo-mudo(a)/mudo(a)) aparentan ser los elementos del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se estarían fusionando con el nivel de la línea base (quechua).

2. **Modelo MF01 (pam ; $v = 10$):**

El modelo de mixtura finita 01 es el más reducido, considerando sólo los modelos de

mixtura finita. Los valores estimados y re-estimados de β_{jk} del modelo se muestran en la Figura 5.7 y la Figura 5.8, respectivamente. A partir de ello, se puede resumir lo siguiente:

- La covariable $\boldsymbol{x_1}$, referente al parentesco con el jefe del hogar, presenta 3 grupos de efectos de nivel. Los efectos de las categorías 3 (hijo(a)/hijastro(a), 4 (yerno/nuera), 5 (padres/suegros) y 6 (otros parientes) se han fusionado en un mismo grupo; los efectos de las categorías 2 (esposo(a)/compañero(a)) y 9 (otros no parientes) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (jefe(a)).
- La covariable $\boldsymbol{x_2}$, referente al rango de edades, presenta 2 grupos de efectos de nivel. El efecto de la categoría 7 (edades en el rango de [70,100] años) es el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (edades en el rango de [14;20[años).
- La covariable $\boldsymbol{x_3}$, referente al nivel educativo, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (sin nivel).
- La covariable $\boldsymbol{x_4}$, referente al empleador, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (Fuerzas Armadas, Policía Nacional del Perú (militares)).
- La covariable $\boldsymbol{x_5}$, referente a la frecuencia de pago, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (diaria).
- La covariable $\boldsymbol{x_6}$, referente al seguro de salud afiliado, presenta 2 grupos de efectos de nivel. El efecto de la categoría 3 (ambos seguros) es el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (EsSalud).
- La covariable $\boldsymbol{x_7}$, referente a las costumbres, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (quechua).
- La covariable $\boldsymbol{x_8}$, referente al idioma o lengua materna, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (quechua).

En total se consiguen 17 grupos de efectos de nivel, considerando los niveles de la línea base de cada covariable $\boldsymbol{x_j}$.

3. **Modelo MF02 (pam ; $v = 10^2$):**
   Los valores estimados y re-estimados de β_{jk} del modelo de mixtura finita 02 se muestran en la Figura 5.9 y la Figura 5.10, respectivamente. A partir de ello, se puede resumir lo siguiente:

- La covariable $\boldsymbol{x_1}$, referente al parentesco con el jefe del hogar, presenta 3 grupos de efectos de nivel. Los efectos de las categorías 3 (hijo(a)/hijastro(a), 4 (yerno/nuera), 5 (padres/suegros) y 6 (otros parientes) se han fusionado en un mismo grupo; los efectos de las categorías 2 (esposo(a)/compañero(a)) y 9 (otros no parientes) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (jefe(a)).
- La covariable $\boldsymbol{x_2}$, referente al rango de edades, presenta 2 grupos de efectos de nivel. El efecto de la categoría 7 (edades en el rango de [70,100] años) es el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (edades en el rango de [14;20[años).
- La covariable $\boldsymbol{x_3}$, referente al nivel educativo, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (sin nivel).
- La covariable $\boldsymbol{x_4}$, referente al empleador, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (Fuerzas Armadas, Policía Nacional del Perú (militares)).
- La covariable $\boldsymbol{x_5}$, referente a la frecuencia de pago, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (diaria).
- La covariable $\boldsymbol{x_6}$, referente al seguro de salud afiliado, presenta 3 grupos de efectos de nivel. El efecto de la categoría 3 (ambos seguros) es el único elemento de un grupo; los efectos de las categorías 4 (otro) y 5 (no está afiliado) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (EsSalud).
- La covariable $\boldsymbol{x_7}$, referente a las costumbres, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (quechua).
- La covariable $\boldsymbol{x_8}$, referente al idioma o lengua materna, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (quechua).

En total se consiguen 18 grupos de efectos de nivel, considerando los niveles de la línea base de cada covariable $\boldsymbol{x_j}$.

4. **Modelo MF03 (pam ; $v = 10^3$):**
Los valores estimados y re-estimados de β_{jk} del modelo de mixtura finita 03 se muestran en la Figura 5.11 y la Figura 5.12, respectivamente. A partir de ello, se puede resumir lo siguiente:

   - La covariable $\boldsymbol{x_1}$, referente al parentesco con el jefe del hogar, presenta 4 grupos de efectos de nivel. Los efectos de las categorías 3 (hijo(a)/hijastro(a)), 4 (yerno/nuera), 5 (padres/suegros) y 6 (otros parientes) se han fusionado en un

mismo grupo; el efecto de la categoría 2 (esposo(a)/compañero(a)) es el único elemento de otro grupo; los efectos de las categorías 10 (nieto(a)) y 11 (hermano(a)) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (jefe(a)).

- La covariable $\boldsymbol{x_2}$, referente al rango de edades, presenta 4 grupos de efectos de nivel. El efecto de la categoría 7 (edades en el rango de [70,100] años) es el único elemento de un grupo; el efecto de la categoría 5 (edades en el rango de [50,60[años) es el único elemento de un grupo; los efectos de las categorías 2 (edades en el rango de [20,30[años), 3 (edades en el rango de [30,40[años) y 4 (edades en el rango de [40,50[años) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (edades en el rango de [14;20[años).
- La covariable $\boldsymbol{x_3}$, referente al nivel educativo, presenta 2 grupos de efectos de nivel. El efecto de la categoría 5 (secundaria incompleta) se ha fusionado con el nivel de la línea base (sin nivel); y el resto de efectos de nivel se han fusionado en un mismo grupo.
- La covariable $\boldsymbol{x_4}$, referente al empleador, presenta 2 grupos de efectos de nivel. El efecto de la categoría 3 (empresa pública) se ha fusionado con el nivel de la línea base (Fuerzas Armadas, Policía Nacional del Perú (militares)); y el resto de efectos de nivel se han fusionado en un mismo grupo.
- La covariable $\boldsymbol{x_5}$, referente a la frecuencia de pago, presenta 2 grupos de efectos de nivel. Todos los efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base (diaria).
- La covariable $\boldsymbol{x_6}$, referente al seguro de salud afiliado, presenta 3 grupos de efectos de nivel. El efecto de la categoría 3 (ambos seguros) es el único elemento de un grupo; los efectos de las categorías 4 (otro) y 5 (no está afiliado) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (EsSalud).
- La covariable $\boldsymbol{x_7}$, referente a las costumbres, presenta 2 grupos de efectos de nivel. El efecto de la categoría 2 (aymara) es el único elemento del grupo distinto al nivel de la línea base; el resto de efectos de nivel se han fusionado con el nivel de la línea base (quechua).
- La covariable $\boldsymbol{x_8}$, referente al idioma o lengua materna, presenta 3 grupos de efectos de nivel. Los efectos de las categorías 2 (aymara) y 6 (otra lengua extranjera) se han fusionado en un mismo grupo; el efecto de la categoría 7 (es sordo-mudo(a)/mudo(a)) es el único elemento de otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (quechua).

En total se consiguen 22 grupos de efectos de nivel, considerando los niveles de la línea base de cada covariable $\boldsymbol{x_j}$.

5. **Modelo MF04 (pam ; $v = 10^4$):**
   Los valores estimados y re-estimados de β_{jk} del modelo de mixtura finita 04 se muestran

en la Figura 5.13 y la Figura 5.14, respectivamente. A partir de ello, se puede resumir lo siguiente:

- La covariable $\boldsymbol{x_1}$, referente al parentesco con el jefe del hogar, presenta 4 grupos de efectos de nivel. Los efectos de las categorías 3 (hijo(a)/hijastro(a)), 4 (yerno/nuera) y 5 (padres/suegros) se han fusionado en un mismo grupo; los efectos de las categorías 2 (esposo(a)/compañero(a)) y 6 (otros parientes) se han fusionado en otro grupo; el efecto de la categoría 9 (otros no parientes) es el único elemento de un grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (jefe(a)).
- La covariable $\boldsymbol{x_2}$, referente al rango de edades, presenta 4 grupos de efectos de nivel. El efecto de la categoría 7 (edades en el rango de [70,100] años) es el único elemento de un grupo; el efecto de la categoría 5 (edades en el rango de [50,60[años) es el único elemento de un grupo; los efectos de las categorías 2 (edades en el rango de [20,30[años), 3 (edades en el rango de [30,40[años), 4 (edades en el rango de [40,50[años) y 6 (edades en el rango de [60,70[años) se han fusionado en otro grupo; y ningún efecto de nivel se fusionó con el nivel de la línea base (edades en el rango de [14;20[años).
- La covariable $\boldsymbol{x_3}$, referente al nivel educativo, presenta 4 grupos de efectos de nivel. El efecto de la categoría 5 (secundaria incompleta) es el único elemento de un grupo; los efectos de las categorías 6 (secundaria completa) y 7 (superior no universitaria incompleta) se han fusionado en otro grupo; el resto de categorías se ha fusionado en un mismo grupo; y ningún efecto de nivel se fusionó con el nivel de la línea base (sin nivel).
- La covariable $\boldsymbol{x_4}$, referente al empleador, presenta 3 grupos de efectos de nivel. Los efectos de las categorías 2 (administración pública) y 3 (empresa pública) se han fusionado en un mismo grupo; los efectos de las categorías 4 (empresas especiales de servicios (service)) y 5 (empresa o patrono privado) se han fusionado en un mismo grupo; y ningún efecto de nivel se fusionó con el nivel de la línea base (Fuerzas Armadas, Policía Nacional del Perú (militares)).
- La covariable $\boldsymbol{x_5}$, referente a la frecuencia de pago, presenta 2 grupos de efectos de nivel. El efecto de la categoría 2 (semanal) es el único elemento del grupo distinto al nivel de la línea base; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (diaria).
- La covariable $\boldsymbol{x_6}$, referente al seguro de salud afiliado, presenta 3 grupos de efectos de nivel. El efecto de la categoría 3 (ambos seguros) es el único elemento de un grupo; los efectos de las categorías 4 (otro) y 5 (no está afiliado) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (EsSalud).
- La covariable $\boldsymbol{x_7}$, referente a las costumbres, presenta 4 grupos de efectos de nivel. El efecto de la categoría 2 (aymara) es el único elemento de un grupo; el efecto de la categoría 8 (no sabe) es el único elemento de un grupo; el efecto de la categoría

3 (nativo o indígena de la Amazonía) se ha fusionado con el nivel de la línea base (quechua); y el resto de efectos de nivel se han fusionado en un grupo distinto al nivel de la línea base.

- La covariable $\boldsymbol{x_8}$, referente al idioma o lengua materna, presenta 2 grupos de efectos de nivel. Los efectos de las categorías 2 (aymara), 6 (otra lengua extranjera) y 7 (es sordo-mudo(a)/mudo(a)) se han fusionado en un mismo grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (quechua).

En total se consiguen 26 grupos de efectos de nivel, considerando los niveles de la línea base de cada covariable $\boldsymbol{x_j}$.

6. **Modelo MF05 (pam ; $v = 10^5$):**
Los valores estimados y re-estimados de β_{jk} del modelo de mixtura finita 05 se muestran en la Figura 5.15 y la Figura 5.16, respectivamente. A partir de ello, se puede resumir lo siguiente:

   - La covariable $\boldsymbol{x_1}$, referente al parentesco con el jefe del hogar, presenta 4 grupos de efectos de nivel. Los efectos de las categorías 3 (hijo(a)/hijastro(a)), 4 (yerno/nuera), 5 (padres/suegros) y 6 (otros parientes) se han fusionado en un mismo grupo; los efectos de las categorías 2 (esposo(a)/compañero(a)) y 9 (otros no parientes) se han fusionado en otro grupo; el efecto de la categoría 10 (nieto (a)) es el único elemento de un grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (jefe(a)).
   - La covariable $\boldsymbol{x_2}$, referente al rango de edades, presenta 4 grupos de efectos de nivel. El efecto de la categoría 7 (edades en el rango de [70,100] años) es el único elemento de un grupo; el efecto de la categoría 5 (edades en el rango de [50,60[años) es el único elemento de un grupo; los efectos de las categorías 3 (edades en el rango de [30,40[años), 4 (edades en el rango de [40,50[años) y 6 (edades en el rango de [60,70[años) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (edades en el rango de [14;20[años).
   - La covariable $\boldsymbol{x_3}$, referente al nivel educativo, presenta 6 grupos de efectos de nivel. Los efectos de las categorías 3 (primaria incompleta), 4 (primaria completa) y 9 (superior universitaria incompleta) se han fusionado en un mismo grupo; los efectos de las categorías 8 (superior no universitaria completa) y 10 (superior universitaria completa) se han fusionado en otro grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (sin nivel).
   - La covariable $\boldsymbol{x_4}$, referente al empleador, presenta 4 grupos de efectos de nivel. Los efectos de las categorías 2 (administración pública) y 5 (empresa o patrono privado) se han fusionado en un mismo grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (Fuerzas Armadas, Policía Nacional del Perú (militares)).

- La covariable $\boldsymbol{x_5}$, referente a la frecuencia de pago, presenta 3 grupos de efectos de nivel. El efecto de la categoría 2 (semanal) es el único elemento de un grupo; el efecto de la categoría 3 (quincenal) es el único elemento de un grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (diaria).
- La covariable $\boldsymbol{x_6}$, referente al seguro de salud afiliado, presenta 3 grupos de efectos de nivel. El efecto de la categoría 3 (ambos seguros) es el único elemento de un grupo; los efectos de las categorías 4 (otro) y 5 (no está afiliado) se han fusionado en otro grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (EsSalud).
- La covariable $\boldsymbol{x_7}$, referente a las costumbres, presenta 4 grupos de efectos de nivel. El efecto de la categoría 2 (aymara) es el único elemento de un grupo; el efecto de la categoría 8 (no sabe) es el único elemento de un grupo; los efectos de las categorías 5 (blanco) y 7 (otro) se han fusionado en un grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (quechua).
- La covariable $\boldsymbol{x_8}$, referente al idioma o lengua materna, presenta 4 grupos de efectos de nivel. Los efectos de las categorías 2 (aymara), 6 (otra lengua extranjera) y 7 (es sordo-mudo(a)/mudo(a)) se han fusionado en un mismo grupo; los efectos de las categorías 4 (castellano) y 5 (portugués) se han fusionado en otro grupo; el efecto de la categoría 3 (otra lengua nativa) es el único elemento de un grupo; y ningún efecto de nivel se fusionó con el nivel de la línea base (quechua).

En total se consiguen 32 grupos de efectos de nivel, considerando los niveles de la línea base de cada covariable $\boldsymbol{x_j}$.

7. **Modelo MF06 (pam ; $v = 10^6$):**
   El modelo de mixtura finita 06 es el más extenso, considerando sólo los modelos de mixtura finita. Los valores estimados y re-estimados de β_{jk} del modelo se muestran en la Figura 5.17 y la Figura 5.18, respectivamente. A partir de ello, se puede resumir lo siguiente:

   - La covariable $\boldsymbol{x_1}$, referente al parentesco con el jefe del hogar, presenta 7 grupos de efectos de nivel. Los efectos de las categorías 3 (hijo(a)/hijastro(a)), 4 (yerno/nuera) y 5 (padres/suegros) se han fusionado en un mismo grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (jefe(a)).
   - La covariable $\boldsymbol{x_2}$, referente al rango de edades, presenta 5 grupos de efectos de nivel. Los efectos de las categorías 3 (edades en el rango de [30,40[años), 4 (edades en el rango de [40,50[años) y 6 (edades en el rango de [60,70[años) se han fusionado en un mismo grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (edades en el rango de [14;20[años).
   - La covariable $\boldsymbol{x_3}$, referente al nivel educativo, presenta 8 grupos de efectos de nivel. Los efectos de las categorías 3 (primaria incompleta) y 4 (primaria completa) se

han fusionado en un mismo grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (sin nivel).

- La covariable $\boldsymbol{x_4}$, referente al empleador, presenta 4 grupos de efectos de nivel. Los efectos de las categorías 4 (empresas especiales de servicios (service)) y 5 (empresa o patrono privado) se han fusionado en un mismo grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (Fuerzas Armadas, Policía Nacional del Perú (militares)).
- La covariable $\boldsymbol{x_5}$, referente a la frecuencia de pago, presenta 3 grupos de efectos de nivel. El efecto de la categoría 2 (semanal) es el único elemento de un grupo; el efecto de la categoría 3 (quincenal) es el único elemento de un grupo; y el resto de efectos de nivel se han fusionado con el nivel de la línea base (diaria).
- La covariable $\boldsymbol{x_6}$, referente al seguro de salud afiliado, presenta 4 grupos de efectos de nivel. Los efectos de las categorías 4 (otro) y 5 (no está afiliado) se han fusionado en un mismo grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (EsSalud).
- La covariable $\boldsymbol{x_7}$, referente a las costumbres, presenta 6 grupos de efectos de nivel. Los efectos de las categorías 4 (negro/mulato/zambo/afroperuano) y 6 (mestizo) se han fusionado en un grupo; el efecto de la categoría 3 (nativo o indígena de la Amazonía) se ha fusionado con el nivel de la línea base (quechua); y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base.
- La covariable $\boldsymbol{x_8}$, referente al idioma o lengua materna, presenta 5 grupos de efectos de nivel. Los efectos de las categorías 2 (aymara), 6 (otra lengua extranjera) y 7 (es sordo-mudo(a)/mudo(a)) se han fusionado en un mismo grupo; y el resto de categorías son elementos únicos de grupos distintos al nivel de la línea base (quechua).

En total se consiguen 42 grupos de efectos de nivel, considerando los niveles de la línea base de cada covariable $\boldsymbol{x_j}$.

Al igual que en el estudio de simulación, con los resultados obtenidos en el presente capítulo, se puede afirmar lo siguiente: 1) mayores variabilidades (menores valores de v) en los componentes de la mixtura de la distribución a priori permiten conseguir modelos de regresión lineal más reducidos; y 2) la técnica de agrupamiento 'pam' no permite que una covariable se excluya del modelo de regresión lineal.

		Valores estimados de los efectos de nivel																				
		Modelo General			Modelo MF01			Modelo MF02			Modelo MF03			Modelo MF04			Modelo MF05			Modelo MF06		
Covariable	Efecto	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)
	β_0	7.18156	7.56400	7.96828	7.46738	7.61600	7.76474	7.53564	7.67500	7.81000	7.65389	7.77200	7.89776	7.44954	7.52300	7.59460	7.43681	7.51700	7.58557	7.52981	7.55700	7.58299
$\boldsymbol{x_1}$	$\beta_{1,2}$	0.29470	0.39100	0.48392	0.34451	0.41070	0.48081	0.35033	0.41580	0.48664	0.29497	0.35860	0.42356	0.38480	0.43320	0.47993	0.35636	0.38390	0.41020	0.37843	0.39680	0.41572
	$\beta_{1,3}$	0.46945	0.56550	0.66665	0.51520	0.58140	0.64786	0.51665	0.58390	0.64866	0.45239	0.51780	0.58399	0.53474	0.57430	0.61560	0.53745	0.55600	0.58603	0.55523	0.56520	0.57215
	$\beta_{1,4}$	0.45868	0.55830	0.66427	0.50582	0.57290	0.64265	0.51153	0.57880	0.64825	0.45182	0.51720	0.58367	0.53469	0.57410	0.61578	0.53749	0.55600	0.58610	0.55527	0.56520	0.57219
	$\beta_{1,5}$	0.47405	0.57820	0.68942	0.51043	0.58300	0.65455	0.51098	0.58120	0.65296	0.45275	0.51780	0.58504	0.53454	0.57420	0.61558	0.53757	0.55600	0.58619	0.55572	0.56600	0.57216
	$\beta_{1,6}$	0.38080	0.49880	0.61692	0.43985	0.52850	0.61195	0.40516	0.54180	0.63903	0.37453	0.49620	0.59022	0.38540	0.43350	0.48049	0.53660	0.55160	0.58264	0.49793	0.50940	0.51964
	$\beta_{1,9}$	0.10331	0.27296	0.44365	0.26509	0.38650	0.52194	0.33477	0.41570	0.55500	-0.00930	0.08711	0.37343	0.15662	0.23850	0.34731	0.30735	0.32800	0.36049	0.28281	0.29130	0.30122
	$\beta_{1,10}$	-0.51073	-0.14920	0.22122	-0.21858	-0.11403	-0.00551	-0.17107	-0.03638	0.03215	-0.14755	-0.02883	0.01400	-0.00459	-0.00008	0.00429	-0.32060	-0.26650	-0.22684	-0.15237	-0.12496	-0.10321
	$\beta_{1,11}$	-0.40031	-0.03776	0.30505	-0.17170	-0.08132	0.00702	-0.09365	-0.03254	0.03218	-0.14798	-0.02917	0.01339	-0.00436	0.00003	0.00446	-0.00139	-0.00002	0.00139	-0.08494	-0.06489	-0.04048
$\boldsymbol{x_2}$	$\beta_{2,2}$	-0.31063	0.04337	0.38391	-0.09487	-0.01803	0.05964	-0.05960	-0.01241	0.04056	-0.14687	-0.02879	0.01297	0.10399	0.14739	0.20142	-0.00138	0.00000	0.00141	0.04319	0.05452	0.06827
	$\beta_{2,3}$	-0.26139	0.09000	0.42359	-0.03832	0.02543	0.08963	-0.03432	0.01446	0.05929	-0.13829	-0.02376	0.01802	0.10388	0.14779	0.20075	0.03340	0.05881	0.08773	0.09313	0.10192	0.11115
	$\beta_{2,4}$	-0.27730	0.07821	0.42118	-0.06790	0.01408	0.09586	-0.04419	0.00345	0.05179	-0.14416	-0.02637	0.01516	0.10378	0.14759	0.20113	0.03334	0.05880	0.08766	0.09310	0.10191	0.11114
	$\beta_{2,5}$	-0.13957	0.21211	0.55003	0.07429	0.14260	0.21410	0.03630	0.12496	0.20381	-0.00623	0.07823	0.18744	0.22881	0.27870	0.34291	0.17041	0.20450	0.24068	0.21739	0.22780	0.23853
	$\beta_{2,6}$	-0.24214	0.10690	0.44837	-0.03671	0.04048	0.11452	-0.03760	0.01941	0.07624	-0.01321	0.00109	0.01557	0.10437	0.14778	0.20144	0.03336	0.05882	0.08769	0.09303	0.10192	0.11105
	$\beta_{2,7}$	0.26855	0.61020	0.95604	0.48741	0.55690	0.62678	0.49178	0.55790	0.61982	0.43482	0.54470	0.61981	0.63425	0.68920	0.74693	0.59486	0.62850	0.65922	0.58430	0.60500	0.63990
$\boldsymbol{x_3}$	$\beta_{3,3}$	-0.22383	-0.17220	-0.12049	-0.20051	-0.16220	-0.12224	-0.18197	-0.15106	-0.11850	-0.19456	-0.16370	-0.13428	-0.20541	-0.18990	-0.17175	-0.17918	-0.16550	-0.14372	-0.16813	-0.16260	-0.15766
	$\beta_{3,4}$	-0.21124	-0.16556	-0.11940	-0.19341	-0.15714	-0.11947	-0.18062	-0.14872	-0.11604	-0.19311	-0.16330	-0.13278	-0.20519	-0.19000	-0.17167	-0.17924	-0.16550	-0.14380	-0.16811	-0.16260	-0.15766
	$\beta_{3,5}$	-0.15011	-0.05555	0.03233	-0.19052	-0.12356	-0.02133	-0.18039	-0.14474	-0.11008	-0.00540	-0.00045	0.00478	-0.13940	-0.09594	-0.05623	-0.10687	-0.08038	-0.05089	-0.06408	-0.05593	-0.04266
	$\beta_{3,6}$	-0.54974	-0.29070	-0.02689	-0.21941	-0.15634	-0.09913	-0.18449	-0.14964	-0.11660	-0.19340	-0.16370	-0.13226	-0.37671	-0.34100	-0.29362	-0.34128	-0.31550	-0.29923	-0.29583	-0.28980	-0.27998
	$\beta_{3,7}$	-0.37273	-0.24471	-0.11005	-0.21903	-0.16512	-0.11286	-0.18524	-0.15037	-0.11716	-0.19359	-0.16370	-0.13290	-0.37714	-0.34100	-0.29402	-0.30559	-0.27100	-0.24300	-0.24881	-0.24580	-0.24228
	$\beta_{3,8}$	-0.24864	-0.12245	0.00957	-0.20629	-0.14902	-0.09455	-0.18300	-0.14844	-0.11515	-0.19383	-0.16380	-0.13240	-0.20609	-0.19000	-0.17250	-0.13103	-0.11800	-0.10319	-0.12472	-0.11960	-0.10942
	$\beta_{3,9}$	-0.33028	-0.20936	-0.08554	-0.21346	-0.16080	-0.10751	-0.18315	-0.14967	-0.11595	-0.19376	-0.16370	-0.13288	-0.20748	-0.19040	-0.17193	-0.17922	-0.16750	-0.15691	-0.23188	-0.22640	-0.22085
	$\beta_{3,10}$	-0.24565	-0.13771	-0.03121	-0.20316	-0.15068	-0.09868	-0.18279	-0.14843	-0.11507	-0.19315	-0.16340	-0.13264	-0.20595	-0.19000	-0.17238	-0.13085	-0.11800	-0.10302	-0.14823	-0.14390	-0.13994
$\boldsymbol{x_4}$	$\beta_{4,2}$	-0.89022	-0.76650	-0.63912	-0.89641	-0.79330	-0.69446	-0.93510	-0.83480	-0.73214	-0.95777	-0.84540	-0.72452	-0.85490	-0.72460	-0.63877	-0.84826	-0.81610	-0.77252	-0.77870	-0.76660	-0.75050
	$\beta_{4,3}$	-0.93949	-0.42750	0.06923	-1.02941	-0.85250	-0.66710	-1.02163	-0.88870	-0.73674	-0.01720	-0.00055	0.01541	-0.74703	-0.46920	-0.15547	-0.53930	-0.42160	-0.35168	-0.49022	-0.46410	-0.44509
	$\beta_{4,4}$	-1.16389	-1.02260	-0.87357	-1.12360	-1.00830	-0.89161	-1.07030	-0.96400	-0.85841	-1.01439	-0.90620	-0.79801	-0.92223	-0.85240	-0.79451	-0.99751	-0.94060	-0.88942	-1.05385	-1.01520	-0.94173
	$\beta_{4,5}$	-1.02511	-0.90890	-0.78673	-1.02959	-0.93440	-0.83959	-1.03091	-0.94260	-0.85447	-1.00975	-0.90470	-0.80177	-0.91778	-0.85120	-0.78902	-0.84889	-0.81630	-0.77307	-0.95705	-0.93080	-0.90654
$\boldsymbol{x_5}$	$\beta_{5,2}$	-0.12396	-0.03900	0.05106	-0.04152	-0.00818	0.01273	-0.02788	-0.00286	0.00609	-0.00156	0.00013	0.00145	-0.06025	-0.04555	-0.02928	-0.04077	-0.03400	-0.02966	-0.04404	-0.04238	-0.04025
	$\beta_{5,3}$	-0.12131	-0.01373	0.09158	-0.03326	-0.00379	0.01707	-0.02434	-0.00216	0.00716	-0.00149	0.00020	0.00149	-0.00040	0.00000	0.00042	-0.02508	-0.02012	-0.01437	-0.01189	-0.01089	-0.00985
	$\beta_{5,4}$	-0.09053	0.00191	0.09505	-0.02850	-0.00005	0.01987	-0.01065	-0.00074	0.00806	-0.00132	0.00024	0.00156	-0.00041	0.00000	0.00041	-0.00013	0.00000	0.00013	-0.00004	0.00000	0.00004
$\boldsymbol{x_6}$	$\beta_{6,2}$	-0.02862	0.14279	0.31021	-0.08289	0.07417	0.23612	-0.18051	-0.00578	0.08320	-0.01880	0.00214	0.02298	-0.00654	0.00022	0.00685	-0.00216	0.00044	0.00217	0.09131	0.12347	0.16151
	$\beta_{6,3}$	0.44823	0.51580	0.58232	0.45741	0.52300	0.58827	0.46568	0.52130	0.57488	0.45842	0.50700	0.55968	0.46790	0.51940	0.57095	0.43723	0.47870	0.51895	0.50525	0.52720	0.54994
	$\beta_{6,4}$	-0.29027	-0.23790	-0.18650	-0.28885	-0.24360	-0.19677	-0.29489	-0.25250	-0.20977	-0.27335	-0.23730	-0.20104	-0.22678	-0.19750	-0.16903	-0.23302	-0.20440	-0.17605	-0.27561	-0.24510	-0.20315
	$\beta_{6,5}$	-0.20776	-0.16046	-0.11141	-0.21147	-0.16670	-0.12209	-0.22808	-0.18765	-0.14665	-0.25160	-0.21580	-0.18144	-0.22378	-0.19500	-0.16586	-0.23206	-0.20420	-0.17521	-0.22234	-0.17980	-0.13596
$\boldsymbol{x_7}$	$\beta_{7,2}$	-0.15126	0.18237	0.50844	-0.04406	0.00247	0.04903	-0.01510	0.00080	0.01564	-0.00671	0.07697	0.28524	-0.00176	0.04974	0.10648	0.16550	0.18590	0.20246	0.16230	0.17180	0.18306
	$\beta_{7,3}$	-0.26384	-0.00818	0.25567	-0.04496	0.00162	0.04821	-0.01428	0.00057	0.01518	-0.00448	-0.00002	0.00443	-0.00141	0.00000	0.00140	-0.00044	0.00000	0.00044	-0.00014	0.00000	0.00014
	$\beta_{7,4}$	-0.08705	0.01475	0.12287	-0.04057	0.00197	0.04420	-0.01539	0.00073	0.01460	-0.00441	0.00000	0.00458	-0.00233	0.00271	0.03274	-0.00043	0.00000	0.00044	0.01002	0.01983	0.02950
	$\beta_{7,5}$	-0.07350	0.02839	0.12944	-0.03709	0.00459	0.04630	-0.01480	0.00096	0.01511	-0.00439	0.00003	0.00449	-0.00226	0.00258	0.03282	0.01992	0.03975	0.05731	0.03404	0.04144	0.04871
	$\beta_{7,6}$	-0.04997	0.01504	0.07712	-0.02481	0.00820	0.04167	-0.01130	0.00343	0.01765	-0.00401	0.00030	0.00484	-0.00226	0.00259	0.03268	-0.00044	0.00000	0.00043	0.01006	0.01983	0.02953
	$\beta_{7,7}$	-0.12132	0.05105	0.21523	-0.04249	0.00344	0.04872	-0.01489	0.00127	0.01620	-0.00444	0.00003	0.00445	-0.00217	0.00257	0.03290	0.02324	0.04350	0.05908	0.05801	0.06065	0.06343
	$\beta_{7,8}$	-0.20096	-0.04243	0.11568	-0.04765	-0.00259	0.04253	-0.01578	0.00026	0.01455	-0.00460	-0.00007	0.00441	-0.11912	-0.07365	-0.03744	-0.06093	-0.04883	-0.03148	-0.04793	-0.04331	-0.03865
$\boldsymbol{x_8}$	$\beta_{8,2}$	-0.77743	-0.33290	0.13764	-0.14800	-0.02171	0.11360	-0.05393	-0.00042	0.04818	-0.56990	-0.34060	-0.15756	-0.48048	-0.37160	-0.26022	-0.38496	-0.35300	-0.32329	-0.37639	-0.35940	-0.34910
	$\beta_{8,3}$	-0.64944	-0.07147	0.49837	-0.13901	-0.00131	0.13473	-0.05662	0.00366	0.05362	-0.01383	0.00002	0.01368	-0.00431	-0.00003	0.00432	-0.22297	-0.17760	-0.11938	-0.09837	-0.06767	-0.04961
	$\beta_{8,4}$	-0.02688	0.06659	0.16072	0.01032	0.08024	0.14641	-0.02040	0.03561	0.09609	-0.01052	0.00345	0.01677	-0.00403	0.00021	0.00457	0.03851	0.08095	0.14143	0.06630	0.08430	0.10118
	$\beta_{8,5}$	-0.15280	0.11617	0.38286	-0.09367	0.03111	0.15530	-0.05066	0.00690	0.05488	-0.01338	0.00054	0.01439	-0.00432	0.00002	0.00427	0.03844	0.08094	0.14128	0.11623	0.12430	0.13305
	$\beta_{8,6}$	-0.75225	-0.33960	0.08149	-0.16340	-0.02839	0.10415	-0.05672	0.00011	0.04909	-0.57423	-0.34382	-0.16230	-0.47972	-0.37160	-0.25911	-0.38499	-0.35300	-0.32322	-0.37635	-0.35940	-0.34899
	$\beta_{8,7}$	-1.04934	-0.35720	0.32459	-0.14668	-0.01013	0.13133	-0.05119	0.00177	0.04930	-0.43592	-0.15946	0.01731	-0.47989	-0.37170	-0.25934	-0.38521	-0.35300	-0.32345	-0.37655	-0.35940	-0.34921

Tabla 5.1: Resultados de los valores estimados de los efectos de nivel del modelo general y los 6 modelos de mixtura finita, en la etapa de aplicación

		Valores re-estimados de los efectos de nivel																	
		Modelo MF01			**Modelo MF02**			**Modelo MF03**			**Modelo MF04**			**Modelo MF05**			**Modelo MF06**		
Covariable	Efecto	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)	Mín. IC (95 %)	Promedio	Máx. IC (95 %)
	β_0	6.92327	7.10900	7.29438	7.37795	7.57400	7.76218	7.60213	7.78500	7.95885	7.32027	7.48900	7.67033	7.33800	7.50200	7.67582	7.14179	7.50500	7.89535
$\boldsymbol{x_1}$	$\beta_{1,2}$	0.41455	0.50960	0.60057	0.34774	0.44520	0.53372	0.25425	0.33440	0.41352	0.34320	0.43210	0.52169	0.30359	0.39310	0.48611	0.31330	0.40160	0.50216
	$\beta_{1,3}$	0.63800	0.72830	0.82254	0.54373	0.63230	0.72682	0.42165	0.50470	0.57797	0.48181	0.57700	0.66327	0.48998	0.58060	0.67375	0.48401	0.57270	0.67084
	$\beta_{1,4}$	0.63800	0.72830	0.82254	0.54373	0.63230	0.72682	0.42165	0.50470	0.57797	0.48181	0.57700	0.66327	0.48998	0.58060	0.67375	0.48401	0.57270	0.67084
	$\beta_{1,5}$	0.63800	0.72830	0.82254	0.54373	0.63230	0.72682	0.42165	0.50470	0.57797	0.48181	0.57700	0.66327	0.48998	0.58060	0.67375	0.48401	0.57270	0.67084
	$\beta_{1,6}$	0.63800	0.72830	0.82254	0.54373	0.63230	0.72682	0.42165	0.50470	0.57797	0.34320	0.43210	0.52169	0.48998	0.58060	0.67375	0.38975	0.50280	0.61697
	$\beta_{1,9}$	0.41455	0.50960	0.60057	0.34774	0.44520	0.53372	0.00000	0.00000	0.00000	0.13686	0.29292	0.45777	0.30359	0.39310	0.48611	0.11457	0.28281	0.44805
	$\beta_{1,10}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.12524	-0.06238	-0.00459	0.00000	0.00000	0.00000	-0.31875	-0.18270	-0.03257	-0.50258	-0.15880	0.23242
	$\beta_{1,11}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.12524	-0.06238	-0.00459	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.38059	-0.03745	0.32691
$\boldsymbol{x_2}$	$\beta_{2,2}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.12524	-0.06238	-0.00459	0.08086	0.15751	0.22883	0.00000	0.00000	0.00000	-0.33623	0.04504	0.36023
	$\beta_{2,3}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.12524	-0.06238	-0.00459	0.08086	0.15751	0.22883	0.02140	0.07922	0.13153	-0.24847	0.09688	0.42689
	$\beta_{2,4}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.12524	-0.06238	-0.00459	0.08086	0.15751	0.22883	0.02140	0.07922	0.13153	-0.24847	0.09688	0.42689
	$\beta_{2,5}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.04163	0.11242	0.18290	0.20121	0.28470	0.36595	0.14046	0.20809	0.27623	-0.12557	0.22120	0.55694
	$\beta_{2,6}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.08086	0.15751	0.22883	0.02140	0.07922	0.13153	-0.24847	0.09688	0.42689
	$\beta_{2,7}$	0.54613	0.59100	0.63203	0.49679	0.53660	0.58028	0.47381	0.54310	0.60854	0.61322	0.69210	0.77955	0.56261	0.62500	0.69203	0.27457	0.62234	0.95896
$\boldsymbol{x_3}$	$\beta_{3,3}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	-0.19085	-0.15747	-0.11950	-0.21639	-0.18040	-0.14456	-0.19104	-0.15517	-0.11670	-0.20553	-0.16800	-0.12659
	$\beta_{3,4}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	-0.19085	-0.15747	-0.11950	-0.21639	-0.18040	-0.14456	-0.19104	-0.15517	-0.11670	-0.20553	-0.16800	-0.12659
	$\beta_{3,5}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	0.00000	0.00000	0.00000	-0.15633	-0.06981	0.02441	-0.13592	-0.04904	0.04243	-0.13818	-0.05785	0.03491
	$\beta_{3,6}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	-0.19085	-0.15747	-0.11950	-0.37969	-0.26143	-0.14564	-0.54372	-0.28460	-0.03332	-0.52064	-0.27140	0.00219
	$\beta_{3,7}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	-0.19085	-0.15747	-0.11950	-0.37969	-0.26143	-0.14564	-0.37568	-0.24383	-0.11257	-0.38213	-0.24528	-0.12415
	$\beta_{3,8}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	-0.19085	-0.15747	-0.11950	-0.21639	-0.18040	-0.14456	-0.20052	-0.11548	-0.03325	-0.26031	-0.12951	0.00212
	$\beta_{3,9}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	-0.19085	-0.15747	-0.11950	-0.21639	-0.18040	-0.14456	-0.19104	-0.15517	-0.11670	-0.32282	-0.20835	-0.08264
	$\beta_{3,10}$	-0.17636	-0.13846	-0.10209	-0.17274	-0.13667	-0.09775	-0.19085	-0.15747	-0.11950	-0.21639	-0.18040	-0.14456	-0.20052	-0.11548	-0.03325	-0.23229	-0.12972	-0.02117
$\boldsymbol{x_4}$	$\beta_{4,2}$	-0.90097	-0.78990	-0.68022	-1.04531	-0.94110	-0.81986	-0.98118	-0.86930	-0.76488	-0.83371	-0.71080	-0.58892	-0.95889	-0.84530	-0.72986	-0.84223	-0.72230	-0.59899
	$\beta_{4,3}$	-0.90097	-0.78990	-0.68022	-1.04531	-0.94110	-0.81986	0.00000	0.00000	0.00000	-0.83371	-0.71080	-0.58892	-0.95090	-0.43850	0.03331	-0.88346	-0.38450	0.11071
	$\beta_{4,4}$	-0.90097	-0.78990	-0.68022	-1.04531	-0.94110	-0.81986	-0.98118	-0.86930	-0.76488	-0.99368	-0.87580	-0.76570	-1.13191	-0.98150	-0.84537	-0.98599	-0.87100	-0.75965
	$\beta_{4,5}$	-0.90097	-0.78990	-0.68022	-1.04531	-0.94110	-0.81986	-0.98118	-0.86930	-0.76488	-0.99368	-0.87580	-0.76570	-0.95889	-0.84530	-0.72986	-0.98599	-0.87100	-0.75965
$\boldsymbol{x_5}$	$\beta_{5,2}$	0.02965	0.11848	0.19984	-0.09043	-0.00392	0.08160	-0.09623	-0.01153	0.07207	-0.07838	-0.03678	0.00601	-0.10674	-0.05944	-0.01839	-0.07916	-0.03638	0.00649
	$\beta_{5,3}$	0.02965	0.11848	0.19984	-0.09043	-0.00392	0.08160	-0.09623	-0.01153	0.07207	0.00000	0.00000	0.00000	-0.10249	-0.03369	0.03114	-0.07671	-0.01024	0.05627
	$\beta_{5,4}$	0.02965	0.11848	0.19984	-0.09043	-0.00392	0.08160	-0.09623	-0.01153	0.07207	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\boldsymbol{x_6}$	$\beta_{6,2}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.02416	0.14587	0.30724
	$\beta_{6,3}$	0.50435	0.56830	0.63767	0.44956	0.51120	0.58507	0.42541	0.49250	0.55637	0.45688	0.52390	0.58845	0.42978	0.49500	0.55918	0.45218	0.51940	0.58674
	$\beta_{6,4}$	0.00000	0.00000	0.00000	-0.27804	-0.24170	-0.20351	-0.25561	-0.22020	-0.18085	-0.23827	-0.19680	-0.15790	-0.23987	-0.19970	-0.15957	-0.23136	-0.18990	-0.14893
	$\beta_{6,5}$	0.00000	0.00000	0.00000	-0.27804	-0.24170	-0.20351	-0.25561	-0.22020	-0.18085	-0.23827	-0.19680	-0.15790	-0.23987	-0.19970	-0.15957	-0.23136	-0.18990	-0.14893
$\boldsymbol{x_7}$	$\beta_{7,2}$	-0.04383	0.01885	0.09220	-0.05026	0.01679	0.07954	-0.12286	0.16189	0.45530	-0.07373	0.19280	0.48799	-0.09835	0.19420	0.45453	-0.11616	0.17380	0.45264
	$\beta_{7,3}$	-0.04383	0.01885	0.09220	-0.05026	0.01679	0.07954	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
	$\beta_{7,4}$	-0.04383	0.01885	0.09220	-0.05026	0.01679	0.07954	0.00000	0.00000	0.00000	-0.01168	0.03922	0.09883	0.00000	0.00000	0.00000	-0.05576	0.01326	0.06894
	$\beta_{7,5}$	-0.04383	0.01885	0.09220	-0.05026	0.01679	0.07954	0.00000	0.00000	0.00000	-0.01168	0.03922	0.09883	-0.05572	0.01627	0.09068	-0.07170	0.03075	0.12968
	$\beta_{7,6}$	-0.04383	0.01885	0.09220	-0.05026	0.01679	0.07954	0.00000	0.00000	0.00000	-0.01168	0.03922	0.09883	0.00000	0.00000	0.00000	-0.05576	0.01326	0.06894
	$\beta_{7,7}$	-0.04383	0.01885	0.09220	-0.05026	0.01679	0.07954	0.00000	0.00000	0.00000	-0.01168	0.03922	0.09883	-0.05572	0.01627	0.09068	-0.13716	0.04091	0.20472
	$\beta_{7,8}$	-0.04383	0.01885	0.09220	-0.05026	0.01679	0.07954	0.00000	0.00000	0.00000	-0.17756	-0.02281	0.12068	-0.18571	-0.05050	0.09251	-0.19651	-0.04168	0.10897
$\boldsymbol{x_8}$	$\beta_{8,2}$	0.03772	0.12813	0.23095	0.03542	0.12527	0.22284	-0.65963	-0.34780	-0.07383	-0.66013	-0.38070	-0.11725	-0.59526	-0.31740	-0.03141	-0.60402	-0.32810	-0.02098
	$\beta_{8,3}$	0.03772	0.12813	0.23095	0.03542	0.12527	0.22284	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.64168	-0.10458	0.45978	-0.64277	-0.07151	0.45919
	$\beta_{8,4}$	0.03772	0.12813	0.23095	0.03542	0.12527	0.22284	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01030	0.09212	0.17044	-0.02409	0.06762	0.16095
	$\beta_{8,5}$	0.03772	0.12813	0.23095	0.03542	0.12527	0.22284	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01030	0.09212	0.17044	-0.13702	0.13709	0.39279
	$\beta_{8,6}$	0.03772	0.12813	0.23095	0.03542	0.12527	0.22284	-0.65963	-0.34780	-0.07383	-0.66013	-0.38070	-0.11725	-0.59526	-0.31740	-0.03141	-0.60402	-0.32810	-0.02098
	$\beta_{8,7}$	0.03772	0.12813	0.23095	0.03542	0.12527	0.22284	-1.04931	-0.37670	0.28801	-0.66013	-0.38070	-0.11725	-0.59526	-0.31740	-0.03141	-0.60402	-0.32810	-0.02098

Tabla 5.2: Resultados de los valores re-estimados de los efectos de nivel de los 6 modelos de mixtura finita, en la etapa de aplicación

		Fusión de efectos de nivel					
Covariable	Nro de grupos	Modelo MF01	Modelo MF02	Modelo MF03	Modelo MF04	Modelo MF05	Modelo MF06
Intercepto		β_0	β_0	β_0	β_0	β_0	β_0
$\boldsymbol{x_1}$	1	$\beta_{1,1}$, $\beta_{1,10}$, $\beta_{1,11}$	$\beta_{1,1}$, $\beta_{1,10}$, $\beta_{1,11}$	$\beta_{1,1}$, $\beta_{1,9}$	$\beta_{1,1}$, $\beta_{1,10}$, $\beta_{1,11}$	$\beta_{1,1}$, $\beta_{1,11}$	$\beta_{1,1}$
	2	$\beta_{1,2}$, $\beta_{1,9}$	$\beta_{1,2}$, $\beta_{1,9}$	$\beta_{1,2}$	$\beta_{1,2}$, $\beta_{1,6}$	$\beta_{1,2}$, $\beta_{1,9}$	$\beta_{1,2}$
	3	$\beta_{1,3}$, $\beta_{1,4}$, $\beta_{1,5}$, $\beta_{1,6}$	$\beta_{1,3}$, $\beta_{1,4}$, $\beta_{1,5}$, $\beta_{1,6}$	$\beta_{1,3}$, $\beta_{1,4}$, $\beta_{1,5}$, $\beta_{1,6}$	$\beta_{1,3}$, $\beta_{1,4}$, $\beta_{1,5}$	$\beta_{1,3}$, $\beta_{1,4}$, $\beta_{1,5}$, $\beta_{1,6}$	$\beta_{1,3}$, $\beta_{1,4}$, $\beta_{1,5}$
	4			$\beta_{1,10}$, $\beta_{1,11}$	$\beta_{1,9}$	$\beta_{1,10}$	$\beta_{1,6}$
	5						$\beta_{1,9}$
	6						$\beta_{1,10}$
	7						$\beta_{1,11}$
$\boldsymbol{x_2}$	1	$\beta_{2,1}$, $\beta_{2,2}$, $\beta_{2,3}$, $\beta_{2,4}$, $\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,1}$, $\beta_{2,2}$, $\beta_{2,3}$, $\beta_{2,4}$, $\beta_{2,5}$, $\beta_{2,6}$	$\beta_{2,1}$, $\beta_{2,6}$	$\beta_{2,1}$	$\beta_{2,1}$, $\beta_{2,2}$	$\beta_{2,1}$
	2	$\beta_{2,7}$	$\beta_{2,7}$	$\beta_{2,2}$, $\beta_{2,3}$, $\beta_{2,4}$	$\beta_{2,2}$, $\beta_{2,3}$, $\beta_{2,4}$, $\beta_{2,6}$	$\beta_{2,3}$, $\beta_{2,4}$, $\beta_{2,6}$	$\beta_{2,2}$
	3			$\beta_{2,5}$	$\beta_{2,5}$	$\beta_{2,5}$	$\beta_{2,3}$, $\beta_{2,4}$, $\beta_{2,6}$
	4			$\beta_{2,7}$	$\beta_{2,7}$	$\beta_{2,7}$	$\beta_{2,5}$
	5						$\beta_{2,7}$
$\boldsymbol{x_3}$	1	$\beta_{3,1}$	$\beta_{3,1}$	$\beta_{3,1}$, $\beta_{3,5}$	$\beta_{3,1}$	$\beta_{3,1}$	$\beta_{3,1}$
	2	$\beta_{3,3}$, $\beta_{3,4}$, $\beta_{3,5}$, $\beta_{3,6}$, $\beta_{3,7}$, $\beta_{3,8}$, $\beta_{3,9}$, $\beta_{3,10}$	$\beta_{3,3}$, $\beta_{3,4}$, $\beta_{3,5}$, $\beta_{3,6}$, $\beta_{3,7}$, $\beta_{3,8}$, $\beta_{3,9}$, $\beta_{3,10}$	$\beta_{3,3}$, $\beta_{3,4}$, $\beta_{3,6}$, $\beta_{3,7}$, $\beta_{3,8}$, $\beta_{3,9}$, $\beta_{3,10}$	$\beta_{3,3}$, $\beta_{3,4}$, $\beta_{3,8}$, $\beta_{3,9}$, $\beta_{3,10}$	$\beta_{3,3}$, $\beta_{3,4}$, $\beta_{3,9}$	$\beta_{3,3}$, $\beta_{3,4}$
	3				$\beta_{3,5}$	$\beta_{3,5}$	$\beta_{3,5}$
	4				$\beta_{3,6}$, $\beta_{3,7}$	$\beta_{3,6}$	$\beta_{3,6}$
	5					$\beta_{3,7}$	$\beta_{3,7}$
	6					$\beta_{3,8}$, $\beta_{3,10}$	$\beta_{3,8}$
	7						$\beta_{3,9}$
	8						$\beta_{3,10}$
$\boldsymbol{x_4}$	1	$\beta_{4,1}$	$\beta_{4,1}$	$\beta_{4,1}$, $\beta_{4,3}$	$\beta_{4,1}$	$\beta_{4,1}$	$\beta_{4,1}$
	2	$\beta_{4,2}$, $\beta_{4,3}$, $\beta_{4,4}$, $\beta_{4,5}$	$\beta_{4,2}$, $\beta_{4,3}$, $\beta_{4,4}$, $\beta_{4,5}$	$\beta_{4,2}$, $\beta_{4,4}$, $\beta_{4,5}$	$\beta_{4,2}$, $\beta_{4,3}$	$\beta_{4,2}$, $\beta_{4,5}$	$\beta_{4,2}$
	3				$\beta_{4,4}$, $\beta_{4,5}$	$\beta_{4,3}$	$\beta_{4,3}$
	4					$\beta_{4,4}$	$\beta_{4,4}$, $\beta_{4,5}$
$\boldsymbol{x_5}$	1	$\beta_{5,1}$	$\beta_{5,1}$	$\beta_{5,1}$	$\beta_{5,1}$, $\beta_{5,3}$, $\beta_{5,4}$	$\beta_{5,1}$, $\beta_{5,4}$	$\beta_{5,1}$, $\beta_{5,4}$
	2	$\beta_{5,2}$, $\beta_{5,3}$, $\beta_{5,4}$	$\beta_{5,2}$, $\beta_{5,3}$, $\beta_{5,4}$	$\beta_{5,2}$, $\beta_{5,3}$, $\beta_{5,4}$	$\beta_{5,2}$	$\beta_{5,2}$	$\beta_{5,2}$
	3					$\beta_{5,3}$	$\beta_{5,3}$
$\boldsymbol{x_6}$	1	$\beta_{6,1}$, $\beta_{6,2}$, $\beta_{6,4}$, $\beta_{6,5}$	$\beta_{6,1}$, $\beta_{6,2}$	$\beta_{6,1}$, $\beta_{6,2}$	$\beta_{6,1}$, $\beta_{6,2}$	$\beta_{6,1}$, $\beta_{6,2}$	$\beta_{6,1}$
	2	$\beta_{6,3}$	$\beta_{6,3}$	$\beta_{6,3}$	$\beta_{6,3}$	$\beta_{6,3}$	$\beta_{6,2}$
	3		$\beta_{6,4}$, $\beta_{6,5}$	$\beta_{6,4}$, $\beta_{6,5}$	$\beta_{6,4}$, $\beta_{6,5}$	$\beta_{6,4}$, $\beta_{6,5}$	$\beta_{6,3}$
	4						$\beta_{6,4}$, $\beta_{6,5}$
$\boldsymbol{x_7}$	1	$\beta_{7,1}$	$\beta_{7,1}$	$\beta_{7,1}$, $\beta_{7,3}$, $\beta_{7,4}$, $\beta_{7,5}$, $\beta_{7,6}$, $\beta_{7,7}$, $\beta_{7,8}$	$\beta_{7,1}$, $\beta_{7,3}$	$\beta_{7,1}$, $\beta_{7,3}$, $\beta_{7,4}$, $\beta_{7,6}$	$\beta_{7,1}$, $\beta_{7,3}$
	2	$\beta_{7,2}$, $\beta_{7,3}$, $\beta_{7,4}$, $\beta_{7,5}$, $\beta_{7,6}$, $\beta_{7,7}$, $\beta_{7,8}$	$\beta_{7,2}$, $\beta_{7,3}$, $\beta_{7,4}$, $\beta_{7,5}$, $\beta_{7,6}$, $\beta_{7,7}$, $\beta_{7,8}$	$\beta_{7,2}$	$\beta_{7,2}$	$\beta_{7,2}$	$\beta_{7,2}$
	3				$\beta_{7,4}$, $\beta_{7,5}$, $\beta_{7,6}$, $\beta_{7,7}$	$\beta_{7,5}$, $\beta_{7,7}$	$\beta_{7,4}$, $\beta_{7,6}$
	4				$\beta_{7,8}$	$\beta_{7,8}$	$\beta_{7,5}$
	5						$\beta_{7,7}$
	6						$\beta_{7,8}$
$\boldsymbol{x_8}$	1	$\beta_{8,1}$	$\beta_{8,1}$	$\beta_{8,1}$, $\beta_{8,3}$, $\beta_{8,4}$, $\beta_{8,5}$	$\beta_{8,1}$, $\beta_{8,3}$, $\beta_{8,4}$, $\beta_{8,5}$	$\beta_{8,1}$	$\beta_{8,1}$
	2	$\beta_{8,2}$, $\beta_{8,3}$, $\beta_{8,4}$, $\beta_{8,5}$, $\beta_{8,6}$, $\beta_{8,7}$	$\beta_{8,2}$, $\beta_{8,3}$, $\beta_{8,4}$, $\beta_{8,5}$, $\beta_{8,6}$, $\beta_{8,7}$	$\beta_{8,2}$, $\beta_{8,6}$	$\beta_{8,2}$, $\beta_{8,6}$, $\beta_{8,7}$	$\beta_{8,2}$, $\beta_{8,6}$, $\beta_{8,7}$	$\beta_{8,2}$, $\beta_{8,6}$, $\beta_{8,7}$
	3			$\beta_{8,7}$		$\beta_{8,3}$	$\beta_{8,3}$
	4					$\beta_{8,4}$, $\beta_{8,5}$	$\beta_{8,4}$
	5						$\beta_{8,5}$

Tabla 5.3: Comparación de la fusión de los efectos de nivel de los 6 modelos de mixtura finita, en la etapa de aplicación

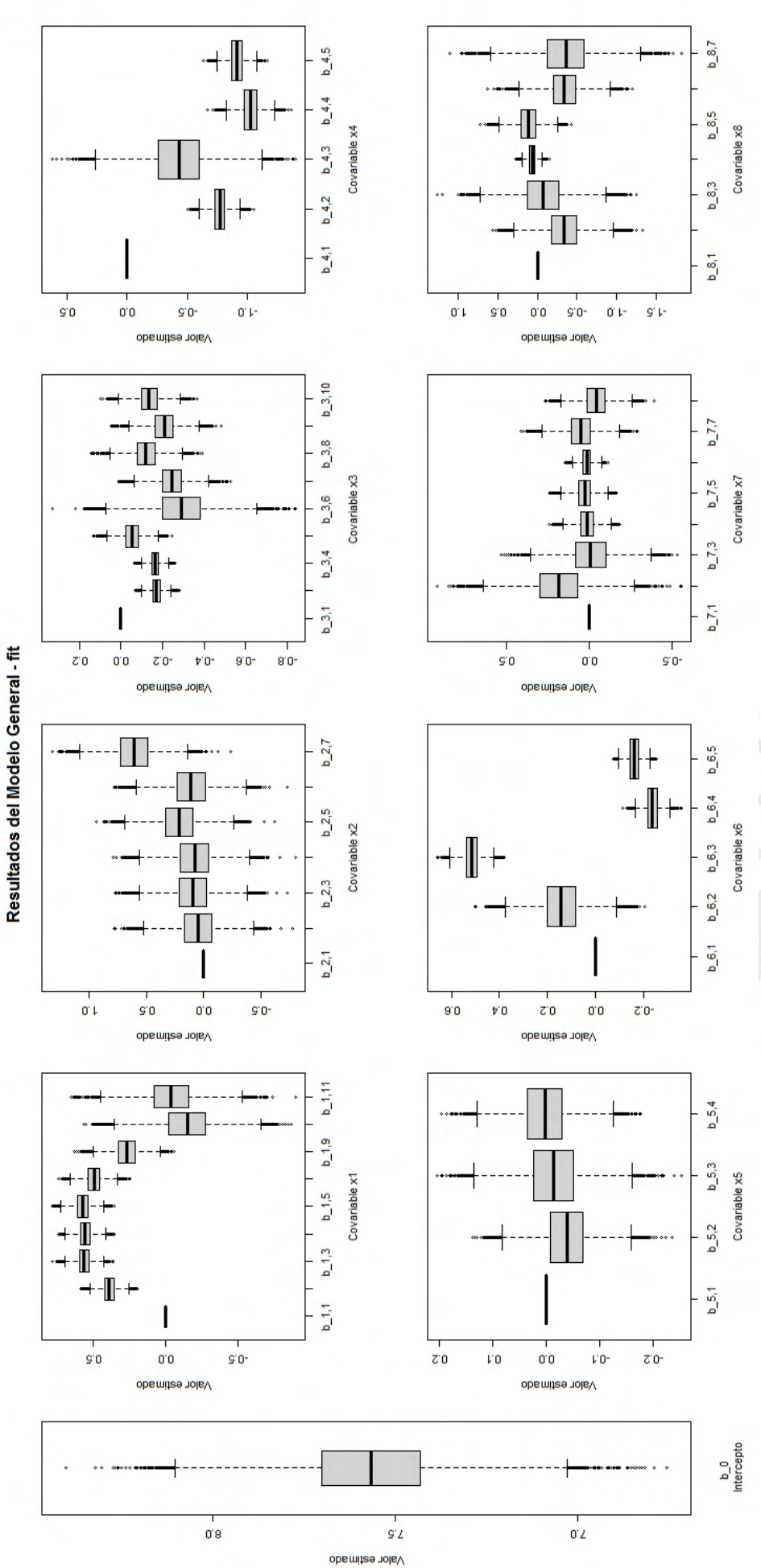

Figura 5.6: Resultados de la estimación de los efectos de nivel en el modelo general

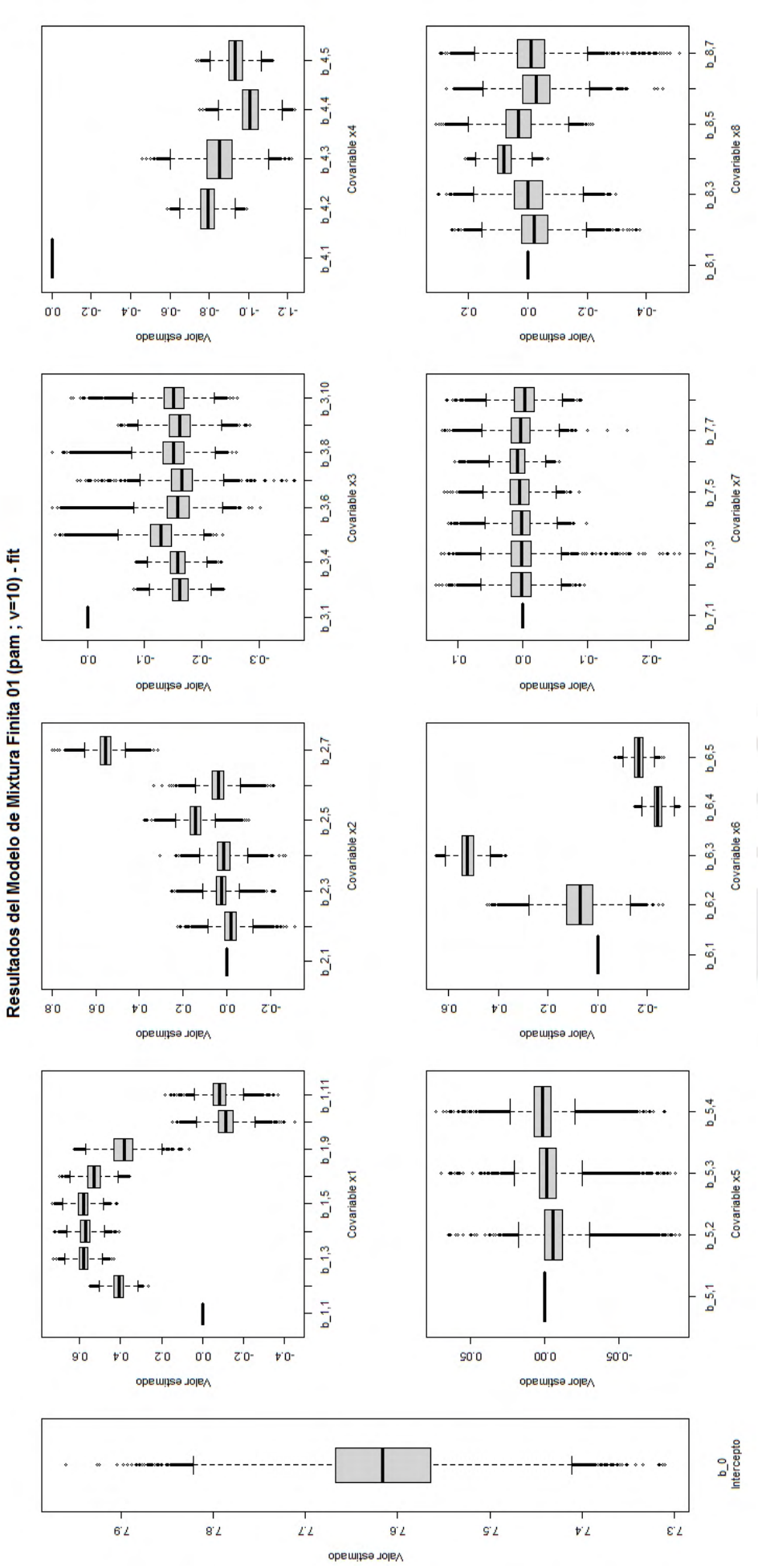

Figura 5.7: Resultados de la estimación de los efectos de nivel en el modelo de mixtura finita 01 (pam ; $v = 10$)

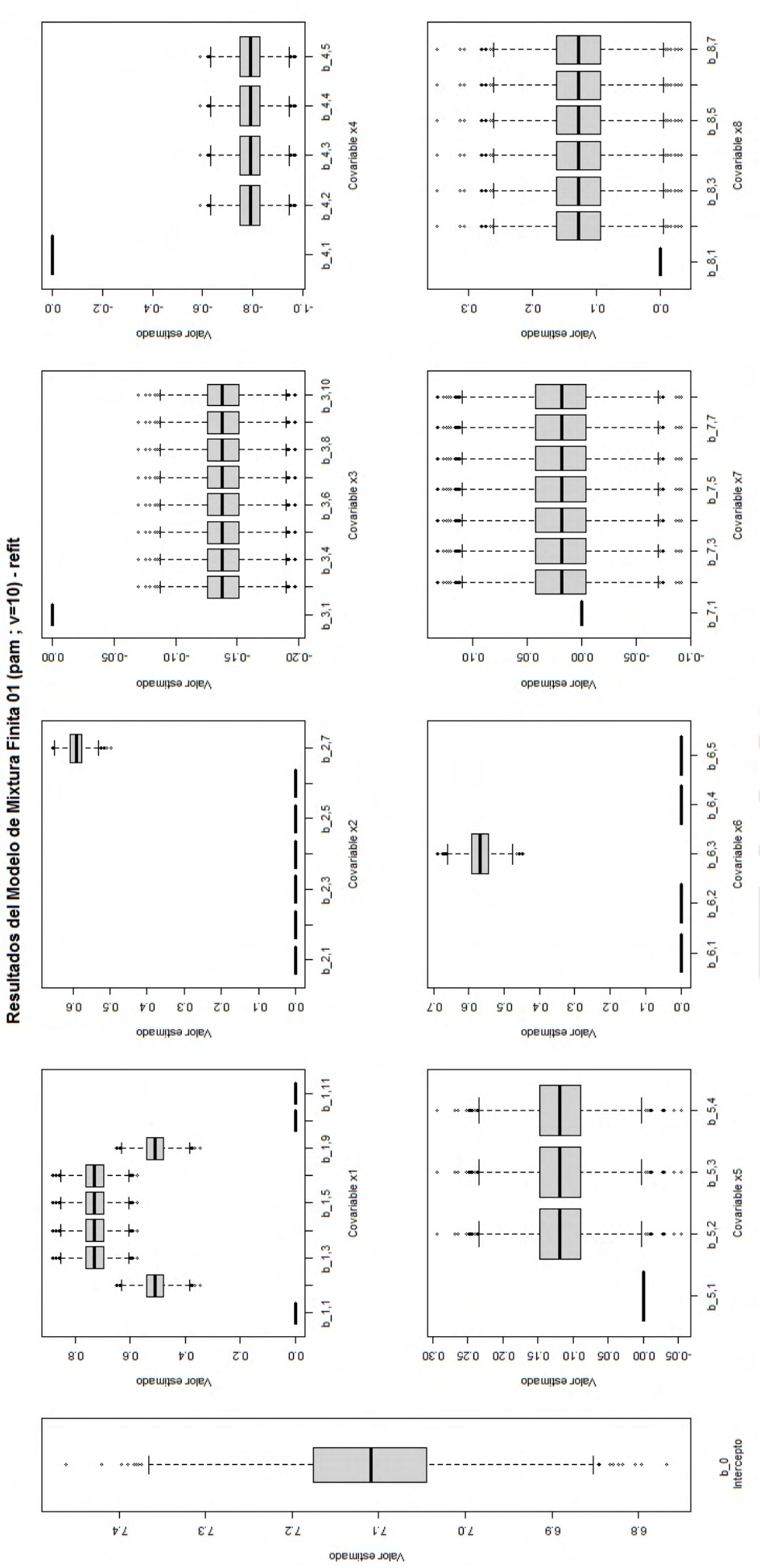

Figura 5.8: Resultados de la re-estimación de los efectos de nivel en el modelo de mixtura finita 01 (pam ; $v = 10$)

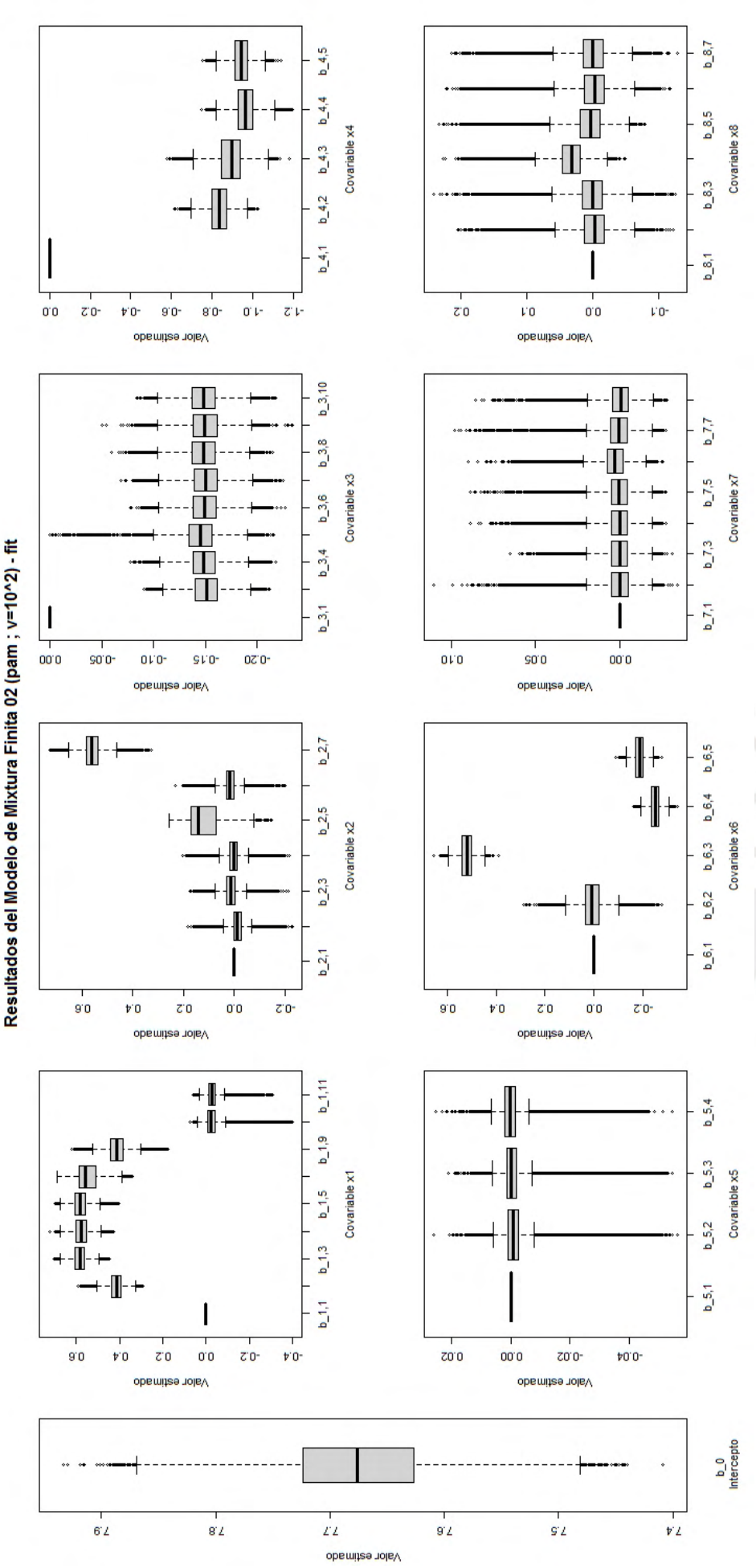

Figura 5.9: Resultados de la estimación de los efectos de nivel en el modelo de mixtura finita 02 (pam ; $v = 10^2$)

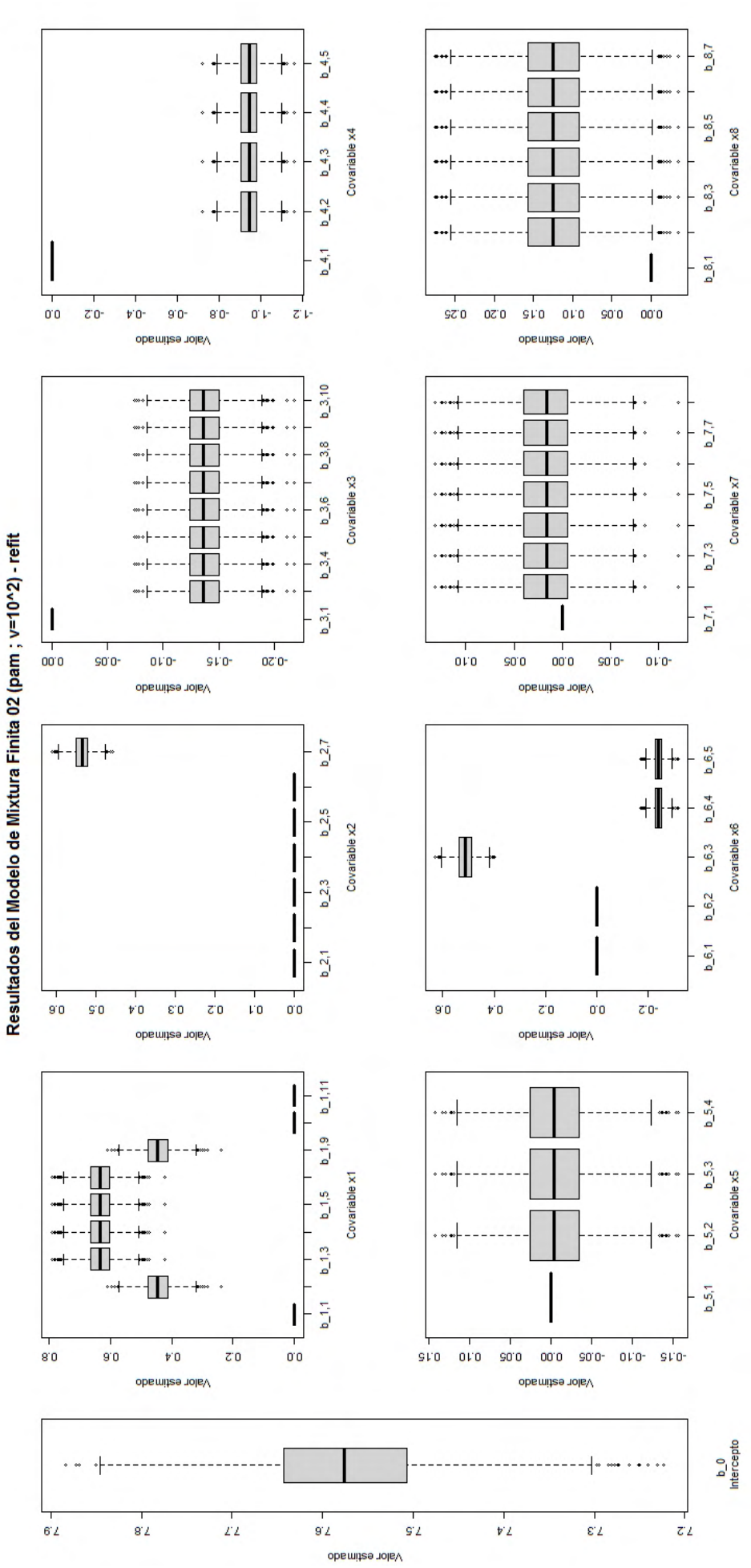

Figura 5.10: Resultados de la re-estimación de los efectos de nivel en el modelo de mixtura finita 02 (pam ; $v = 10^2$)

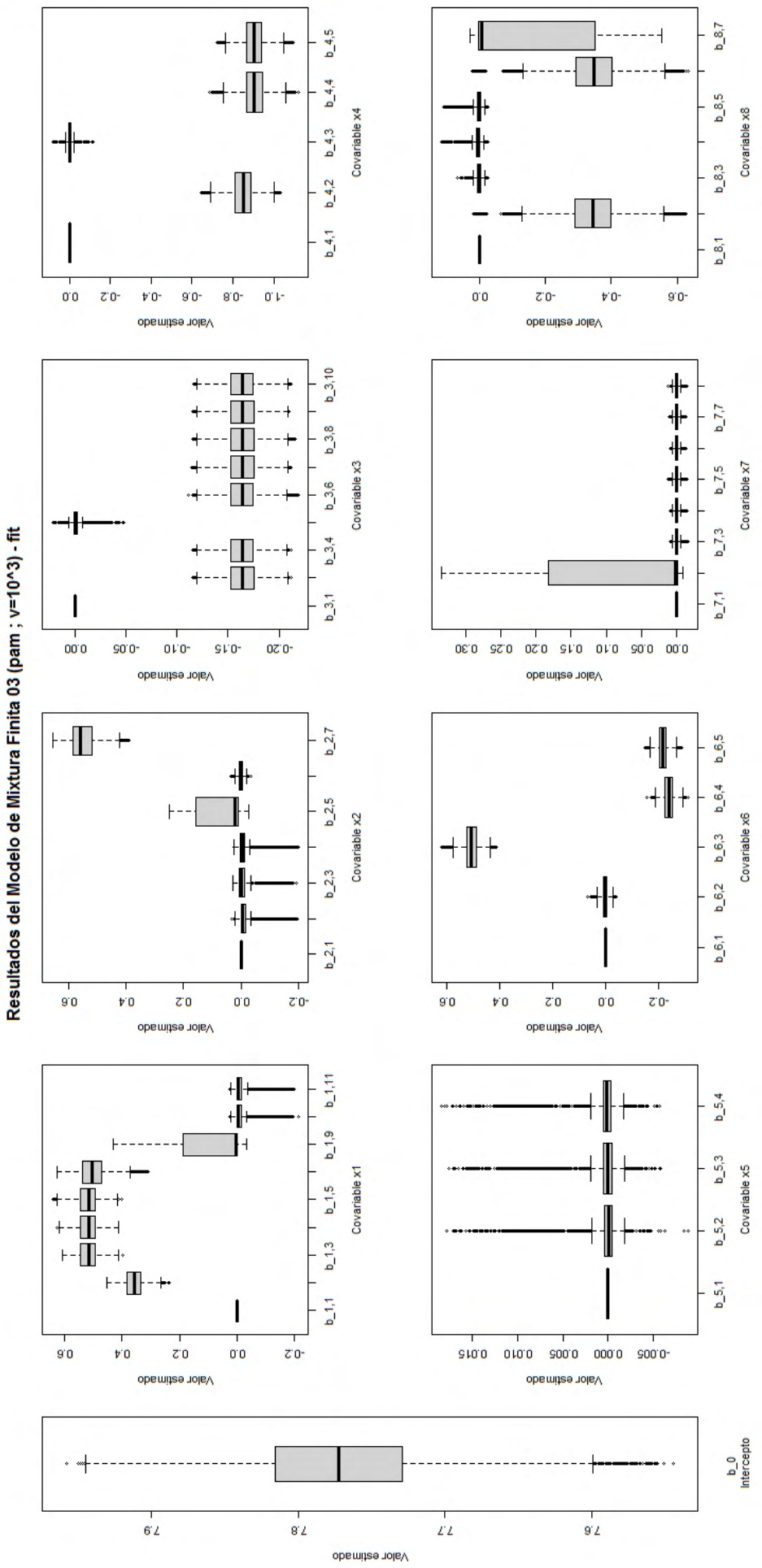

Figura 5.11: Resultados de la estimación de los efectos de nivel en el modelo de mixtura finita 03 (pam ; $v = 10^3$)

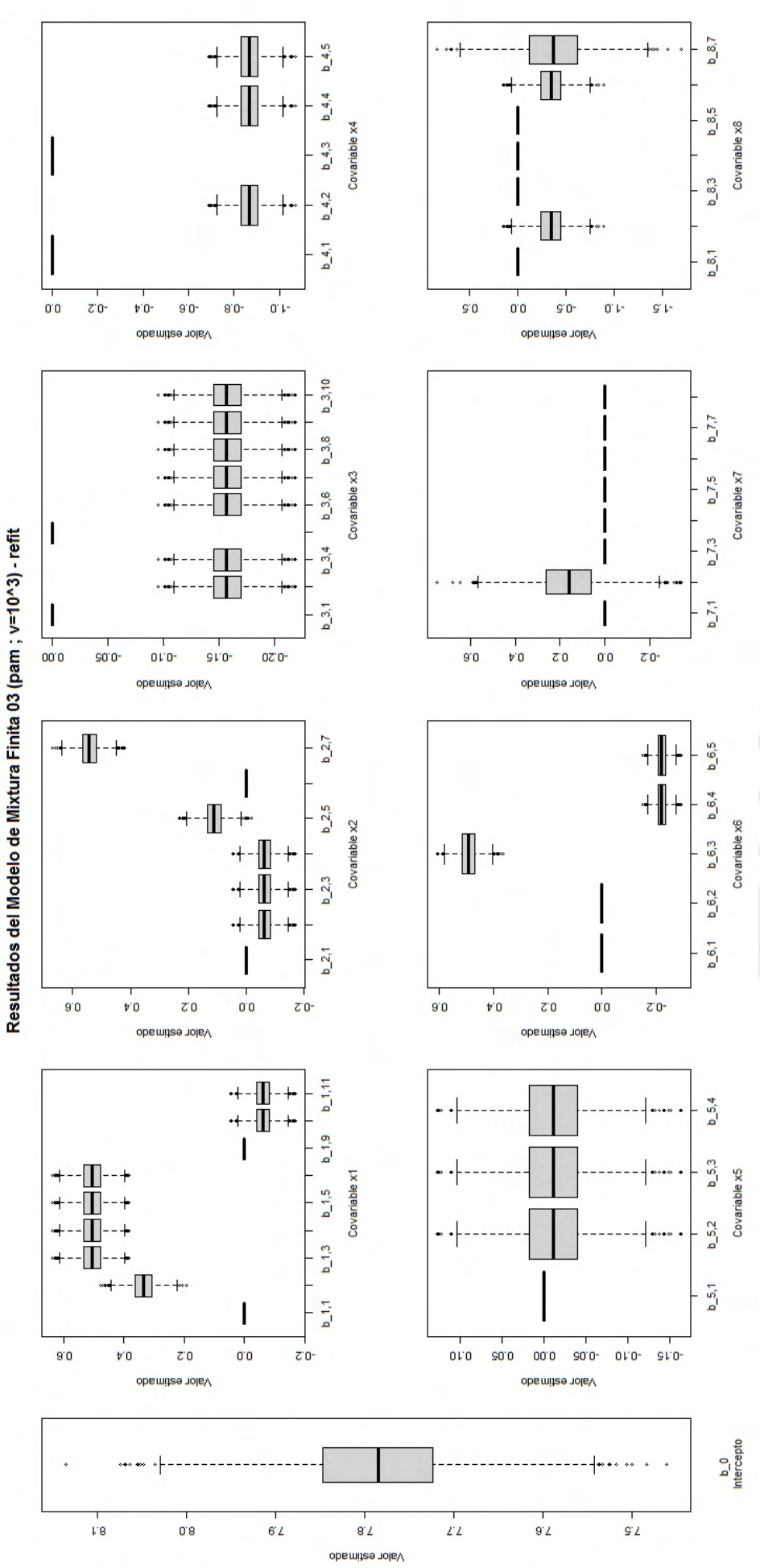

Figura 5.12: Resultados de la re-estimación de los efectos de nivel en el modelo de mixtura finita 03 (pam ; $v = 10^3$)

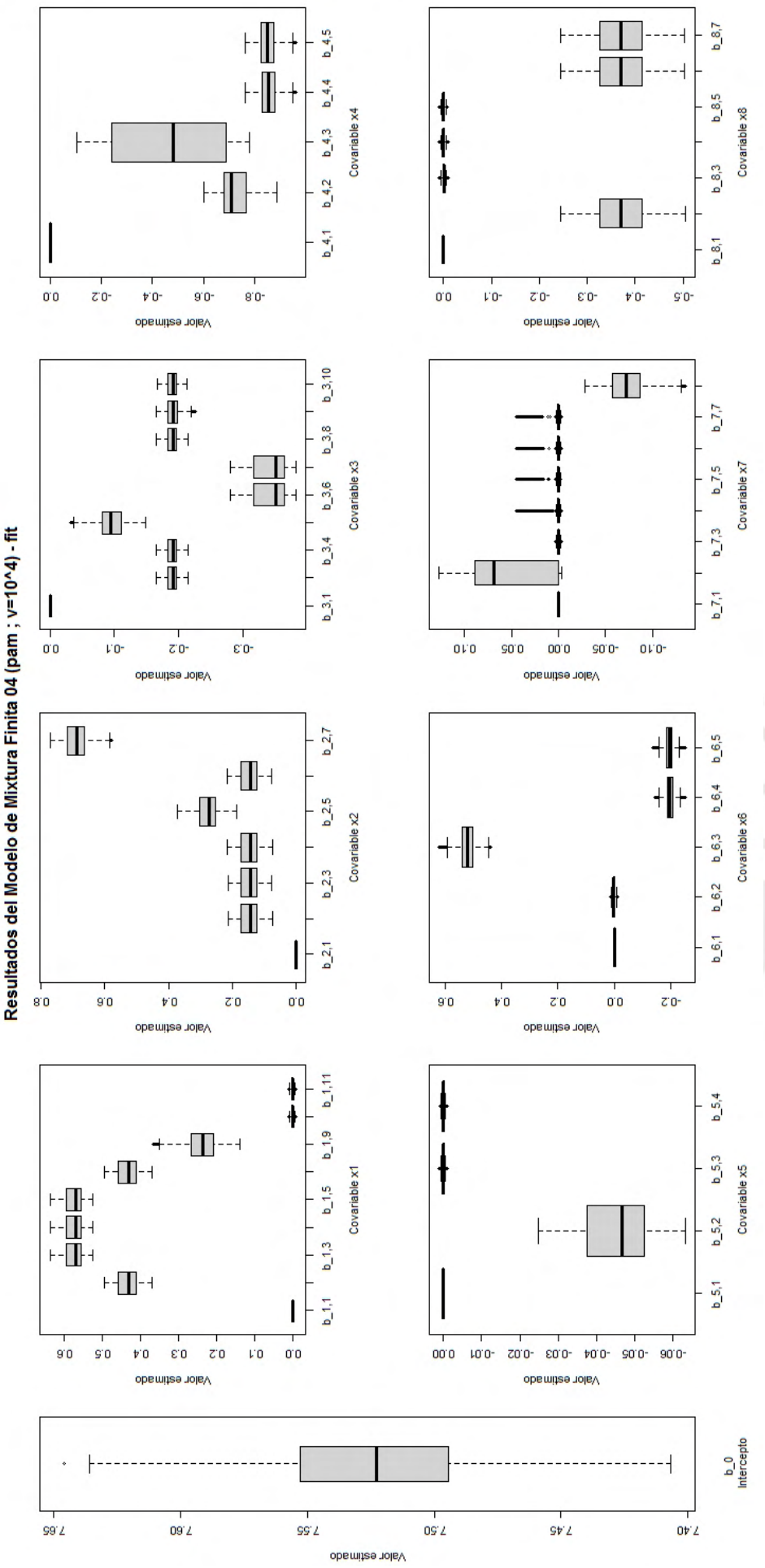

Figura 5.13: Resultados de la estimación de los efectos de nivel en el modelo de mixtura finita 04 (pam ; $v = 10^4$)

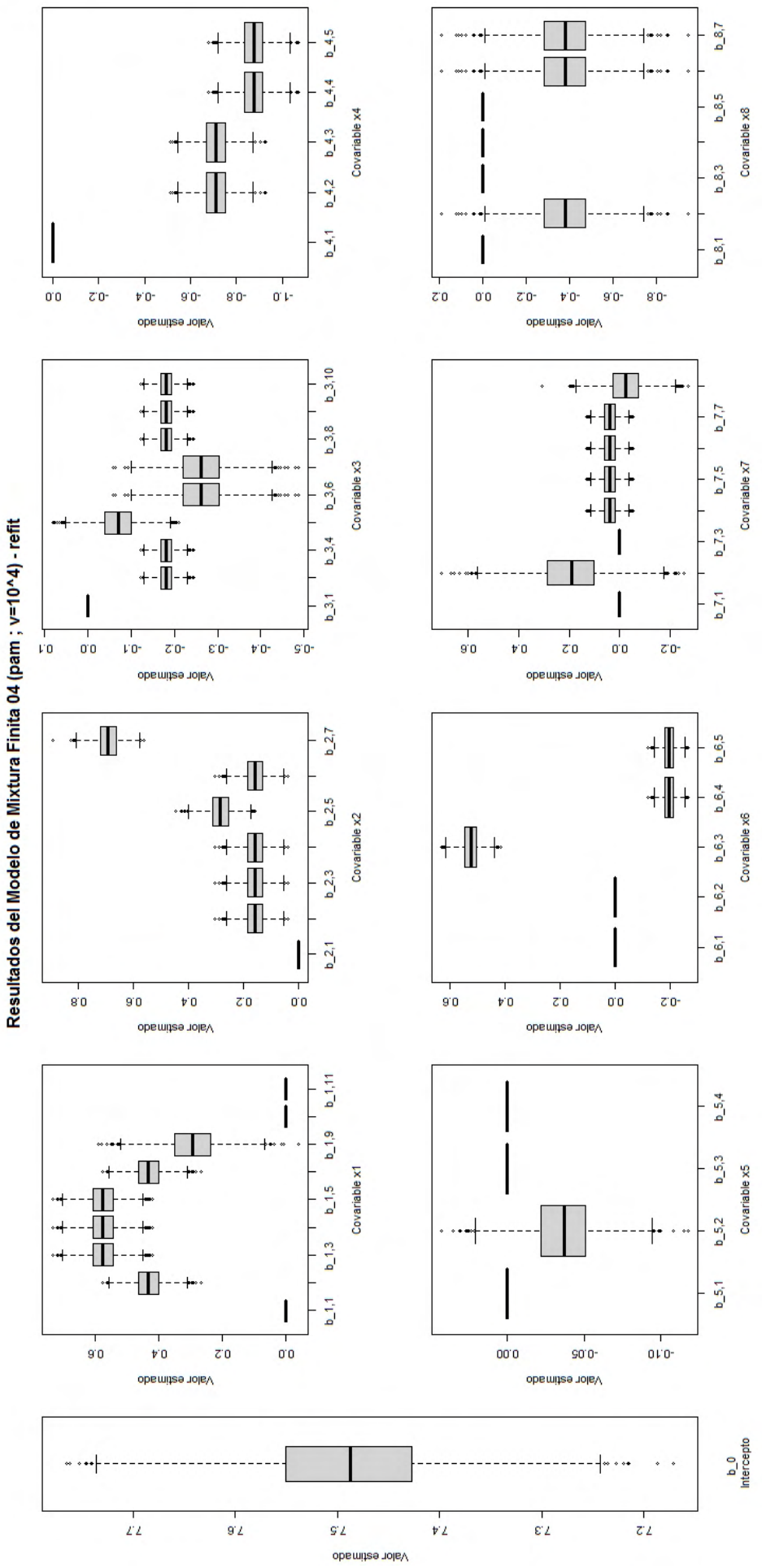

Figura 5.14: Resultados de la re-estimación de los efectos de nivel en el modelo de mixtura finita 04 (pam ; $v = 10^4$)

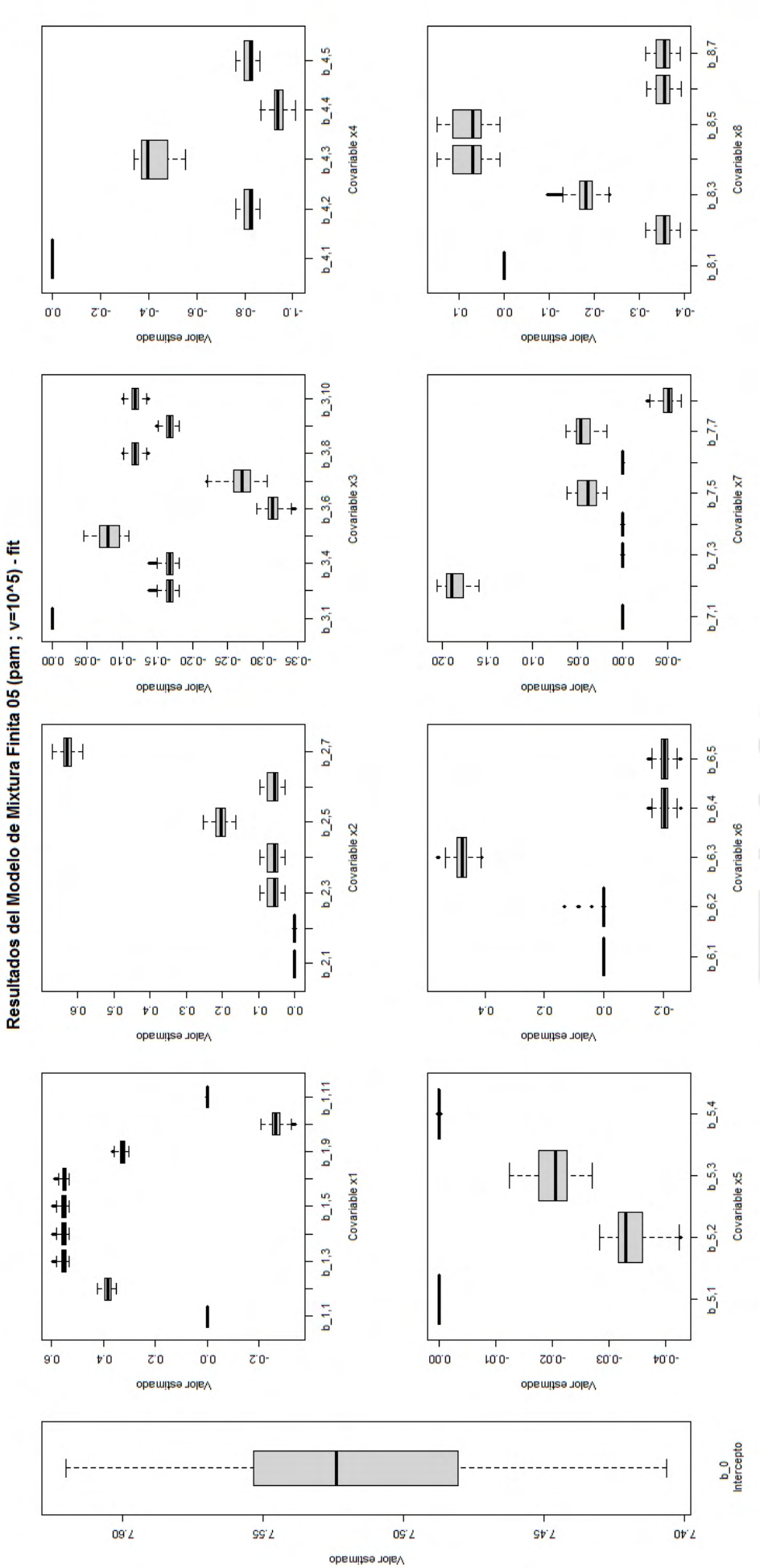

Figura 5.15: Resultados de la estimación de los efectos de nivel en el modelo de mixtura finita 05 (pam ; $v = 10^5$)

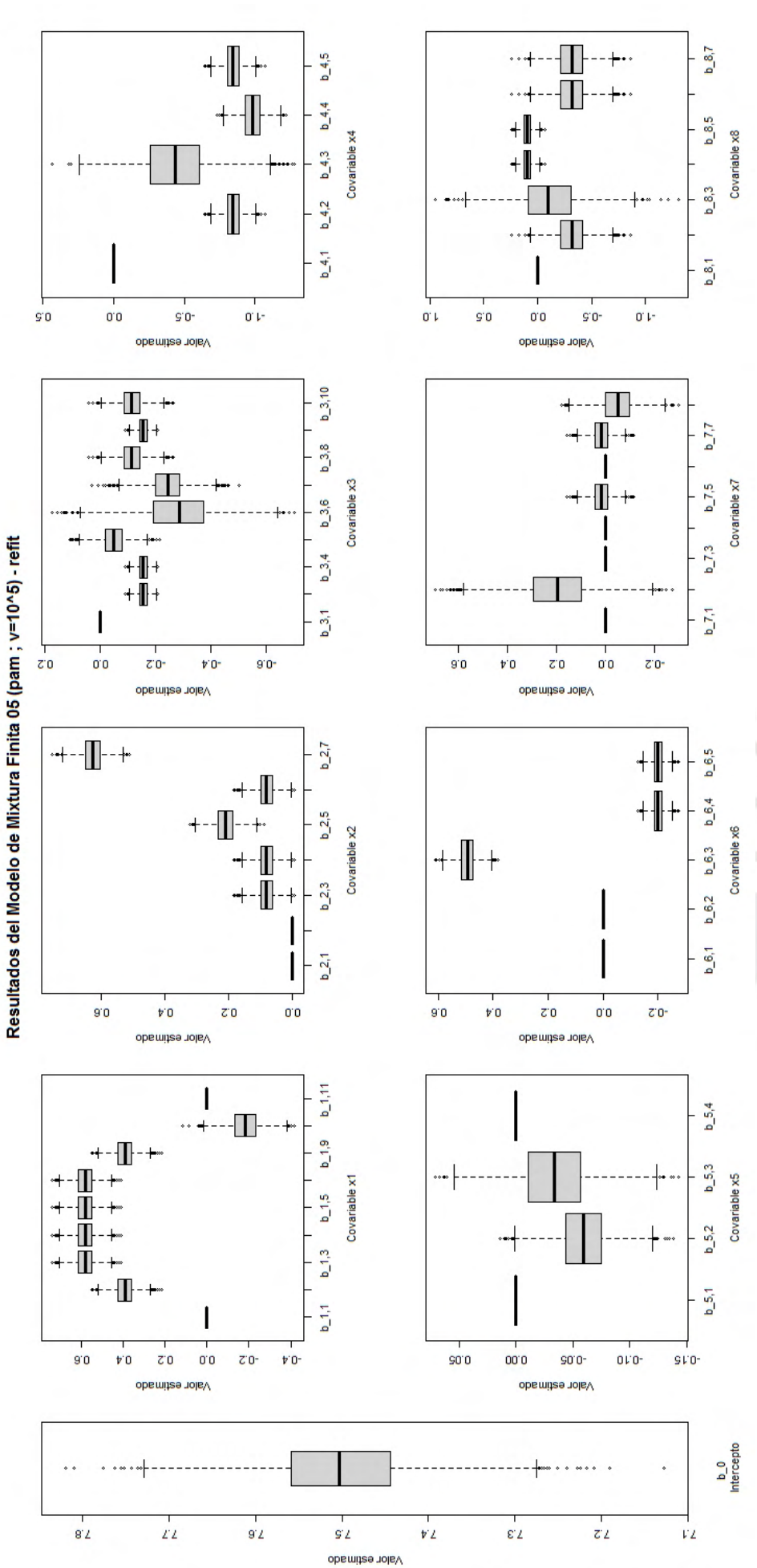

Figura 5.16: Resultados de la re-estimación de los efectos de nivel en el modelo de mixtura finita 05 (pam ; $v = 10^5$)

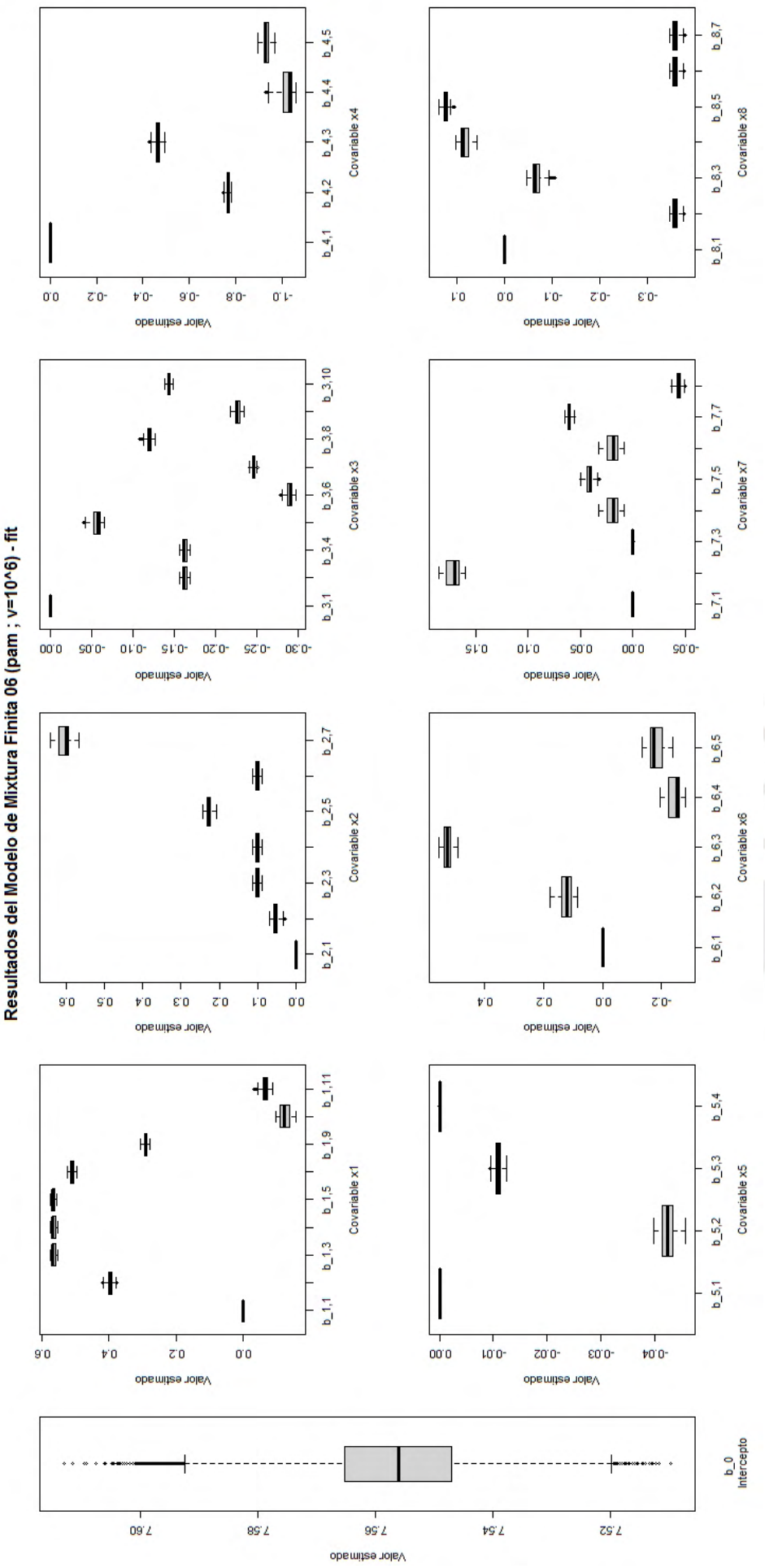

Figura 5.17: Resultados de la estimación de los efectos de nivel en el modelo de mixtura finita 06 (pam ; $v = 10^6$)

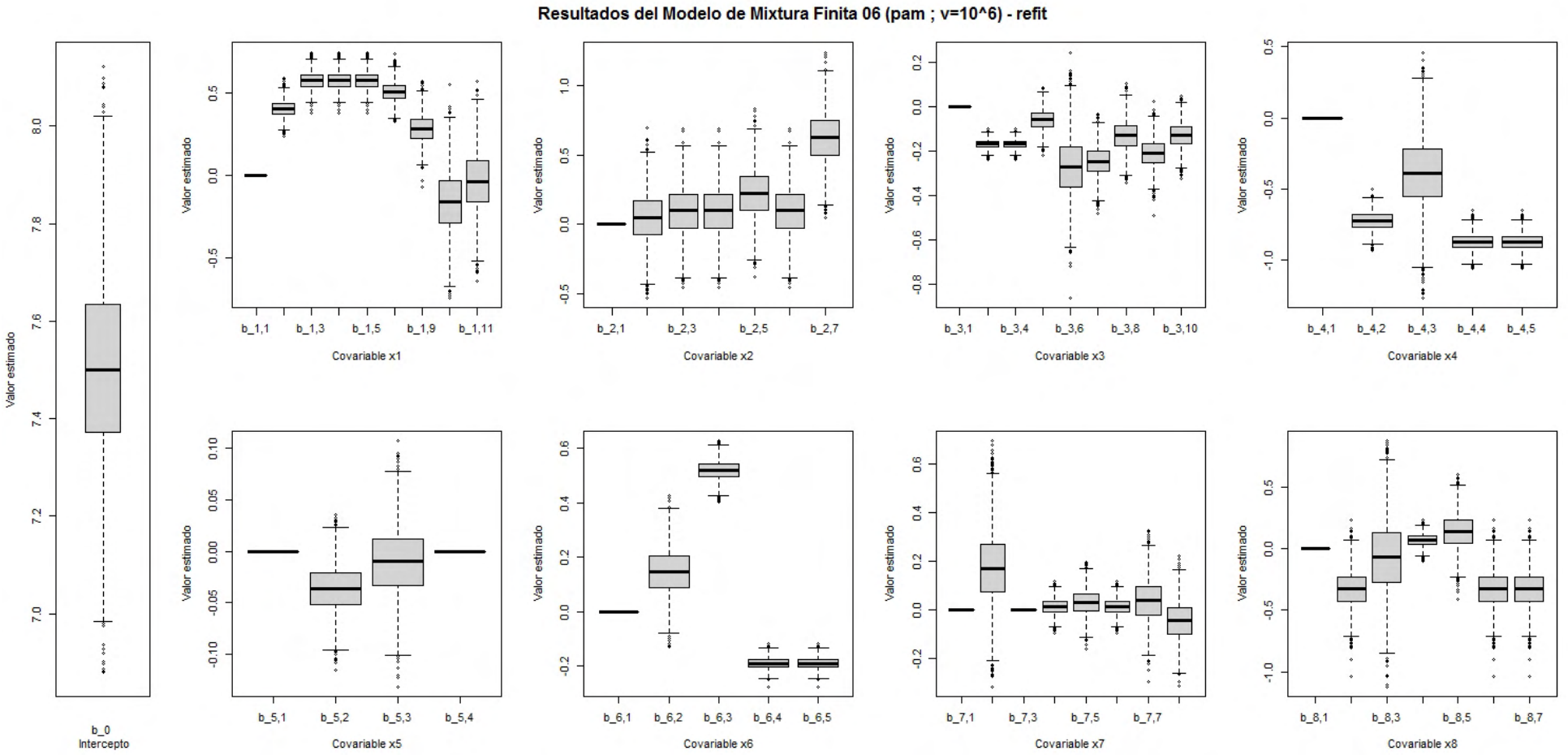

Figura 5.18: Resultados de la re-estimación de los efectos de nivel en el modelo de mixtura finita 06 (pam ; $v = 10^6$)

5.4. Comparación de los Modelos de Regresión

Los resultados del criterio de información de desvío (DIC) permitirán identificar el modelo de regresión lineal con mejor ajuste. A diferencia del estudio de simulación, de la Tabla 5.4 se aprecia que el modelo MF04 (pam ; $v = 10^4$) es quien proporciona el mejor ajuste.

Criterio	Modelo MF01	Modelo MF02	Modelo MF03	Modelo MF04	Modelo MF05	Modelo MF06
DIC	5504.250	5348.126	5312.516	5252.537	5273.431	5264.182

Tabla 5.4: Comparación del DIC de los modelos de regresión lineal en la etapa de aplicación

A partir de los resultados obtenidos en la aplicación, se puede concluir los siguientes puntos:

- El modelo MF01 (pam ; $v = 10$) presenta la forma más reducida del modelo de regresión lineal (17 grupos de efectos), sin embargo, el valor del DIC es mayor que el resto de modelos.
- El modelo MF04 (pam ; $v = 10^4$) presenta un mejor ajuste que los otros modelos de regresión lineal, debido a su menor valor en el DIC, y presenta una extensión intermedia del modelo de regresión lineal (26 grupos de efectos), comparándolos con los otros modelos.
- El modelo MF06 (pam ; $v = 10^6$) presenta la forma más extensa del modelo de regresión lineal (42 grupos de efectos), sin embargo , el DIC es el segundo menor valor de todos los modelos.

En ese sentido, el modelo MF04 (pam ; $v = 10^4$), no sólo presenta el mejor ajuste al modelo de regresión lineal debido a su menor valor de DIC, sino que también ofrece una expresión más reducida de la ecuación (5.1):

$$\begin{aligned} y_i \;=\; & \beta_0 + \\ & (x_{i1,2}\beta_{1,2} + x_{i1,3}\beta_{1,3} + x_{i1,4}\beta_{1,4} + x_{i1,5}\beta_{1,5} + x_{i1,6}\beta_{1,6} + x_{i1,9}\beta_{1,9}) + \\ & (x_{i2,2}\beta_{2,2} + x_{i2,3}\beta_{2,3} + x_{i2,4}\beta_{2,4} + x_{i2,5}\beta_{2,5} + x_{i2,6}\beta_{2,6} + x_{i2,7}\beta_{2,7}) + \\ & (x_{i3,3}\beta_{3,3} + x_{i3,4}\beta_{3,4} + x_{i3,5}\beta_{3,5} + x_{i3,6}\beta_{3,6} + x_{i3,7}\beta_{3,7} + x_{i3,8}\beta_{3,8} + \\ & x_{i3,9}\beta_{3,9} + x_{i3,10}\beta_{3,10}) + \\ & (x_{i4,2}\beta_{4,2} + x_{i4,3}\beta_{4,3} + x_{i4,4}\beta_{4,4} + x_{i4,5}\beta_{4,5}) + \\ & (x_{i5,2}\beta_{5,2}) + \\ & (x_{i6,3}\beta_{6,3} + x_{i6,4}\beta_{6,4} + x_{i6,5}\beta_{6,5}) + \\ & (x_{i7,2}\beta_{7,2} + x_{i7,4}\beta_{7,4} + x_{i7,5}\beta_{7,5} + x_{i7,6}\beta_{7,6} + x_{i7,7}\beta_{7,7} + x_{i7,8}\beta_{7,8}) + \\ & (x_{i8,2}\beta_{8,2} + x_{i8,6}\beta_{8,6} + x_{i8,7}\beta_{8,7}) + \\ & \epsilon_i \end{aligned} \tag{5.2}$$

De lo anterior, se puede observar que las covariables categóricas $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ y $\boldsymbol{x_4}$ no presentan ninguna reducción, es decir, sus categorías no se han fusionado.

Finalmente, de los resultados del estudio de simulación y de la aplicación, se puede afirmar lo siguiente:

- Los modelos de regresión lineal más reducidos se obtienen cuando se trabaja con mayor variabilidad (menores valores de v) en los componentes de la mixtura de la distribución a priori.
- Los modelos de regresión lineal con mejor ajuste a los datos se obtienen cuando se trabaja con menor variabilidad (mayores valores de v) en los componentes de la mixtura de la distribución a priori.
- Se comprueba que: 1) la técnica de agrupamiento 'pam' no permite que una covariable se excluya del modelo de regresión lineal, es decir, la técnica 'pam' no permite que todas las categorías de una covariable se fusionen al nivel de la línea base; y 2) el método de fusión de efectos considera a todas las covariables categóricas como nominales, es decir, durante la aplicación del método no se toma en cuenta el orden de las categorías de las covariables ordinales.

Capítulo 6

Conclusiones y Recomendaciones

6.1. Conclusiones

El presente trabajo de tesis tuvo como principal objetivo el desarrollo del método de fusión de efectos de nivel de covariables categóricas usando la técnica de agrupamiento PAM, propuesto por Malsiner-Walli, Pauger y Wagner (2018), y aplicarlo en un conjunto de datos reales relacionados a los ingresos monetarios de la población de Lima Metropolitana y Callao en el primer trimestre del 2020. Por ende, en el desarrollo de la investigación para el cumplimiento del objetivo principal, se llegaron a las siguientes conclusiones:

- Los modelos de regresión lineal con mejor ajuste a los datos se obtienen cuando se trabaja con menores valores de variabilidad (mayores valores de v) en los componentes de la mixtura de la distribución a priori, y viceversa. Por ello, en el estudio de simulación (Capítulo 4) y en la aplicación (Capítulo 5) se obtuvieron menores valores de DIC cuando se usa valores de $v \geq 10^4$.

- Los modelos de regresión lineal más reducidos se obtienen cuando se trabaja con mayor variabilidad (menores valores de v) en los componentes de la mixtura de la distribución a priori, y viceversa. Por ello, en el estudio de simulación (Capítulo 4) y en la aplicación (Capítulo 5) se obtuvo que la forma más reducida del modelo de regresión resulta cuando se usa el valor de $v = 10$ y la forma más extensa del modelo de regresión resulta cuando se usa el valor de $v = 10^6$.

- En ese sentido, se puede afirmar que a mayor valor de v, se consigue que los componentes de la mixtura presenten menor dispersión, y por ende se puedan identificar con mayor facilidad los diferentes grupos de efectos de nivel; obteniéndose de esta forma, modelos con un mayor número de efectos. Por esta razón es que se puede afirmar que la varianza en los componentes de la mixtura a priori controla la precisión de la división de los efectos de nivel.

- Finalmente, las desventajas del método propuesto por Malsiner-Walli, Pauger y Wagner (2018), se corroboraron tanto en el estudio de simulación (Capítulo 4) como en la aplicación (Capítulo 5), los cuales son: 1) todas las covariables categóricas son tratadas como nominales, es decir, para la fusión de efectos no se considera el orden de las

categorías en las covariables ordinales; y 2) todos los efectos de nivel de una covariable categórica no se pueden fusionar al nivel de la línea base, en otras palabras, una covariable categórica no se podrá excluir del modelo.

6.2. Recomendaciones

A consecuencia del trabajo de tesis realizado, se recomiendan los siguientes temas para investigaciones futuras:

- Realizar un estudio de simulación y aplicación con covariables cuantitativas continuas usando el método de mixtura finita con la técnica de agrupamiento 'pam', para observar la eficiencia del paquete effectFusion al trabajar con este tipo de covariables.
- Realizar un estudio de simulación y aplicación usando el método de mixtura finita con la técnica de agrupamiento 'binder' en lugar de 'pam', recomendado en Pauger D., Leitner M., Wagner H. y Malsiner-Walli G. (2019), con la finalidad de comparar qué técnica de agrupamiento brinda un mejor ajuste al modelo de regresión lineal. Es importante mencionar que la técnica 'binder' si permite fusionar toda una covariable categórica a la categoría de referencia.
- Realizar un estudio de simulación y aplicación usando el método de 'spike and slab' con la técnica de agrupamiento 'binder', usado en Malsiner-Walli-Wagner(2011), con la finalidad de comparar qué método de fusión de efectos de nivel brinda un mejor ajuste al modelo de regresión lineal.

Apéndice A

Programa Computacional para la Simulación

A.1. Código para Estudio de Simulación usando el Paquete effectFusion

El código computacional para la implementación del paquete effectFusion del software R en el estudio de simulación es el siguiente:

 # SIMULACIÓN DE MODELO CON 04 COVARIABLES

 # Simulación de datos

```
n=1000

x1=sample(1:5,n,replace=TRUE)
x2=sample(1:6,n,replace=TRUE)
x3=sample(1:7,n,replace=TRUE)
x4=sample(1:8,n,replace=TRUE)

beta0=0
beta1=c(0,0,0,0.5)
beta2=c(0,0.5,0.5,1,1)
beta3=c(0,0.5,0.5,0.75,0.75,1)
beta4=c(0,0.5,0.5,0.75,0.75,1,1)

beta01=c(beta0,beta1)
beta02=c(beta0,beta2)
beta03=c(beta0,beta3)
beta04=c(beta0,beta4)

e=rnorm(n,0,0.5)

y=beta01[x1]+beta02[x2]+beta03[x3]+beta04[x4]+e
```

 # Definición de inputs

```
y=matrix(y)
x=data.frame(var1=factor(x1),var2=factor(x2),var3=factor(x3),var4=factor(x4))
beta=c(beta01,beta02,beta03,beta04)
types=c("n","n","n","n")
data=list(y=y,x=x,beta=beta,types=types)
summary(data)
View(data)

# Resumen de variables dependiente e independientes
summary(x)
summary(y)

# Definición de modelo general (method = NULL)

## Modelo general (method = NULL)
library(coda)
library(effectFusion)
modelogeneral=effectFusion(y,x,types,method=NULL,mcmc=c(M=30000,burnin=15000))
print(modelogeneral)
summary(modelogeneral)
plot(modelogeneral,4)
dic(modelogeneral)
View(modelogeneral)

## Resumen de resultados del modelo general (method = NULL)
summary(modelogeneral$fit$beta)
HPDinterval(as.mcmc(modelogeneral$fit$beta))

# Definición del modelo de mixtura finita (method = 'pam')

## Modelo de mixtura finita 01 (pam ; v=10)
modeloMF01=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=10),mcmc=c(M=30000,burnin=15000))
print(modeloMF01)
summary(modeloMF01)
plot(modeloMF01,4)
model(modeloMF01)
dic(modeloMF01)
View(modeloMF01)

## Resumen de resultados del modelo de mixtura finita 01 (pam ; v=10)
summary(modeloMF01$fit$beta)
```

```
HPDinterval(as.mcmc(modeloMF01$fit$beta))
summary(modeloMF01$refit$beta)
HPDinterval(as.mcmc(modeloMF01$refit$beta))

## Modelo de mixtura finita 02 (pam ; v=10^2)
modeloMF02=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=100),mcmc=c(M=30000,burnin=15000))
print(modeloMF02)
summary(modeloMF02)
plot(modeloMF02,4)
model(modeloMF02)
dic(modeloMF02)
View(modeloMF02)

## Resumen de resultados del modelo de mixtura finita 02 (pam ; v=10^2)
summary(modeloMF02$fit$beta)
HPDinterval(as.mcmc(modeloMF02$fit$beta))
summary(modeloMF02$refit$beta)
HPDinterval(as.mcmc(modeloMF02$refit$beta))

## Modelo de mixtura finita 03 (pam ; v=10^3)
modeloMF03=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=1000),mcmc=c(M=30000,burnin=15000))
print(modeloMF03)
summary(modeloMF03)
plot(modeloMF03,4)
model(modeloMF03)
dic(modeloMF03)
View(modeloMF03)

## Resumen de resultados del modelo de mixtura finita 03 (pam ; v=10^3)
summary(modeloMF03$fit$beta)
HPDinterval(as.mcmc(modeloMF03$fit$beta))
summary(modeloMF03$refit$beta)
HPDinterval(as.mcmc(modeloMF03$refit$beta))

## Modelo de mixtura finita 04 (pam ; v=10^4)
modeloMF04=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=10000),mcmc=c(M=30000,burnin=15000))
print(modeloMF04)
summary(modeloMF04)
plot(modeloMF04,4)
```

```
model(modeloMF04)
dic(modeloMF04)
View(modeloMF04)

## Resumen de resultados del modelo de mixtura finita 04 (pam ; v=10^4)
summary(modeloMF04$fit$beta)
HPDinterval(as.mcmc(modeloMF04$fit$beta))
summary(modeloMF04$refit$beta)
HPDinterval(as.mcmc(modeloMF04$refit$beta))

## Modelo de mixtura finita 05 (pam ; v=10^5)
modeloMF05=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=100000),mcmc=c(M=30000,burnin=15000))
print(modeloMF05)
summary(modeloMF05)
plot(modeloMF05,4)
model(modeloMF05)
dic(modeloMF05)
View(modeloMF05)

## Resumen de resultados del modelo de mixtura finita 05 (pam ; v=10^5)
summary(modeloMF05$fit$beta)
HPDinterval(as.mcmc(modeloMF05$fit$beta))
summary(modeloMF05$refit$beta)
HPDinterval(as.mcmc(modeloMF05$refit$beta))

## Modelo de mixtura finita 06 (pam ; v=10^6)
modeloMF06=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=1000000),mcmc=c(M=30000,burnin=15000))
print(modeloMF06)
summary(modeloMF06)
plot(modeloMF06,4)
model(modeloMF06)
dic(modeloMF06)
View(modeloMF06)

## Resumen de resultados del modelo de mixtura finita 06 (pam ; v=10^6)
summary(modeloMF06$fit$beta)
HPDinterval(as.mcmc(modeloMF06$fit$beta))
summary(modeloMF06$refit$beta)
HPDinterval(as.mcmc(modeloMF06$refit$beta))
```

 # Comparación de modelos

dic=data.frame(dic(modelogeneral),dic(modeloMF01),dic(modeloMF02),dic(modeloMF03),
dic(modeloMF04),dic(modeloMF05),dic(modeloMF06))
dic

Apéndice B

Programa Computacional para la Aplicación

B.1. Código para Aplicación usando el Paquete effectFusion

El código computacional para la implementación del paquete effectFusion del software R en la aplicación del conjunto de datos reales es el siguiente:

 # APLICACIÓN DE MODELO CON 08 COVARIABLES

 # Lectura de data
library(haven)
Trim_Ene_Feb_Mar20=read_sav("G:/Mi unidad/.../Trim Ene-Feb-Mar20.sav")

 # Selección de variables dependiente e independientes
y=matrix(Trim_Ene_Feb_Mar20$ingtot)
x=data.frame(var1=factor(Trim_Ene_Feb_Mar20$p103),
var2=factor(Trim_Ene_Feb_Mar20$p108),var3=factor(Trim_Ene_Feb_Mar20$p109a),
var4=factor(Trim_Ene_Feb_Mar20$P206AA),var5=factor(Trim_Ene_Feb_Mar20$p210),
var6=factor(Trim_Ene_Feb_Mar20$p222),var7=factor(Trim_Ene_Feb_Mar20$P224),
var8=factor(Trim_Ene_Feb_Mar20$P225))

x$var2=cut(as.numeric(x$var2),breaks=c(0,20,30,40,50,60,70,100),labels=c(1,2,3,4,5,6,7),
right=FALSE,include.lowest=TRUE)

 # Resumen inicial de variables dependiente e independientes
summary(y)
summary(x)

 # Eliminación de NaN y valores atípicos
y[y==0]=NA
z=data.frame(y,x)
View(z)
summary(z)

```
nan=sapply(z, function(x) sum(is.na(x)))
z=na.omit(z)
View(z)

# Resumen final de variables dependiente e independientes
summary(z)

# Definición de inputs (logaritmo a variable dependiente)
y=matrix(log(z$y))
x=data.frame(var1=factor(z$var1),var2=factor(z$var2),var3=factor(z$var3),var4=factor(z$var4),
var5=factor(z$var5),var6=factor(z$var6),var7=factor(z$var7),var8=factor(z$var8))
types=c("n","o","o","n","n","n","n","n")
data=list(y=y,x=x,types=types)
summary(data)
View(data)

# Resumen final de variables dependiente (logaritmo) e independientes
summary(y)
summary(x)

# Definición de modelo general (method = NULL)

## Modelo general (method = NULL)
library(coda)
library(effectFusion)
modelogeneral=effectFusion(y,x,types,method=NULL,mcmc=c(M=30000,burnin=15000))
print(modelogeneral)
summary(modelogeneral)
plot(modelogeneral,8)
dic(modelogeneral)
View(modelogeneral)

## Resumen de resultados del modelo general (method = NULL)
summary(modelogeneral$fit$beta)
HPDinterval(as.mcmc(modelogeneral$fit$beta))

# Definición del modelo de mixtura finita (method = 'pam')

## Modelo de mixtura finita 01 (pam ; v=10)
modeloMF01=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=10),mcmc=c(M=30000,burnin=15000))
print(modeloMF01)
```

```
summary(modeloMF01)
plot(modeloMF01,8)
model(modeloMF01)
dic(modeloMF01)
View(modeloMF01)

## Resumen de resultados del modelo de mixtura finita 01 (pam ; v=10)
summary(modeloMF01$fit$beta)
HPDinterval(as.mcmc(modeloMF01$fit$beta))
summary(modeloMF01$refit$beta)
HPDinterval(as.mcmc(modeloMF01$refit$beta))

## Modelo de mixtura finita 02 (pam ; v=10^2)
modeloMF02=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=100),mcmc=c(M=30000,burnin=15000))
print(modeloMF02)
summary(modeloMF02)
plot(modeloMF02,8)
model(modeloMF02)
dic(modeloMF02)
View(modeloMF02)

## Resumen de resultados del modelo de mixtura finita 02 (pam ; v=10^2)
summary(modeloMF02$fit$beta)
HPDinterval(as.mcmc(modeloMF02$fit$beta))
summary(modeloMF02$refit$beta)
HPDinterval(as.mcmc(modeloMF02$refit$beta))

## Modelo de mixtura finita 03 (pam ; v=10^3)
modeloMF03=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=1000),mcmc=c(M=30000,burnin=15000))
print(modeloMF03)
summary(modeloMF03)
plot(modeloMF03,8)
model(modeloMF03)
dic(modeloMF03)
View(modeloMF03)

## Resumen de resultados del modelo de mixtura finita 03 (pam ; v=10^3)
summary(modeloMF03$fit$beta)
HPDinterval(as.mcmc(modeloMF03$fit$beta))
summary(modeloMF03$refit$beta)
```

```
HPDinterval(as.mcmc(modeloMF03$refit$beta))

## Modelo de mixtura finita 04 (pam ; v=10^4)
modeloMF04=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=10000),mcmc=c(M=30000,burnin=15000))
print(modeloMF04)
summary(modeloMF04)
plot(modeloMF04,8)
model(modeloMF04)
dic(modeloMF04)
View(modeloMF04)

## Resumen de resultados del modelo de mixtura finita 04 (pam ; v=10^4)
summary(modeloMF04$fit$beta)
HPDinterval(as.mcmc(modeloMF04$fit$beta))
summary(modeloMF04$refit$beta)
HPDinterval(as.mcmc(modeloMF04$refit$beta))

## Modelo de mixtura finita 05 (pam ; v=10^5)
modeloMF05=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=100000),mcmc=c(M=30000,burnin=15000))
print(modeloMF05)
summary(modeloMF05)
plot(modeloMF05,8)
model(modeloMF05)
dic(modeloMF05)
View(modeloMF05)

## Resumen de resultados del modelo de mixtura finita 05 (pam ; v=10^5)
summary(modeloMF05$fit$beta)
HPDinterval(as.mcmc(modeloMF05$fit$beta))
summary(modeloMF05$refit$beta)
HPDinterval(as.mcmc(modeloMF05$refit$beta))

## Modelo de mixtura finita 06 (pam ; v=10^6)
modeloMF06=effectFusion(y,x,types,method='FinMix',modelSelection='pam',
prior=list(p=1000000),mcmc=c(M=30000,burnin=15000))
print(modeloMF06)
summary(modeloMF06)
plot(modeloMF06,8)
model(modeloMF06)
dic(modeloMF06)
```

```
View(modeloMF06)

## Resumen de resultados del modelo de mixtura finita 06 (pam ; v=10^6)
summary(modeloMF06$fit$beta)
HPDinterval(as.mcmc(modeloMF06$fit$beta))
summary(modeloMF06$refit$beta)
HPDinterval(as.mcmc(modeloMF06$refit$beta))

# Comparación de modelos
dic=data.frame(dic(modelogeneral),dic(modeloMF01),dic(modeloMF02),dic(modeloMF03),
dic(modeloMF04),dic(modeloMF05),dic(modeloMF06))
dic
```

Bibliografía

Malsiner-Walli, G., Pauger, D. y Wagner, H. (2018). Effect fusion using model-based clustering, Statistical Modelling 18(2): 175-196.

Pauger D., Leitner M., Wagner H. y Malsiner-Walli G. (2019). effectFusion: Bayesian Effect Fusion for Categorical Predictors. R package version 1.1.2.
https://CRAN.R-project.org/package=effectFusion

Bouveyron, C., Celeux, G., Murphy, T. B. y Raftery. A. E. (2019). Model-Based Clustering and Classification for Data Science With Applications in R, Cambridge University Press.

Congdon, P. (2005). Bayesian Models for Categorical Data, Wiley.

Frühwirth-Schnatter, S., Celeux, G. y Robert, C. (2018). Handbook of Mixture Analysis, Chapman & Hall - CRC Press.

Kaufman, L. y Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons.

McLachlan, G. y Peel, D. (2000). Finite Mixture Models, Wiley.

Blei, D., Ng, A. y Jordan, M. (2003). Latent Dirichlet Allocation, Journal of Machine Learning Research 3: 993-1022.

Bondell H. y Reich J. (2009). Simultaneous factor selection and collapsing levels in ANOVA, Biometrics 65: 169–77.

Chipman H. (1996). Bayesian variable selection with related predictors, Canadian Journal of Statistics 1: 17–36.

Dunson D., Herring A. y Engel S. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes, Journal of the American Statistical Association 103: 534–46.

George E. y McCulloch R. (1993). Variable selection via Gibbs sampling, Journal of the American Statistical Association 88: 881–89.

George E. y McCulloch R. (1997) Approaches for Bayesian variable selection, Statistica Sinica 7: 339–73.

Gertheiss J., Hogger S., Oberhauser C. y Tutz G. (2011). Selection of ordinally scaled independent variables with application to international classication of functioning score sets, Journal of Royal Statistical Society Series C (60): 377–95.

Gertheiss J. y Tutz G. (2009). Penalized regression with ordinal predictors, International Statistical Review 77: 345–65.

Gertheiss J. y Tutz G. (2010) Sparse modelling of categorical explanatory variables. The Annals of Applied Statistics 4: 2150–80.

Griffin J. y Brown P. (2010) Inference with normal-gamma prior distributions in regression problems, Bayesian Analysis 5: 171–88.

Ishwaran H. y Rao J. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies, The Annals of Statistics 33: 730–73.

Kyung M., Gill J., Ghosh M. y Casella G. (2010). Penalized regression, standard errors, and Bayesian lasso, Bayesian Analysis 5, 369–412.

MacLehose R. y Dunson D. (2010). Bayesian semiparametric multiple shrinkage, Biometrics 66: 455–62.

Malsiner-Walli, G., Frühwirth-Schnatter, S. y Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures, Stat Comput 26: 303–324.

Malsiner-Walli, G., Frühwirth-Schnatter, S. y Grün, B. (2017). Identifying mixtures of mixtures using Bayesian estimation, Journal of Computational and Graphical Statistics 26: 285–95.

Malsiner-Walli G. y Wagner H. (2011). Comparing spike and slab priors for Bayesian variable selection, Austrian Journal of Statistics 40: 241–64.

Mitchell T. y Beauchamp J. (1988). Bayesian variable selection in linear regression, Journal of the American Statistical Association 83: 1023–32.

Park T. y Casella G. (2008). The Bayesian lasso, Journal of the American Statistical Association 103: 681–86.

Pauger, D. y Wagner, H. (2017). Bayesian effect fusion for categorical predictors, Bayesian Analysis 14(2): 341-369.

Raman S., Fuchs T., Wild P., Dahl E. y Roth V. (2009). The Bayesian group-lasso for analyzing contingency tables, Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009), 881–888.

Ronning, G. (1989). Maximum-likelihood estimation of dirichlet distributions. Journal of Statistical Computation and Simulation 32: 215–221.

Simon N., Friedman J., Hastie T. y Tibshirani R. (2013). A sparse-group lasso, Journal of Computational and Graphical Statistics 22: 231–45.

Spiegelhalter, D., Best, N., Carlin, B., y Van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit, Journal of the Royal Statistical Society, Series B 64(4): 583–639.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso, Journal of Royal

Statistical Society Series B (58): 267–88.

Tibshirani R., Saunders M., Rosset S., Zhu J. y Kneight K. (2005). Sparsity and smoothness via the fused lasso, Journal of Royal Statistical Society Series B (67): 91–108.

Tutz G. y Gertheiss J. (2016). Regularized regression for categorical data, Statistical Modelling 16: 161–200.

Yengo L., Jacques J. y Biernacki C. (2014). Variable clustering in high dimensional linear regression models, Journal de la Societe Francaise de Statistique 155: 38–56.

Yengo L., Jacques J., Biernacki C. y Canouil M. (2016). Variable clustering in highdimensional linear regression: The R package clere, The R Journal 8: 92–106.

Yuan M. y Lin Y. (2006). Model selection and estimation in regression with grouped variables, Journal of Royal Statistical Society Series B (68): 49–67.

Zou H. y Hastie T. (2005): Regularization and variable selection via the elastic net, Journal of Royal Statistical Society Series B (67): 301–20.