



**Universidad Nacional de Colombia**

Facultad de Ciencias  
Departamento de Estadística

## **Segundo caso de estudio**

### **Estadística Bayesiana**

Juan Andrés Camacho Zárate  
[jucamachoz@unal.edu.co](mailto:jucamachoz@unal.edu.co)  
Camilo Alejandro Raba Gomez  
[craba@unal.edu.co](mailto:craba@unal.edu.co)

**Docente**  
Juan Camilo Sosa Martínez

Febrero 2025

# Índice

<b>Prueba Saber 11 2022-2: Una perspectiva multinivel</b>	<b>3</b>
<b>Tratamiento de datos</b>	<b>4</b>
<b>Modelos</b>	<b>4</b>
$M_1$ : Modelo t con medias específicas por departamento . . . . .	4
$M_2$ : Modelo t con medias y varianzas específicas por departamento . . . . .	5
$M_3$ : Modelo t con medias específicas por municipio y departamento . . . . .	6
$M_4$ : Modelo t con medias específicas por municipio y departamento . . . . .	6
<b>Desarrollo metodológico</b>	<b>7</b>
<b>Preguntas</b>	<b>8</b>
Punto 1 . . . . .	8
Punto 2 . . . . .	8
Punto 3 . . . . .	9
Punto 4 . . . . .	10
Punto 5 . . . . .	11
Punto 6 . . . . .	12
Punto 7 . . . . .	13
Punto 8 . . . . .	15
Punto 9 . . . . .	17

Punto 10 . . . . .	17
Punto 11 . . . . .	18
Punto 12 . . . . .	19
Punto 13 . . . . .	20
Punto 14 . . . . .	22
Punto 15 . . . . .	23
Punto 16 . . . . .	24
Punto 17 . . . . .	26
<b>Apendice</b>	<b>27</b>
Distribuciones posteriores y condicionales . . . . .	27
Primer Modelo . . . . .	27
Segundo Modelo . . . . .	28
Tercer Modelo . . . . .	29
Cuarto Modelo . . . . .	29
Coeficientes de variación de Montecarlo . . . . .	31
Primer Modelo . . . . .	31
Segundo Modelo . . . . .	31
Tercer Modelo . . . . .	31
Cuarto Modelo . . . . .	32
<b>Referencias</b>	<b>32</b>

## Prueba Saber 11 2022-2: Una perspectiva multinivel

La base de datos **Saber 11 2022-2** contiene los resultados de la prueba **Saber 11 del segundo semestre de 2022**. Estos datos son de carácter público y pueden descargarse de forma gratuita a través de este [enlace](#).

Según la *Guía de Usuario del Examen Saber 11*, esta prueba estandarizada, administrada semestralmente por el Icfes, tiene como objetivos principales: servir como criterio de admisión para las Instituciones de Educación Superior, monitorear la calidad de la formación impartida en los establecimientos de educación media, y proporcionar información para estimar el valor agregado de la educación superior.

De acuerdo con la *Documentación del examen Saber 11*, esta prueba genera resultados a nivel individual para estudiantes próximos a finalizar la educación media. Los resultados incluyen puntajes obtenidos en cada una de las cinco pruebas genéricas: Matemáticas (M), Lectura (L), Ciencias (C), Sociales (S) e Inglés (I). Estos puntajes se presentan en una escala estándar definida desde la segunda aplicación del año 2014, con una media de 50 y una desviación estándar de 10. Esta normalización establece una línea de base y un punto de referencia para las estimaciones. Además, se calcula un puntaje global (PG) como un promedio ponderado de los puntajes en las cinco pruebas genéricas, según la fórmula:

$$PG = 5 \cdot \frac{5 \cdot M + 3 \cdot L + 3 \cdot C + 3 \cdot S + 1 \cdot I}{13}.$$

El puntaje global está diseñado para oscilar entre 0 y 500 puntos, con una media de 250 puntos y una desviación estándar de 50 puntos, proporcionando una medida integral del desempeño académico del evaluado.

El objetivo de este trabajo es ajustar modelos multinivel Bayesianos utilizando como datos de entrenamiento el **puntaje global** de los estudiantes, con el propósito de modelar los resultados de la prueba a nivel nacional por **municipio** y **departamento**. Específicamente, se busca:

- Establecer un *ranking* y una segmentación probabilística de los departamentos según su puntaje global promedio.
- Establecer un *ranking* y una segmentación probabilística de los municipios según su puntaje global promedio.
- Desarrollar un modelo predictivo para la **incidencia de la pobreza monetaria** a partir del puntaje global promedio por departamento.

- Desarrollar un modelo predictivo para la **cobertura neta secundaria** a partir del puntaje global promedio por municipio.

Este enfoque permitirá explorar patrones regionales en el desempeño académico y su relación con indicadores socioeconómicos clave.

## Tratamiento de datos

Para ajustar los modelos propuestos, se consideran exclusivamente los estudiantes que cumplen con los siguientes criterios:

- Nacionalidad colombiana.
- Residencia en Colombia.
- Proceso de investigación en el Icfes con estado “Publicar”.
- Colegio ubicado fuera del departamento de San Andrés.
- Sin datos faltantes en la ubicación del colegio por municipio, la ubicación del colegio por departamento y el puntaje global.

La base de datos resultante, tras aplicar estos filtros, contiene un total de **525,061 registros**. Se recomienda utilizar el diccionario de variables para garantizar la correcta aplicación de estos criterios.

## Modelos

### **M<sub>1</sub>: Modelo t con medias específicas por departamento**

**Distribución muestral:**

$$y_{i,j} \mid \theta_j, \sigma^2 \stackrel{\text{ind}}{\sim} t_v(\theta_j, \sigma^2),$$

donde  $t_v(\theta, \sigma^2)$  denota la distribución t con  $v$  grados de libertad con media  $\theta$ , para  $v > 1$ , y varianza  $\frac{v}{v-2} \sigma^2$ , para  $v > 2$ .

La variable aleatoria  $X$  tiene distribución t con parámetros  $v \in \mathbb{N}$ ,  $-\infty < \theta < \infty$ ,  $\sigma^2 > 0$ , i.e.,  $X | v, \theta, \sigma^2 \sim t_v(\theta, \sigma^2)$ , si su función de densidad de probabilidad es

$$p(x | v, \theta, \sigma^2) = \frac{1}{\sqrt{\pi v \sigma^2}} \frac{\Gamma((v+1)/2)}{\Gamma(v/2)} \left(1 + \frac{(x-\theta)^2}{v\sigma^2}\right)^{-(v+1)/2}, \quad -\infty < x < \infty.$$

Esta distribución es útil para modelar *outliers* y se encuentra implementada en el paquete `metRology` de R (ver <https://rdrr.io/cran/metRology/man/dt.scaled.html>).

Para ajustar este modelo de manera directa utilizando el muestreador de Gibbs, se debe tener en cuenta que la distribución muestral  $y_{i,j} | \theta_j, \sigma^2 \stackrel{\text{iid}}{\sim} t_v(\theta_j, \sigma^2)$  es equivalente a la distribución jerárquica dada por

$$y_{i,j} | \theta_j, \zeta_{i,j}^2 \stackrel{\text{ind}}{\sim} N(\theta_j, \zeta_{i,j}^2), \quad \zeta_{i,j}^2 | \sigma^2 \stackrel{\text{iid}}{\sim} GI\left(\frac{v}{2}, \frac{v\sigma^2}{2}\right),$$

donde las variables  $\zeta_{i,j}^2$  son variables auxiliares (variables latentes) cuyo objetivo es facilitar la implementación del muestreador de Gibbs.

Si no se consideran las variables  $\zeta_{i,j}^2$  en el modelo, la implementación del muestreador de Gibbs requeriría de otros métodos numéricos más sofisticados como el algoritmo de Metropolis-Hastings o el algoritmo de Monte Carlo Hamiltoniano, dado que la distribuciones condicionales completas tanto de  $\theta_j$  como  $\sigma^2$  no tendrían forma probabilística conocida.

Esta misma consideración acerca de las variables auxiliares se debe tener en cuenta para la implementación computacional de los modelos demás modelos.

### Distribución previa:

$$\begin{aligned} \theta_j | \mu, \tau^2 &\stackrel{\text{iid}}{\sim} N(\mu, \tau^2), & \mu &\sim N(\mu_0, \gamma_0^2), & \tau^2 &\sim GI\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right), \\ \sigma^2 &\sim G\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right), \end{aligned}$$

donde  $v, \mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu_0, \sigma_0^2$  son los hiperparámetros del modelo y  $GI(\alpha, \beta)$  denota la distribución Gamma-Inversa con media  $\frac{\beta}{\alpha-1}$ , para  $\alpha > 1$ , y varianza  $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ , para  $\alpha > 2$ .

## M<sub>2</sub>: Modelo t con medias y varianzas específicas por departamento

### Distribución muestral:

$$y_{i,j} | \theta_j, \sigma_j^2 \stackrel{\text{ind}}{\sim} t_v(\theta_j, \sigma_j^2).$$

**Distribución previa:**

$$\begin{aligned}\theta_j \mid \mu, \tau^2 &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), & \mu &\sim \mathcal{N}(\mu_0, \gamma_0^2), & \tau^2 &\sim \text{GI}\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right), \\ \sigma_j^2 \mid \nu, \sigma^2 &\sim \text{G}\left(\frac{\nu}{2}, \frac{\nu \sigma^2}{2}\right), & p(\nu) &\propto e^{-\lambda_0 \nu} & \sigma^2 &\sim \text{G}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right),\end{aligned}$$

donde  $v, \mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu, \alpha_0, \beta_0$  son los hiperparámetros del modelo y  $\text{G}(\alpha, \beta)$  denota la distribución Gamma con media  $\frac{\alpha}{\beta}$  y varianza  $\frac{\alpha}{\beta^2}$ . Además, el parámetro  $\nu$  está restringido a tomar valores en los números enteros positivos.

**Nota:** ¡Cuidado! La parametrización de la previa de  $\sigma^2$  es  $\text{G}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right)$  en lugar de  $\text{G}(\alpha_0, \beta_0)$ .

### **M<sub>3</sub>: Modelo t con medias específicas por municipio y departamento)**

**Distribución muestral:**

$$y_{i,j,k} \mid \zeta_{j,k}, \kappa^2 \stackrel{\text{ind}}{\sim} \text{t}_v(\zeta_{j,k}, \kappa^2),$$

para  $i = 1, \dots, n_{j,k}$ ,  $j = 1, \dots, n_k$  y  $k = 1, \dots, m$ , donde  $y_{i,j,k}$  es el puntaje global del estudiante  $i$  en el municipio  $j$  del departamento  $k$ .

**Distribución previa:**

$$\begin{aligned}\zeta_{j,k} \mid \theta_k, \sigma^2 &\stackrel{\text{ind}}{\sim} \mathcal{N}(\theta_k, \sigma^2), & \kappa^2 &\sim \text{G}\left(\frac{\xi_0}{2}, \frac{\xi_0 \kappa_0^2}{2}\right), \\ \theta_k \mid \mu, \tau^2 &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), & \mu &\sim \mathcal{N}(\mu_0, \gamma_0^2), & \tau^2 &\sim \text{GI}\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right), \\ \sigma^2 &\sim \text{GI}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right),\end{aligned}$$

donde  $v, \xi_0, \kappa_0^2, \mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu_0, \sigma_0^2$  son los hiperparámetros del modelo.

### **M<sub>4</sub>: Modelo t con medias específicas por municipio y departamento**

**Distribución muestral:**

$$y_{i,j,k} \mid \zeta_{j,k}, \kappa^2 \stackrel{\text{ind}}{\sim} \text{t}_v(\zeta_{j,k}, \kappa^2).$$

**Distribución previa:**

$$\zeta_{j,k} \mid \theta_k, \sigma_k^2 \stackrel{\text{ind}}{\sim} \mathcal{N}(\theta_k, \sigma_k^2), \quad \kappa^2 \sim \text{G}\left(\frac{\xi_0}{2}, \frac{\xi_0 \kappa_0^2}{2}\right),$$

$$\begin{aligned} \theta_k | \mu, \tau^2 &\stackrel{\text{iid}}{\sim} \mathsf{N}(\mu, \tau^2), & \mu &\sim \mathsf{N}(\mu_0, \gamma_0^2), & \tau^2 &\sim \mathsf{Gl}\left(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right), \\ \sigma_k^2 | \nu, \sigma^2 &\sim \mathsf{Gl}\left(\frac{\nu}{2}, \frac{\nu \sigma^2}{2}\right), & p(\nu) &\propto e^{-\lambda_0 \nu} & \sigma^2 &\sim \mathsf{G}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right), \end{aligned}$$

donde  $v, \xi_0, \kappa_0^2, \mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu, \alpha_0, \beta_0$  son los hiperparámetros del modelo.

## Desarrollo metodológico

Los modelos presentados se ajustan mediante **muestreadores de Gibbs** con un total de 110,000 iteraciones, desglosadas en 10,000 iteraciones iniciales correspondientes al periodo de calentamiento, las cuales se descartan para evitar el efecto de las condiciones iniciales en la inferencia. Posteriormente, para reducir la autocorrelación en la cadena de muestras, se emplea un muestreo sistemático con una amplitud de 10. De este modo, la cadena utilizada para realizar inferencias sobre la distribución posterior de los parámetros de cada modelo consta de  $B = 10,000$  iteraciones seleccionadas de manera espaciada.

Se emplean distribuciones previas difusas, definidas a partir de los siguientes hiperparámetros basados en la información de la prueba, para garantizar un enfoque no informativo que permita explorar las distribuciones posteriores de los parámetros de cada modelo:

- $M_1$ :  $v = 3, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu_0 = 1, \sigma_0^2 = 50^2$ .
- $M_2$ :  $v = 3, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \lambda_0 = 1, \alpha_0 = 1, \beta_0 = 1/50^2$ .
- $M_3$ :  $v = 3, \xi_0 = 1, \kappa_0^2 = 50^2, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu_0 = 1, \sigma_0^2 = 50^2$ .
- $M_4$ :  $v = 3, \xi_0 = 1, \kappa_0^2 = 50^2, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \lambda_0 = 1, \alpha_0 = 1, \beta_0 = 1/50^2$ .

Estas elecciones permiten ajustar los modelos de forma flexible, sin imponer restricciones fuertes sobre las distribuciones a priori, mientras reflejan la escala y características generales de los datos observados.

# Preguntas

1. En un gráfico con dos paneles ( $1 \times 2$ ), hacer un mapa de Colombia por **departamentos**, donde se desplieguen los valores de la media muestral del puntaje global (panel 1, izquierda) y la **incidencia de la pobreza monetaria en 2018** (panel 2, derecha). Interpretar los resultados obtenidos (máximo 100 palabras).

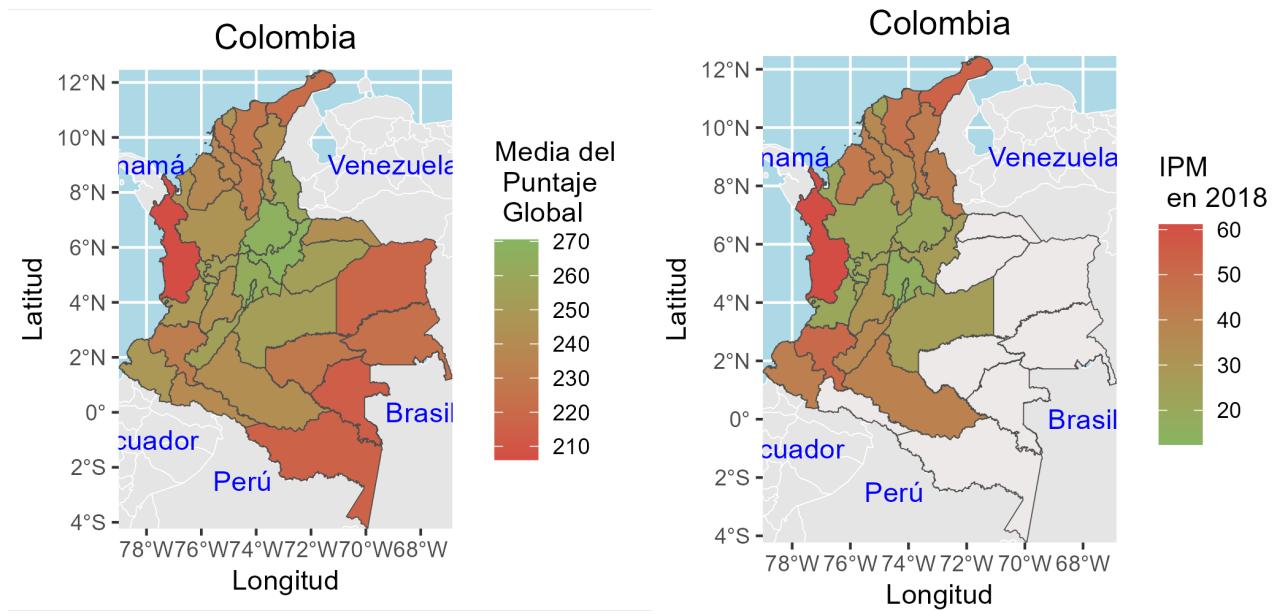


Figura 1: Media muestral e IPM en 2018 por departamento.

Se observa que las regiones del suroccidente, oriente y algunas del norte presentan los valores más bajos en el puntaje global y, a su vez, las mayores tasas de pobreza, evidenciadas en colores rojos intensos. En contraste, las zonas con puntajes más altos, como el centro del país, exhiben menores índices de pobreza. Esta relación sugiere una correlación negativa entre el desempeño en el puntaje global y la pobreza monetaria. Este hallazgo destaca la desigualdad regional en Colombia, donde el acceso a mejores condiciones socioeconómicas parece estar vinculado a un mejor desempeño en la prueba.

2. En un gráfico con dos paneles ( $1 \times 2$ ), hacer un mapa de Colombia por **municipios**, donde se desplieguen los valores de la media muestral del puntaje global (panel 1, izquierda) y

la **cobertura neta secundaria en 2022** (panel 2, derecha). Interpretar los resultados obtenidos (máximo 100 palabras).

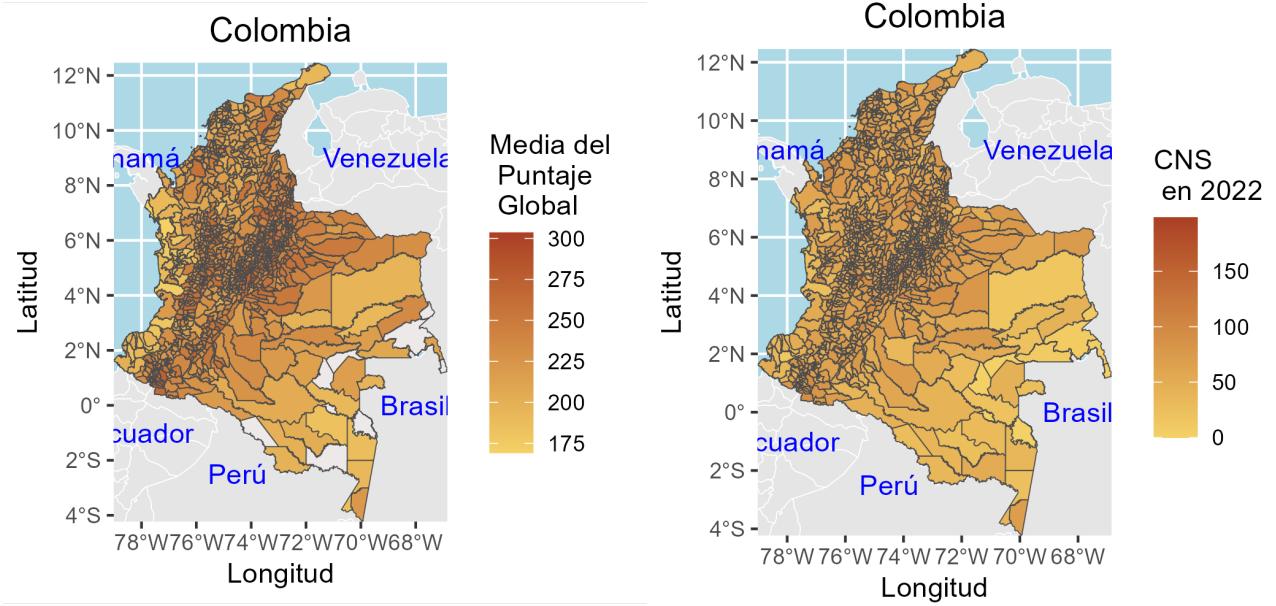


Figura 2: Media muestral y CNS en 2022 por municipio.

Al igual que en el análisis anterior, se observa que los municipios con mayores puntajes globales están concentrados en el centro del país, mientras que en las periferias los puntajes son más bajos. La cobertura neta secundaria sigue un patrón similar: las regiones con mejor desempeño en el puntaje global suelen tener mayores niveles de cobertura secundaria. Esto sugiere que la educación secundaria influye directamente en el desempeño académico. Municipios con menor cobertura secundaria, marcados en tonos más claros, coinciden con los de menor puntaje global, lo que refuerza la idea de desigualdad en el acceso y calidad educativa, afectando el rendimiento académico.

3. Hacer una tabla donde se presente el número de parámetros (incluyendo las variables auxiliares) y de hiperparámetros de cada modelo.

Modelo	Nº Parámetros	Nº V. AUX	Nº Hiperparametros
<b>M1</b>	35 $\{\theta_j, \sigma^2, \mu, \tau^2\}$	525061 $\{\zeta_{i,j}^2\}$	7 $\{v, \mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu_0, \sigma_0^2\}$
<b>M2</b>	68 $\{\theta_j, \sigma_j^2, \mu, \sigma^2, \tau^2, \nu\}$	525061 $\{\zeta_{i,j}^2\}$	8 $\{v, \mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu, \alpha_0, \beta_0\}$
<b>M3</b>	1148 $\{\zeta_{j,k}, \theta_k, \mu, \tau^2, \sigma^2, \kappa^2\}$	525061 $\{\xi_{i,j,k}\}$	9 $\{v, \xi_0, \kappa_0^2, \mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu_0, \sigma_0^2\}$
<b>M4</b>	1181 $\{\zeta_{j,k}, \theta_k, \sigma_k^2, \mu, \tau^2, \sigma^2, \kappa^2, \nu\}$	525061 $\{\xi_{i,j,k}\}$	10 $\{v, \xi_0, \kappa_0^2, \mu_0, \gamma_0^2, \eta_0, \lambda_0, \alpha_0, \beta_0, \tau_0^2\}$

Tabla 1: Número de parámetros, variables auxiliares e hiperparámetros por modelo

4. Demostrar que la distribución t se puede expresar como una mezcla de distribuciones Normales ponderadas por una distribución Gamma-Inversa. Es decir, demostrar que la distribución muestral  $y_i \mid \theta, \sigma^2 \stackrel{\text{iid}}{\sim} t_v(\theta, \sigma^2)$ , para  $i = 1, \dots, n$ , es equivalente a la distribución jerárquica dada por

$$y_i \mid \theta, V_i \stackrel{\text{ind}}{\sim} N(\theta, V_i), \quad V_i \mid \sigma^2 \stackrel{\text{iid}}{\sim} GI\left(\frac{v}{2}, \frac{v\sigma^2}{2}\right).$$

**Solucion:**

$$\begin{aligned} p(y_i \mid \theta, \sigma^2) &= \int_0^\infty p(y_i, V_i \mid \theta, \sigma^2) dV_i = \int_0^\infty p(y_i \mid \theta, V_i)p(V_i \mid \sigma^2) dV_i \\ &= \int_0^\infty \frac{\exp\left\{-\frac{1}{2V_i}(y_i - \theta)^2\right\}}{\sqrt{2\pi V_i}} \times \frac{\left(\frac{k\sigma^2}{2}\right)^{k/2}}{\Gamma(k/2)} V_i^{-(k/2+1)} \exp\left\{-\frac{k\sigma^2}{2V_i}\right\} dV_i \\ &= \frac{\left(\frac{k\sigma^2}{2}\right)^{k/2}}{\Gamma(k/2)\sqrt{2\pi}} \int_0^\infty V_i^{-\left(\frac{k+1}{2}+1\right)} \exp\left\{-\frac{(y_i - \theta)^2 + k\sigma^2}{2V_i}\right\} dV_i. \end{aligned}$$

Este integrando es el núcleo de una Gamma-Inversa.

Tenemos que si  $X \sim GI(\alpha, \beta)$ , entonces

$$\int_0^\infty x^{-\alpha+1} \exp\left\{-\frac{\beta}{x}\right\} dx = \frac{\Gamma(\alpha)}{\beta^\alpha},$$

En este caso, los parámetros son:

$$\alpha = \frac{k+1}{2}, \quad \beta = \frac{(y_i - \theta)^2 + k\sigma^2}{2}.$$

Luego,

$$p(y_i | \theta, \sigma^2) = \frac{\left(\frac{k\sigma^2}{2}\right)^{k/2}}{\Gamma(k/2)\sqrt{2\pi}} \times \Gamma\left(\frac{k+1}{2}\right) \times \left(\frac{(y_i - \theta)^2 + k\sigma^2}{2}\right)^{-\frac{k+1}{2}}.$$

Multiplicando por 1 y cancelando términos:

$$\begin{aligned} &= \frac{(k\sigma^2)^{k/2}\Gamma\left(\frac{k+1}{2}\right)}{\Gamma(k/2)\sqrt{\pi}} \times \frac{(k\sigma^2)^{1/2}}{(k\sigma^2)^{1/2}} \times \left(\frac{(y_i - \theta)^2 + k\sigma^2}{2}\right)^{-\frac{k+1}{2}} \\ &= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma(k/2)\sqrt{\pi k\sigma^2}} \left(1 + \frac{(y_i - \theta)^2}{k\sigma^2}\right)^{-\frac{k+1}{2}}. \end{aligned}$$

Por lo tanto, podemos concluir que la distribución marginal de  $y_i$  sigue una distribución  $t$  con  $k$  grados de libertad, media  $\theta$  y escala  $\sigma^2$ . Es decir:

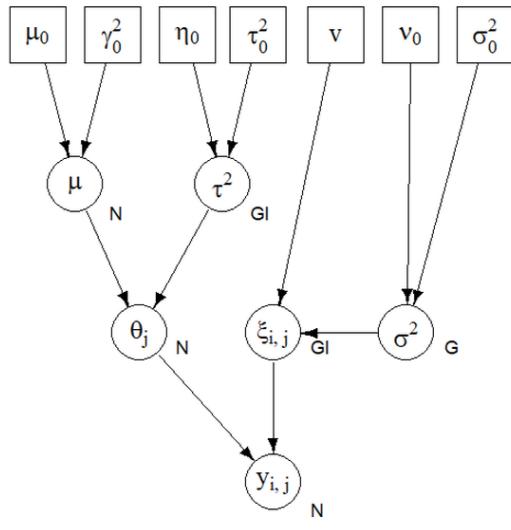
$$y_i | \theta, \sigma^2 \sim t_k(\theta, k\sigma^2).$$

Lo que concluye la prueba.

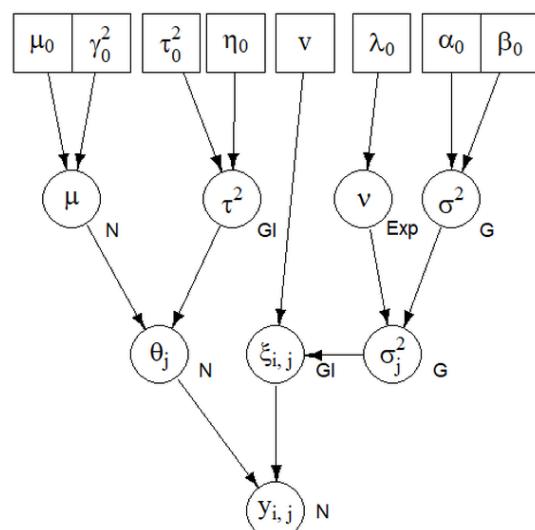
■

5. En un gráfico con cuatro paneles ( $2 \times 2$ ), hacer el DAG (incluyendo las variables auxiliares) de  $M_1$  (panel 1, esquina superior izquierda),  $M_2$  (panel 2, esquina superior derecha),  $M_3$  (panel 3, esquina inferior izquierda) y  $M_4$  (panel 4, esquina inferior derecha).

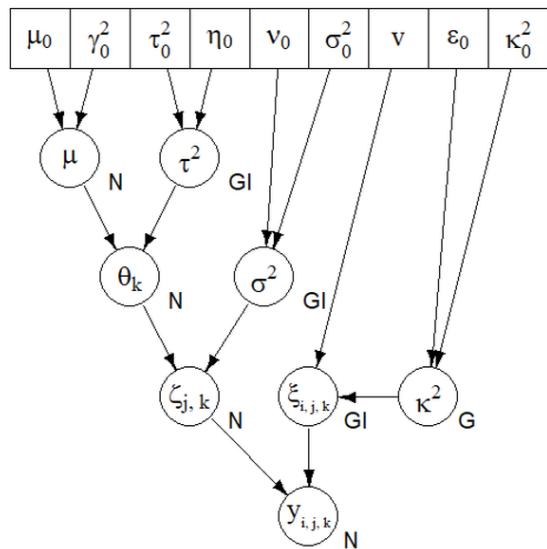
**Modelo 1: DAG**



**Modelo 2: DAG**



**Modelo 3: DAG**



**Modelo 4: DAG**

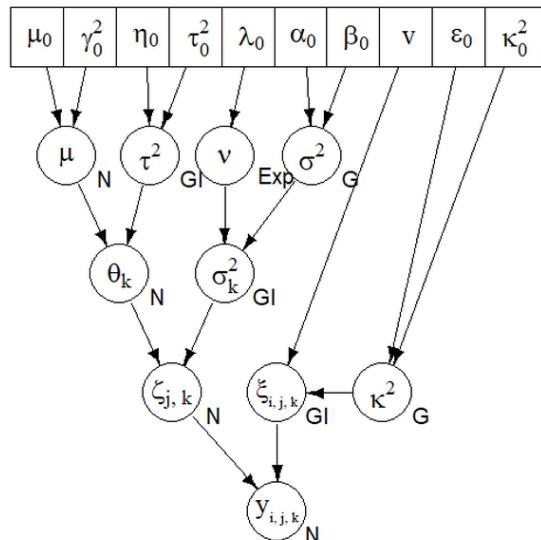


Figura 3: Grafos Acíclicos Dirigidos (DAG) de los modelos.

6. En un gráfico con cuatro paneles ( $2 \times 2$ ), dibujar la cadena de la log-verosimilitud de  $M_1$  (panel 1, esquina superior izquierda),  $M_2$  (panel 2, esquina superior derecha),  $M_3$  (panel

3, esquina inferior izquierda) y  $M_4$  (panel 4, esquina inferior derecha). No es necesario que los gráficos estén en la misma escala. Interpretar los resultados obtenidos (máximo 100 palabras).

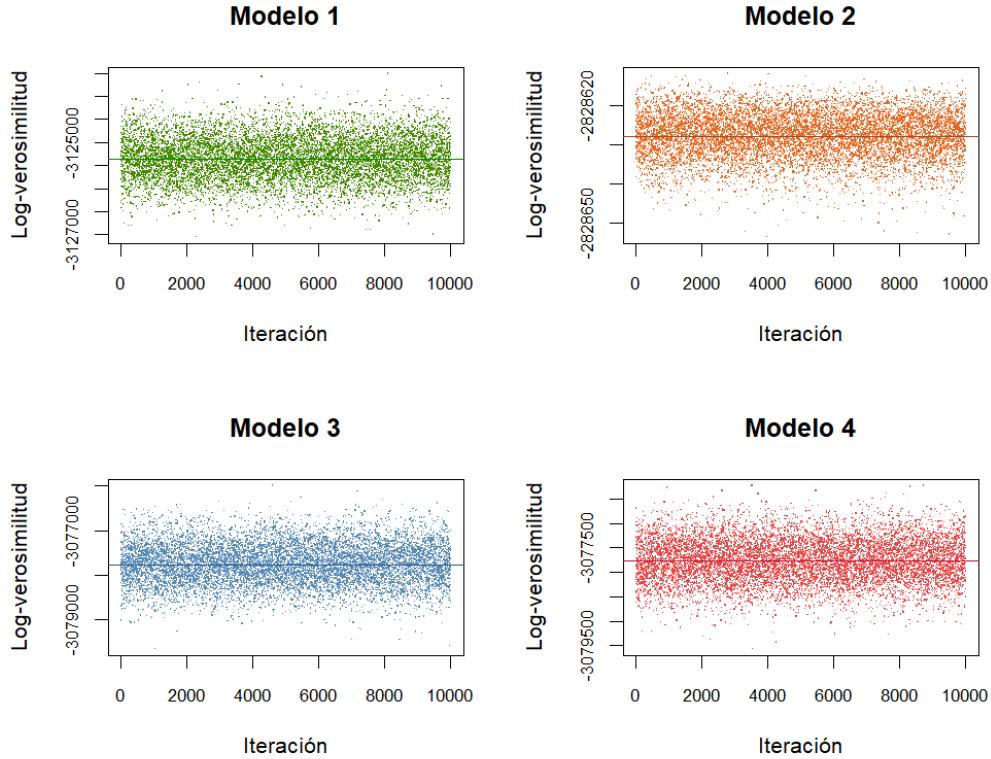


Figura 4: Cadenas de log-verosimilitud de los modelos.

Se observa que todas las cadenas presentan estabilidad, lo que sugiere convergencia. Sin embargo, hay diferencias en los valores de log-verosimilitud entre modelos:  $M_2$  tiene el valor más alto, lo que indica un mejor ajuste a los datos, mientras que  $M_3$  y  $M_4$  tienen los valores más bajos, sugiriendo un peor ajuste. Además, las fluctuaciones en cada cadena parecen mantenerse dentro de un rango estable, lo que indica una buena exploración del espacio de parámetros.

7. Para cada modelo, hacer un resumen de cinco números (mínimo, cuartil 1, mediana, media, cuaril 3 y máximo) del tamaño efectivo muestra de cada parámetro del modelo. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\theta_j$	9216	9496	9585	9648	9874	10179
$\sigma^2$			10080			
$\mu$			9667			
$\tau^2$			10000			

Tabla 2: Resumen de los tamaños efectivos de muestra del modelo 1.

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\theta_j$	9122	10000	10000	10091	10000	12523
$\sigma_j^2$	7069	7685	7868	7882	7966	8705
$\mu$			10000			
$\tau^2$			10370			
$\sigma^2$			9965			
$\nu$			10000			

Tabla 3: Resumen de los tamaños efectivos de muestra del modelo 2.

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\zeta_{j,k}$	6360	9553	10000	9779	10000	11790
$\theta_k$	9069	10000	10000	9967	10000	10621
$\mu$			10000			
$\tau^2$			10000			
$\sigma^2$			10000			
$\kappa^2$			10664			

Tabla 4: Resumen de los tamaños efectivos de muestra del modelo 3.

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\zeta_{j,k}$	6360	9553	10000	9779	10000	11790
$\theta_k$	9069	10000	10000	9967	10000	10621
$\sigma^2$	9385	10000	10000	9982	10000	10516
$\mu$			10045			
$\tau^2$			10000			
$\sigma^2$			10000			
$\kappa^2$			10000			
$\nu$			10000			

Tabla 5: Resumen de los tamaños efectivos de muestra del modelo 4.

Para la gran mayoría de los parámetros, los tamaños efectivos de muestra oscilan alrededor de 10,000, lo que indica una adecuada exploración de la distribución posterior. Sin embargo, se identifican dificultades en la exploración de los parámetros asociados a las medias por municipio  $\xi_{i,j}$  ya que en los modelos 3 y 4 sus valores mínimos caen por debajo de 7,000 muestras efectivas. Por otro lado, los valores para los 4 modelos de  $\mu$  y  $\tau^2$  sugieren que estos parámetros fueron bien explorados por la cadena de Gibbs, con baja autocorrelación y una buena convergencia.

8. Para cada modelo, hacer un resumen de cinco números (mínimo, cuartil 1, mediana, media, cuaril 3 y máximo) del error estándar de Monte Carlo de cada parámetro del modelo. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\theta_j$	0.001	0.003	0.003	0.005	0.006	0.017
$\sigma^2$			0.040			
$\mu$			0.004			
$\tau^2$			1.369			

Tabla 6: Resumen del error estandar de Montecarlo del modelo 1.

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\theta_j$	0.002	0.004	0.005	0.007	0.008	0.025
$\sigma_j^2$	0.123	0.256	0.312	0.431	0.485	1.603
$\mu$			0.037			
$\tau^2$			1.199			
$\sigma^2$			0.000			
$\nu$			0.033			

Tabla 7: Resumen del error estandar de Montecarlo del modelo 2.

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\zeta_{j,k}$	0.001	0.018	0.027	0.030	0.038	0.120
$\theta_k$	0.016	0.027	0.034	0.043	0.049	0.136
$\mu$			0.037			
$\tau^2$			1.180			
$\sigma^2$			0.140			
$\kappa^2$			0.003			

Tabla 8: Resumen del error estandar de Montecarlo del modelo 3.

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\zeta_{j,k}$	0.001	0.018	0.026	0.030	0.038	0.118
$\theta_k$	0.016	0.024	0.030	0.042	0.051	0.151
$\sigma_k^2$	0.377	0.524	0.727	1.141	1.180	8.760
$\mu$			0.037			
$\tau^2$			1.169			
$\sigma^2$			0.294			
$\kappa^2$			0.003			
$\nu$			0.019			

Tabla 9: Resumen del error estandar de Montecarlo del modelo 4.

La mayoría de los parámetros presentan valores bajos de error estándar, una señal de que la inferencia en general es estable y precisa. Sin embargo, los errores estándar de las varianzas

específicas por departamento en los modelos 2 y 4 son notablemente altos, lo que implica una mayor incertidumbre en la inferencia y estimación de estos parámetros. Además, el error estándar de las medias específicas por municipio muestra una alta dispersión (mínimos muy pequeños y máximos elevados), lo que sugiere que ciertas regiones del espacio de parámetros son más difíciles de explorar para el algoritmo.

9. Calcular el DIC y el WAIC de cada modelo. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Modelo	DIC	WAIC
M1	6250744.64	6250845.99
M2	5657319.50	5657316.24
M3	6156595.95	6159276.99
M4	6156617.54	6159294.73

Tabla 10: Valores de DIC y WAIC para los diferentes modelos

Los valores de DIC y WAIC indican la calidad del ajuste de cada modelo, donde menores valores reflejan un mejor equilibrio entre ajuste y complejidad. El modelo M2 presenta los valores más bajos, lo que sugiere que es el mejor modelo entre los evaluados. Por otro lado, M1 tiene los valores más altos, indicando el peor desempeño en términos de ajuste. Los modelos M3 y M4 tienen valores similares, sugiriendo un desempeño intermedio. Esto confirma la tendencia observada en las cadenas de log-verosimilitud, donde M2 también mostró los valores más altos, respaldando su selección como el modelo más adecuado.

10. Calcular la media posterior, el coeficiente de variación y el intervalo de credibilidad al 95% basado en percentiles de  $\mu$  de cada modelo. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Modelo	Media Posterior	CV (%)	L. Inf (2.5%)	L. Sup (97.5%)
M1	235.412	1.690	227.603	243.117
M2	238.408	1.573	231.138	245.774
M3	227.278	1.640	219.874	234.561
M4	227.226	1.620	219.968	234.584

Tabla 11: Media posterior, coeficiente de variacion e intervalos de credibilidad para  $\mu$ .

Los resultados muestran que el modelo M2 tiene la mayor media posterior (238.408) con el menor coeficiente de variación (1.573%), lo que sugiere mayor estabilidad y precisión en la estimación de  $\mu$ . Además, su intervalo de credibilidad al 95% es relativamente estrecho, lo que indica menor incertidumbre en la estimación. En contraste, los modelos M3 y M4 presentan valores de media más bajos y coeficientes de variación ligeramente superiores, sugiriendo menor precisión. M1 tiene un desempeño intermedio, con una media posterior de 235.412. Estos resultados refuerzan la idea de que el mejor modelo es el M2.

11. Usando  $M_4$ , hacer el *ranking* de los departamentos basado las medias específicas de los departamentos. Hacer una visualización del *ranking* Bayesiano. La visualización debe incluir simultáneamente las estimaciones puntuales y los intervalos de credibilidad al 95%. Interpretar los resultados obtenidos (máximo 100 palabras).

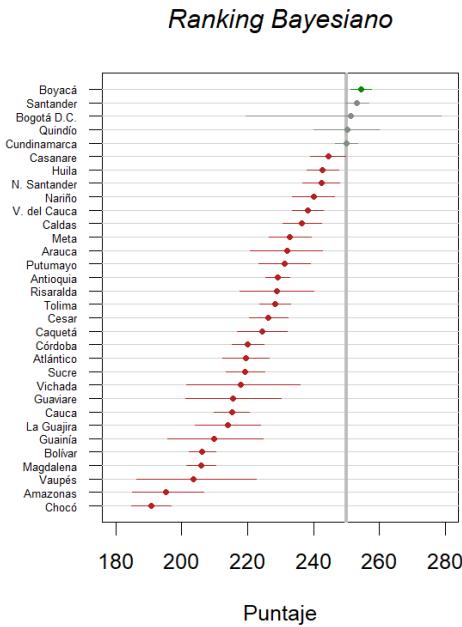


Figura 5: Ranking Bayesiano de los departamentos basado en las medias específicas.

El ranking bayesiano muestra la estimación puntual y el intervalo de credibilidad al 95% para cada departamento. Boyacá, Santander y Bogotá D.C. tienen los puntajes más altos, con menor incertidumbre (intervalos más estrechos), lo que sugiere un desempeño consistentemente superior. En contraste, Chocó, Amazonas y Vaupés presentan los puntajes más bajos, con intervalos más amplios, indicando mayor variabilidad en sus estimaciones. La tendencia general revela una marcada desigualdad entre regiones, donde

los departamentos con mayor desarrollo suelen tener mejores puntajes. Esta clasificación proporciona información clave para orientar políticas públicas y estrategias de mejora educativa en las regiones más rezagadas.

12. Usando  $M_4$ , hacer una segmentación de los departamentos usando las medias específicas de los departamentos, por medio del método de agrupamiento de  $K$ -medias (usar un método apropiado para seleccionar el número de grupos). Presentar los resultados obtenidos visualmente a través de una matriz de incidencia organizada a partir del *ranking* Bayesiano del numeral anterior y de un mapa que señale los departamentos que pertenecen al mismo grupo. Interpretar los resultados obtenidos (máximo 100 palabras).

**Nota:** Para obtener la matriz de incidencia, llevar a cabo la segmentación en cada iteración de la cadena de Markov asociada con  $M_4$ , y en cada iteración, establecer los departamentos que pertenecen al mismo grupo. Para llevar a cabo la visualización del mapa se recomienda utilizar una segmentación de las medias posteriores de las medias específicas de los departamentos.

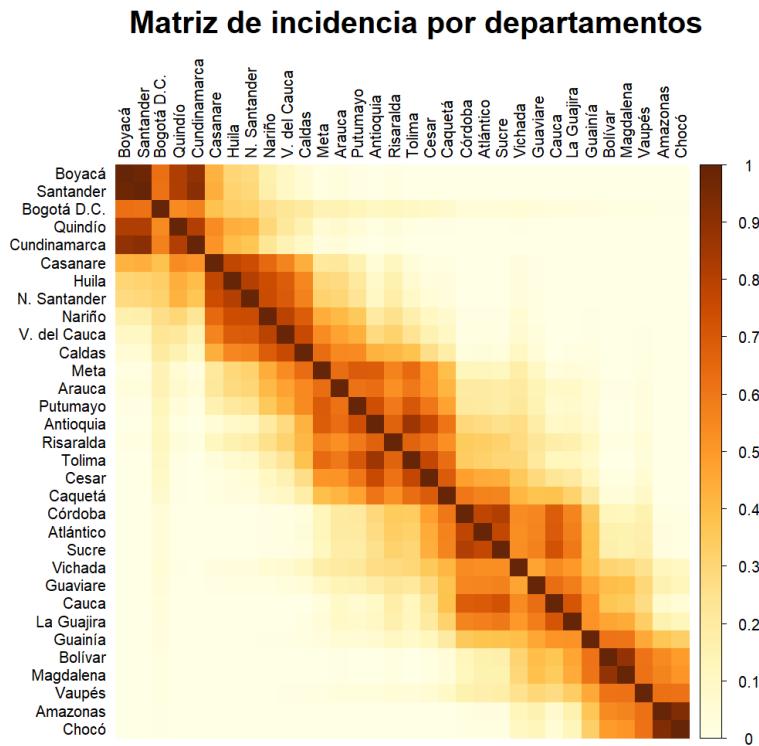


Figura 6: Matriz de incidencia por departamentos.

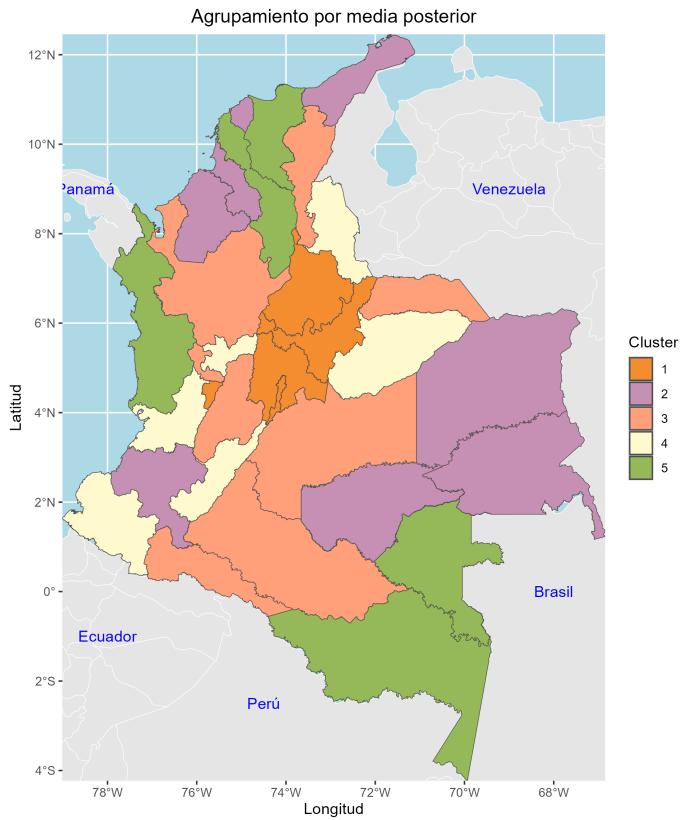


Figura 7: Mapa de agrupamiento de departamentos por media posterior.

En la matriz de incidencia se identifican cinco grupos distintos, algunos con una relación más marcada que otros. Los departamentos con puntajes globales más altos, ubicados en el centro del país, forman un grupo homogéneo con alta correlación, posiblemente debido a similitudes en desarrollo económico o densidad poblacional. Por otro lado, los departamentos periféricos también presentan una relación fuerte entre sí y están asociados a puntajes más bajos. Esta diferenciación sugiere contrastes significativos en acceso a servicios, condiciones socioeconómicas y dinámicas geográficas en las distintas regiones del país.

13. Calcular la media posterior y un intervalo de credibilidad al 95% de la **incidencia de la pobreza monetaria en 2018** (IPM) para todos los departamentos que no fueron medidos por el **DANE**, por medio de una regresión lineal simple de la IPM frente a las medias específicas de los departamentos de  $M_4$ . Presentar los resultados tabularmente (organizados descendente de acuerdo con la media posterior) y visualmente (por medio de un mapa usando la media posterior). Interpretar los resultados obtenidos (máximo 100 palabras).

Departamento	Media Posterior	L. Inf (2.5%)	L. Sup (97.5%)
Amazonas	51.422	45.133	57.421
Vaupés	47.149	36.917	56.559
Guainía	43.778	35.834	51.534
Guaviare	40.814	33.141	48.540
Vichada	39.555	30.134	48.481
Putumayo	32.623	28.449	36.775
Arauca	32.256	26.489	38.190
Casanare	25.691	22.558	28.721

Tabla 12: Media posterior e intervalos de credibilidad para los departamentos sin IPM en 2018.

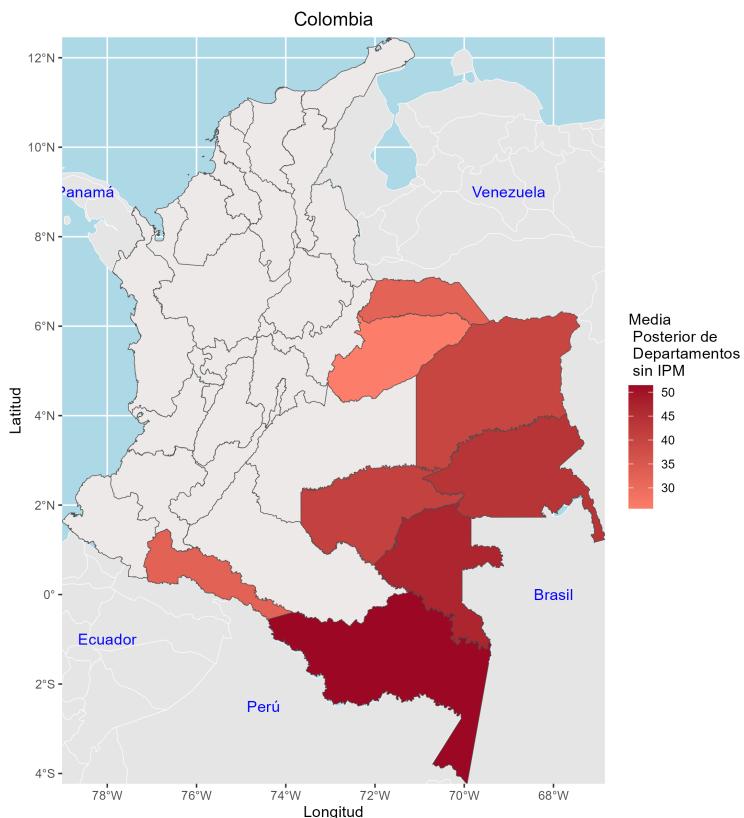


Figura 8: Media posterior de departamentos sin IPM en 2018.

Los resultados muestran las predicciones del IPM en 2018 para los departamentos no medidos por el DANE, utilizando una regresión en cada iteración de la cadena de Markov basada en M4. Amazonas presenta el mayor valor (51.42), seguido de Vaupés y Guainía,

indicando altos niveles de pobreza monetaria. En contraste, Casanare tiene la menor estimación (25.69), sugiriendo mejores condiciones económicas. El mapa confirma esta tendencia, con tonos más oscuros en regiones amazónicas y orientales, indicando mayor IPM. Estos hallazgos resaltan la vulnerabilidad de estos departamentos y la necesidad de políticas enfocadas en reducir la pobreza en estas zonas marginadas.

14. Usando  $M_4$ , hacer el *ranking* de los municipios basado las medias específicas de los municipios (no es preciso visualizar el *ranking* debido a la gran cantidad de municipios). Luego, hacer una segmentación de los municipios usando las medias específicas de los municipios, por medio del método de agrupamiento de  $K$ -medias (usar un método apropiado para seleccionar el número de grupos). Presentar los resultados obtenidos visualmente a través de una matriz de incidencia organizada a partir del *ranking* Bayesiano de los municipios obtenido inicialmente y de un mapa que señale los municipios que pertenecen al mismo grupo. Interpretar los resultados obtenidos (máximo 100 palabras).

**Matriz de incidencia por municipios**

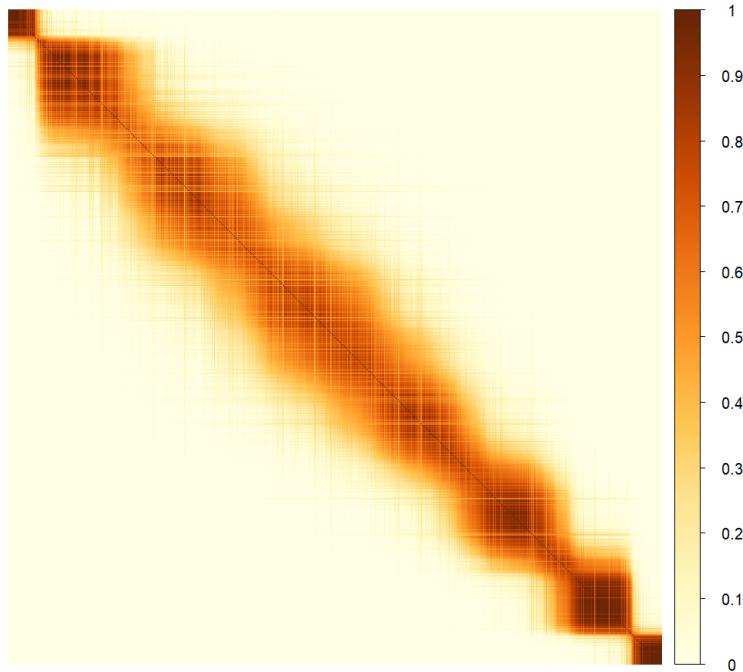


Figura 9: Matriz de incidencia por municipios.

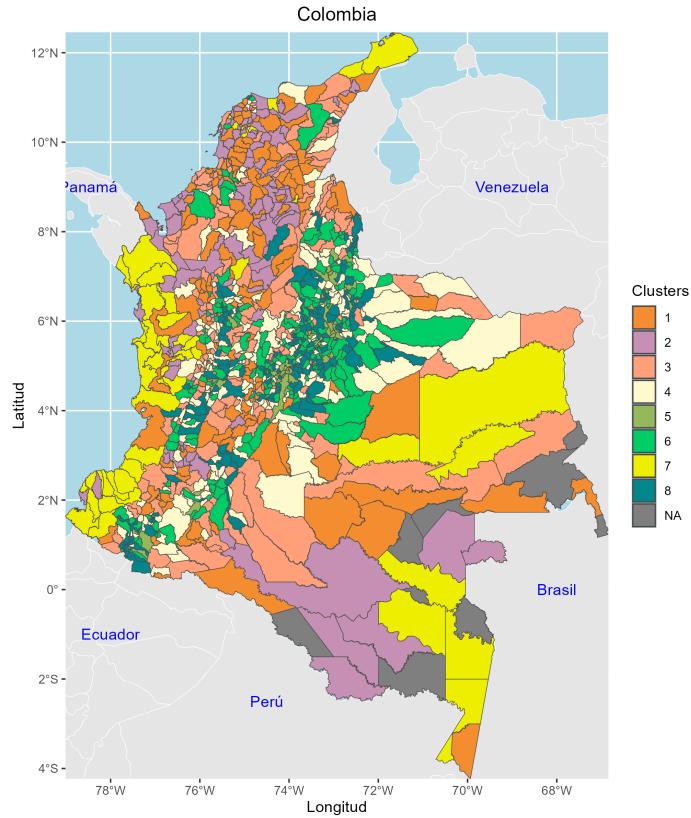


Figura 10: Agrupamiento de municipios por media posterior.

Se observa un patrón general, donde municipios cercanos tienden a pertenecer al mismo grupo, aunque algunos presentan características atípicas. En el centro y occidente, hay mayor heterogeneidad, mientras que, en regiones periféricas como Amazonas, Chocó y La Guajira, los municipios son más homogéneos y sus valores de puntaje global tienden a ser más bajos. Esto sugiere la influencia de factores como el desarrollo económico, acceso a servicios y condiciones geográficas, que generan diferencias marcadas con el resto del país y caracterizan fuertemente estas regiones.

15. Calcular la media posterior y un intervalo de credibilidad al 95% de la **cobertura neta secundaria en 2022** (CNS) para todos los municipios que no fueron medidos por el MEN, por medio de una regresión lineal simple de la CNS frente a las medias específicas de los municipios de  $M_5$ . Presentar los resultados tabularmente (organizados descendente de acuerdo con la media posterior) y visualmente (por medio de un mapa usando la media posterior). Interpretar los resultados obtenidos (máximo 100 palabras).

Municipio	Media Posterior	L. Inf (2.5%)	L. Sup (97.5%)
Belén de Bajirá	58.738	57.324	60.162
Mapiripana	55.916	52.271	59.568

Tabla 13: Media posterior e intervalos de credibilidad para los municipios sin CNS en 2022.

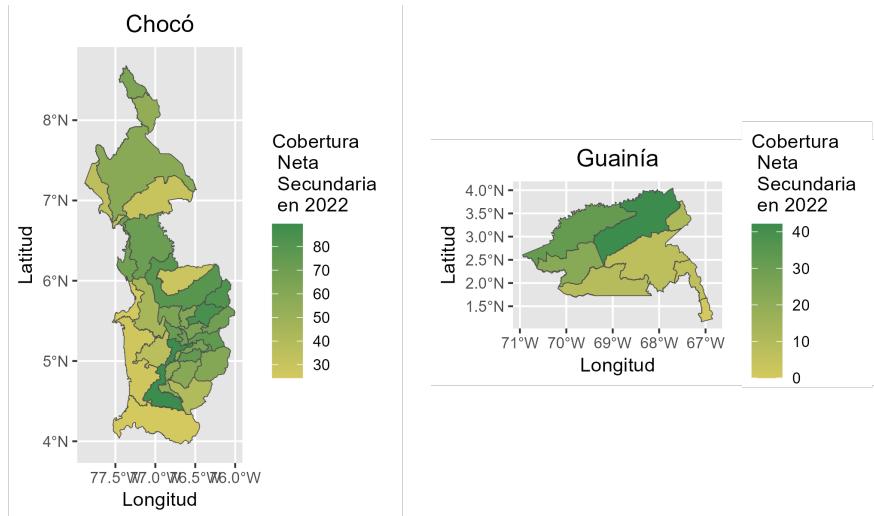


Figura 11: Media posterior en los departamentos con municipios sin CNS en 2022.

No fue posible encontrar los shapes de los dos municipios a los cuales se les imputó el valor de CNS para 2022. Belén de Bajirá ha estado en disputa territorial entre Chocó y Antioquia, lo que ha afectado su representación en bases de datos geográficas oficiales. Por su parte, Mapiripana, ubicado en Guainía, pertenece a un departamento con pocos municipios y extensas áreas rurales. Es posible que los datos geográficos disponibles lo incluyan dentro de una región más amplia, sin límites municipales bien definidos. No obstante, para contextualizar el análisis, se graficaron los departamentos a los cuales están asociados estos municipios.

Se observa que el valor de la media posterior para la Cobertura Neta Secundaria de Belén de Bajirá se asemeja al promedio de su departamento. Por otro lado, Mapiripana destaca dentro Guainía, ya que la predicción de su media posterior es considerablemente más alta en comparación con la mayoría de sus municipios vecinos.

16. Usando  $M_4$ , hacer un el top 5 de departamentos con mayor proporción de observaciones clasificadas como atípicas. Presentar los resultados tabularmente y visualmente por medio de un mapa. Interpretar los resultados obtenidos (máximo 100 palabras).

En el modelo jerárquico basado en la distribución  $t_v$ , las variables auxiliares  $\varsigma_{i,j}^2$  desempeñan un papel clave en la identificación de *outliers*. Estas variables representan varianzas locales específicas para cada observación  $y_{i,j}$ , reflejando la discrepancia de  $y_{i,j}$  con respecto a la tendencia general del modelo. Un criterio para identificar *outliers* es el siguiente: una observación  $y_{i,j}$  se clasifica como atípica si la media posterior de  $\varsigma_{i,j}^2$  excede el percentil 95 de la distribución de las medias posteriores de todas las  $\varsigma_{i,j}^2$ . Este enfoque aprovecha la información posterior para establecer un umbral adaptativo basado en las características generales del modelo y los datos.

<b>Departamento</b>	<b>Prop. Outliers (%)</b>
<b>Bolívar</b>	7.11
<b>Atlántico</b>	6.94
<b>Quindío</b>	6.34
<b>Sucre</b>	6.11
<b>Sandander</b>	6.06

Tabla 14: Top 5 departamentos con mayor proporción de observaciones atípicas (M4)

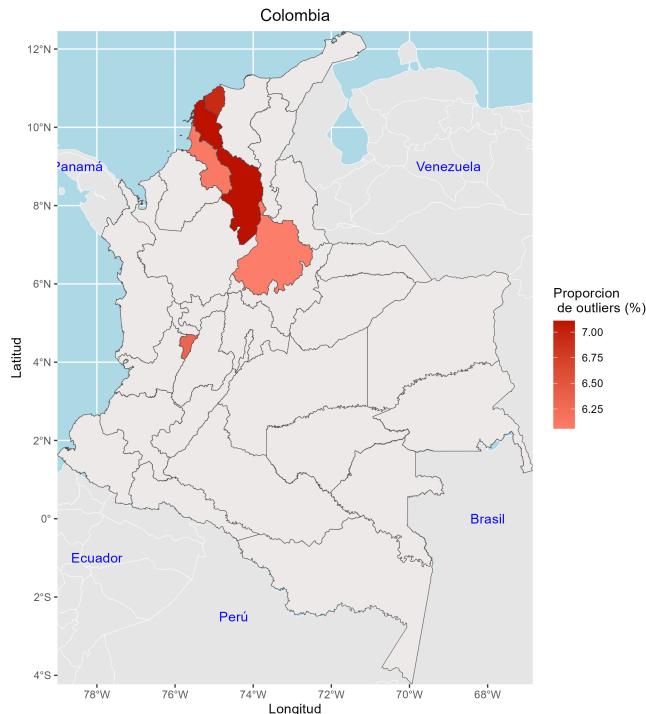


Figura 12: Top 5 departamentos con mayor proporción de observaciones atípicas (M4)

El análisis de outliers basado en M4 muestra que Bolívar (7.11%) y Atlántico (6.94%) presentan la mayor proporción de observaciones atípicas, seguidos por Quindío, Sucre y Santander. Estos resultados sugieren que en estos departamentos hay una mayor variabilidad o presencia de valores extremos en los datos analizados. Al analizar el mapa, se destaca que las regiones costeras presentan una mayor concentración de outliers. Este comportamiento puede deberse a diferencias estructurales en las condiciones socioeconómicas o en la calidad de los datos recolectados, lo que indica la necesidad de un análisis más profundo para identificar posibles causas subyacentes.

17. Validar la bondad ajuste de  $M_4$  por medio de la distribución predictiva posterior en cada municipio, utilizando como estadístico de prueba la media. Presentar los resultados visualmente. Interpretar los resultados obtenidos (máximo 100 palabras).

A continuación, validaremos la bondad de ajuste del modelo por municipios usando el modelo 4, para esto emplearemos como estadístico de prueba la media. Se calcularon los valores p predictivo posterior - PPP para la media en cada municipio y se obtuvieron los siguientes resultados:

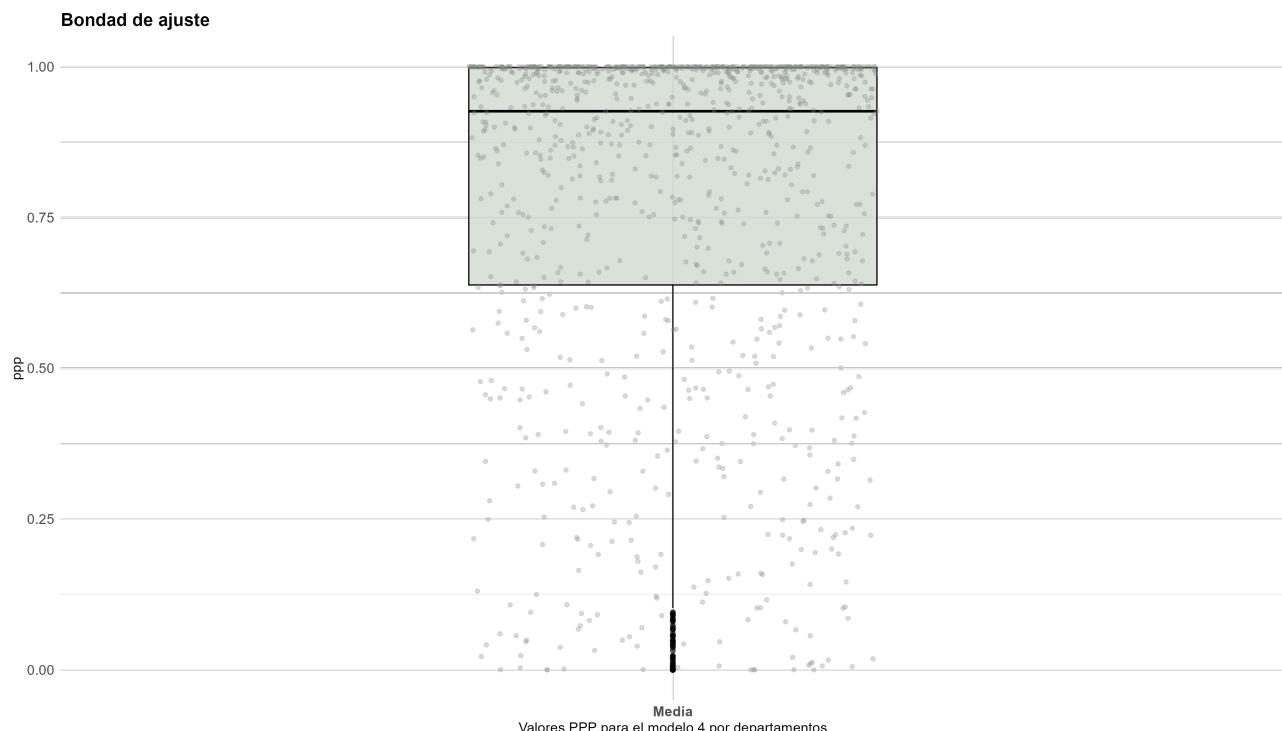


Figura 13: Bondad de Ajuste M4.

El análisis de los valores PPP muestra que el modelo M4 tiende a sobreajustar los datos y

a subestimar de la incertidumbre. Este sesgo puede afectar la capacidad del modelo para generalizar correctamente a nuevos datos. En un modelo bien ajustado, los valores PPP deberían distribuirse más uniformemente alrededor de 0.5. La revisión de los supuestos del modelo o una mayor flexibilidad en la estructura jerárquica podrían mejorar su desempeño y evitar este sobreajuste.

## Apendice

### Distribuciones posteriores y condicionales

#### Primer Modelo

##### Distribución Posterior:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \prod_j \prod_i \mathsf{N}(y_{ij} \mid \theta_j, \xi_{ij}^2) \cdot \mathsf{GI}\left(\xi_{ij}^2 \mid \frac{v}{2}, \frac{v\sigma^2}{2}\right) \times \\ \prod_j \mathsf{N}(\theta_j \mid \mu, \tau^2) \times \mathsf{G}\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \times \\ \mathsf{N}(\mu \mid \mu_0, \gamma_0^2) \times \mathsf{GI}\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right)$$

##### Condicionales completas:

$$\theta_j \mid resto \sim \mathsf{N}\left(\frac{\mu/\tau^2 + \sum_i y_{ij}/\xi_{ij}^2}{1/\tau^2 + \sum_i 1/\xi_{ij}^2}, \frac{1}{1/\tau^2 + \sum_i 1/\xi_{ij}^2}\right) \\ \xi_{ij}^2 \mid resto \sim \mathsf{GI}\left(\frac{v+1}{2}, \frac{(y_{ij} - \theta_j)^2 + v\sigma^2}{2}\right) \\ \mu \mid resto \sim \mathsf{N}\left(\frac{\sum_j \theta_j/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2}\right) \\ \tau^2 \mid resto \sim \mathsf{GI}\left(\frac{\eta_0 + m}{2}, \frac{\sum_j (\theta_j - \mu)^2 + \eta_0\tau_0^2}{2}\right) \\ \sigma^2 \mid resto \sim \mathsf{G}\left(\frac{nv + \nu_0}{2}, \frac{v \sum_{ij} 1/\xi_{ij}^2 + \nu_0\sigma_0^2}{2}\right)$$

## Segundo Modelo

Distribución Posterior:

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \tilde{\mathbf{y}}) \propto & \prod_j \prod_i \mathsf{N}(y_{ij} \mid \theta_j, \xi_{ij}^2) \cdot \mathsf{GI}\left(\xi_{ij}^2 \mid \frac{v}{2}, \frac{v\sigma_j^2}{2}\right) \times \\
& \prod_j \mathsf{N}(\theta_j \mid \mu, \tau^2) \times \mathsf{N}(\mu \mid \mu_0, \gamma_0^2) \times \\
& \mathsf{GI}\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right) \times \prod_j \mathsf{G}\left(\sigma_j^2 \mid \frac{\nu}{2}, \frac{\nu\sigma^2}{2}\right) \times \\
& p(v) \times \mathsf{G}\left(\sigma^2 \mid \frac{\alpha_0}{2}, \frac{\beta_0}{2}\right)
\end{aligned}$$

Concicionales completas:

$$\begin{aligned}
\theta_j \mid resto &\sim \mathsf{N}\left(\frac{\mu/\tau^2 + \sum_i y_{ij}/\xi_{ij}^2}{1/\tau^2 + \sum_i 1/\xi_{ij}^2}, \frac{1}{1/\tau^2 + \sum_i 1/\xi_{ij}^2}\right) \\
\mu \mid resto &\sim \mathsf{N}\left(\frac{\sum_j \theta_j/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2}\right) \\
\tau^2 \mid resto &\sim \mathsf{GI}\left(\frac{\eta_0 + m}{2}, \frac{\sum_j (\theta_j - \mu)^2 + \eta_0\tau_0^2}{2}\right) \\
\xi_{ij}^2 \mid resto &\sim \mathsf{GI}\left(\frac{v+1}{2}, \frac{(y_{ij} - \theta_j)^2 + v\sigma_j^2}{2}\right) \\
\sigma_j^2 \mid resto &\sim \mathsf{G}\left(\frac{n_j v + \nu}{2}, \frac{v \sum_{ij} 1/\xi_{ij}^2 + \nu\sigma^2}{2}\right) \\
\sigma^2 \mid resto &\sim \mathsf{G}\left(\frac{\alpha_0 + \nu m}{2}, \frac{\beta_0 + \nu \sum_j \sigma_j^2}{2}\right)
\end{aligned}$$

Para  $\nu$  se tiene que:

$$p(\nu) \propto \left(\frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma(\nu/2)}\right)^m \prod_j (\sigma_j^2)^{\nu/2} \exp\left\{-\nu \left(\lambda_0 + \frac{\sigma^2 \sum_j \sigma_j^2}{2}\right)\right\}$$

$$\ell(\nu) \propto \frac{\nu m}{2} \log\left(\frac{\nu\sigma^2}{2}\right) - m \log \Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2}\right) \sum_j \log(\sigma_j^2) - \nu \left(\lambda_0 + \frac{\sigma^2}{2} \sum_j \sigma_j^2\right)$$

## Tercer Modelo

Distribución Posterior:

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \tilde{\mathbf{y}}) \propto & \prod_{k,j,i} \mathsf{N}(y_{ijk} \mid \zeta_{jk}, \xi_{ijk}^2) \cdot \mathsf{GI}\left(\xi_{ijk}^2 \mid \frac{v}{2}, \frac{vk^2}{2}\right) \times \\
& \prod_{k,j} \mathsf{N}(\zeta_{jk} \mid \theta_k, \sigma^2) \times \mathsf{G}\left(k^2 \mid \frac{\epsilon_0}{2}, \frac{\epsilon_0 k_0^2}{2}\right) \times \\
& \prod_k \mathsf{N}(\theta_k \mid \mu, \tau^2) \times \mathsf{N}(\mu \mid \mu_0, \gamma_0^2) \times \\
& \mathsf{GI}\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right) \times \mathsf{GI}\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)
\end{aligned}$$

Condicionales Completas:

$$\begin{aligned}
\zeta_{jk} \mid resto &\sim \mathsf{N}\left(\frac{\sum_i y_{ijk}/\xi_{ijk}^2 + \theta_k/\sigma^2}{\sum_i 1/\xi_{ijk}^2 + 1/\sigma^2}, \frac{1}{\sum_i 1/\xi_{ijk}^2 + 1/\sigma^2}\right) \\
\xi_{ijk}^2 \mid resto &\sim \mathsf{GI}\left(\frac{v+1}{2}, \frac{(y_{ijk} - \zeta_{jk})^2 + v\kappa^2}{2}\right) \\
\theta_k \mid resto &\sim N\left(\frac{\sum_j \zeta_{j,k}/\sigma^2 + \mu/\tau^2}{n_k/\sigma^2 + 1/\tau^2}, \frac{1}{n_k/\sigma^2 + 1/\tau^2}\right) \\
\sigma^2 \mid resto &\sim \mathsf{GI}\left(\frac{\sum_k n_k + \nu_0}{2}, \frac{\sum_{j,k} (\zeta_{j,k} - \theta_k)^2 + \nu_0 \sigma_0^2}{2}\right) \\
\mu \mid resto &\sim \mathsf{N}\left(\frac{\sum_k \theta_k/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2}\right) \\
\tau^2 \mid resto &\sim \mathsf{GI}\left(\frac{\eta_0 + m}{2}, \frac{\sum_k (\theta_k - \mu)^2 + \eta_0 \tau_0^2}{2}\right) \\
\kappa^2 \mid resto &\sim \mathsf{G}\left(\frac{v \sum_k n_k + \epsilon_0}{2}, \frac{v \sum_{j,k,i} 1/\xi_{ijk}^2 + \epsilon_0 \kappa_0^2}{2}\right)
\end{aligned}$$

## Cuarto Modelo

Distribución Posterior:

$$p(\boldsymbol{\theta} \mid \tilde{\mathbf{y}}) \propto \prod_{k,j,i} \mathsf{N}(y_{ij,k} \mid \xi_{jk}, \zeta_{ijk}) \cdot \mathsf{GI}\left(\xi_{ij,k}^2 \mid \frac{v}{2}, \frac{vk^2}{2}\right) \times$$

$$\begin{aligned}
& \prod_{k,j} \mathsf{N}(\zeta_{j,k} | \theta_k, \sigma_k^2) \times \prod_k \mathsf{N}(\theta_k | \mu, \tau^2) \times \\
& \prod_k \mathsf{Gl}\left(\sigma_k^2 | \frac{\nu}{2}, \frac{\nu\sigma^2}{2}\right) \times \mathsf{N}(\mu | \mu_0, \tau_0^2) \times \\
& \mathsf{Gl}\left(\tau^2 | \frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right) \times p(\nu) \times \\
& \mathsf{G}\left(\sigma^2 | \frac{\alpha_0}{2}, \frac{\beta_0}{2}\right) \times \mathsf{G}\left(\kappa^2 | \frac{\epsilon_0}{2}, \frac{\kappa_0^2\epsilon_0}{2}\right)
\end{aligned}$$

**Condicionales Completas:**

$$\begin{aligned}
\xi_{ijk}^2 | resto &\sim \mathsf{Gl}\left(\frac{\nu + 1}{2}, \frac{\nu\kappa^2 + (y_{ijk} - \zeta_{j,k})^2}{2}\right) \\
\theta_k | resto &\sim \mathsf{N}\left(\frac{\mu/\tau^2 + \sum_j \zeta_{j,k}/\sigma_k^2}{1/\tau^2 + n_k/\sigma_k^2}, \frac{1}{1/\tau^2 + n_k/\sigma_k^2}\right) \\
\sigma_k^2 | resto &\sim \mathsf{Gl}\left(\frac{\nu + nk}{2}, \frac{\nu\sigma^2 + \sum_j (\zeta_{j,k} - \theta_k)^2}{2}\right) \\
\zeta_{j,k} | resto &\sim \mathsf{N}\left(\frac{\sum_i y_{ijk}/\xi_{ijk}^2 + \theta_k/\sigma_k^2}{\sum_i 1/\xi_{ijk}^2 + 1/\sigma_k^2}, \frac{1}{\sum_i 1/\xi_{ijk}^2 + 1/\sigma_k^2}\right) \\
\mu | resto &\sim \mathsf{N}\left(\frac{\sum_k \theta_k/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2}\right) \\
\tau^2 | resto &\sim \mathsf{Gl}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_k (\theta_k - \mu)^2}{2}\right) \\
\sigma^2 | resto &\sim \mathsf{G}\left(\frac{\alpha_0 + m\nu}{2}, \frac{\beta_0 + \nu \sum_k 1/\sigma_k^2}{2}\right) \\
\kappa^2 | resto &\sim \mathsf{G}\left(\frac{\epsilon_0 + n\nu}{2}, \frac{\epsilon_0\kappa_0^2 + \nu \sum_{jki} 1/\xi_{ijk}^2}{2}\right)
\end{aligned}$$

Para  $\nu$  se tiene que:

$$\begin{aligned}
p(\nu) &\propto \left(\frac{(\nu\sigma^2)^{\nu/2}}{\Gamma(\nu/2)}\right)^m \prod_k (\sigma_k^2)^{-\nu/2} \exp\left\{-\nu\left(\lambda_0 + \frac{\sigma^2}{2} \sum_k \frac{1}{\sigma_k^2}\right)\right\} \\
\ell(\nu) &\propto \frac{m\nu}{2} \ln\left(\frac{\nu\sigma^2}{2}\right) - m \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \left(\frac{\nu}{2}\right) \sum_k \ln(\sigma_k^2) - \nu\left(\lambda_0 + \frac{\sigma^2}{2} \sum_k \frac{1}{\sigma_k^2}\right)
\end{aligned}$$

## Coeficientes de variación de Montecarlo

### Primer Modelo

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\theta_j$	0.000%	0.001%	0.001%	0.002%	0.002%	0.008%
$\sigma^2$				0.001%		
$\mu$				0.017%		
$\tau^2$				0.264%		

### Segundo Modelo

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\theta_j$	0.001%	0.002%	0.002%	0.003%	0.003%	0.012%
$\sigma_j^2$	0.008%	0.016%	0.019%	0.030%	0.032%	0.110%
$\mu$				0.016%		
$\tau^2$				0.262%		
$\sigma^2$				0.069%		
$\nu$				0.240%		

### Tercer Modelo

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\zeta_{j,k}$	0.000%	0.008%	0.012%	0.013%	0.016%	0.048%
$\theta_k$	0.006%	0.012%	0.016%	0.019%	0.021%	0.054%
$\mu$				0.016%		
$\tau^2$				0.288%		
$\sigma^2$				0.045%		
$\kappa^2$				0.001%		

## Cuarto Modelo

Estadístico	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
$\zeta_{j,k}$	0.000%	0.008%	0.011%	0.013%	0.016%	0.051%
$\theta_k$	0.006%	0.011%	0.014%	0.019%	0.024%	0.060%
$\sigma_k^2$	0.129%	0.218%	0.272%	0.374%	0.382%	2.147%
$\mu$				0.016%		
$\tau^2$				0.283%		
$\sigma^2$				0.124%		
$\kappa^2$				0.001%		
$\nu$				0.279%		

## Referencias.

1. Sosa, J. (2024). Estadística Bayesiana. [Google Sites](#).
2. Martínez, D., & Peña, D. (2018). El nivel socioeconómico de los estudiantes y su efecto en la calidad medida por pruebas estandarizadas: Bogotá.
3. Victor, Abadía & Fabio, Cañavera & Chavez, Paula & Jose, Villadiego. (2024). Análisis del Impacto del Estrato Económico en los Resultados del ICFES.
4. Rodríguez, L. F., & Gómez, J. P. (2023). Análisis estadístico de los factores socioeconómicos que influyen en el rendimiento académico en Colombia. Revista de Estadística, 45(2), 123-145.
5. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian Data Analysis. CRC Press.