



Universidad Nacional de Colombia

FACULTAD DE CIENCIAS  
DEPARTAMENTO DE ESTADÍSTICA

ESTADÍSTICA BAYESIANA

*Parcial 2*

Alejandro Urrego López  
[aurrego@unal.edu.co](mailto:aurrego@unal.edu.co)

Noviembre 2023

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

La base de datos **Saber 11 2022-2.csv** que se encuentra disponible en la página web del curso, corresponde a los resultados de la prueba Saber 11 del segundo semestre de 2022. Los datos son de carácter público y se pueden descargar de manera gratuita en este enlace.

De acuerdo con la Guía de Usuario examen Saber 11, el examen Saber 11 “es una evaluación estandarizada realizada semestralmente por el Icfes, que tiene como objetivos: servir de criterio para la entrada de estudiantes a las Instituciones de Educación Superior, monitorear la calidad de la formación que ofrecen los establecimientos de educación media y producir información para la estimación del valor agregado de la educación superior.”

Siguiendo la Documentación del examen Saber 11, “este examen produce resultados a nivel individual de estudiantes que están próximos a culminar la educación media. Los resultados contienen puntajes del evaluado en cada una de las cinco pruebas genéricas (*Matemáticas, Lectura, Ciencias, Sociales, Inglés*) en una escala fijada en la segunda aplicación del año 2014 con promedio 50 y desviación estándar 10 (fijar la media y desviación estándar permite establecer una línea de base y tener un punto de referencia para las estimaciones) y un puntaje global, construido a partir de un promedio ponderado de los puntajes en las cinco pruebas genéricas”.

Así el puntaje global (PG) de la prueba se encuentra dado por

$$PG = \frac{5}{13} (5M + 3L + 3C + 3S + I),$$

donde  $M$ ,  $L$ ,  $C$ ,  $S$ , e  $I$  son los puntajes en las pruebas de Matemáticas, Lectura, Ciencias, Sociales, e Inglés, respectivamente. Por lo tanto, el puntaje global está diseñado de forma que asuma valores entre 0 puntos y 500 puntos, con una media de 250 puntos y una desviación estándar de 50 puntos.

El objetivo de este trabajo es ajustar modelos multinivel Bayesianos, tomando como datos de entrenamiento el puntaje global de los estudiantes, con el fin de modelar los resultados de la prueba a nivel nacional por municipio y departamento, para:

- 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 de la incidencia de la pobreza monetaria a partir del puntaje global promedio por departamento.
- Desarrollar un modelo predictivo de la cobertura neta secundaria a partir del puntaje global promedio por municipio.

## Tratamiento de Datos

Para ajustar los modelos propuestos, se consideran únicamente los estudiantes que cumplen con las siguientes condiciones:

- Nacionalidad colombiana.
- Residencia en Colombia.
- Proceso de investigación en el Icfes en estado de "Publicar".
- Ubicación del colegio no es 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 así conformada contiene 525,061 registros. Se utilizará el diccionario de variables para realizar este proceso.

# Modelos

## M1: Modelo Normal

### Distribución muestral

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

para  $i = 1, \dots, n_j$  y  $j = 1, \dots, m$ , donde  $y_{i,j}$  es el puntaje global del estudiante  $i$  en el departamento  $j$ , y  $N(\theta, \sigma^2)$  denota la distribución Normal con media  $\theta$  y varianza  $\sigma^2$ .

### Distribución previa

$$\theta \sim N(\mu_0, \gamma_0^2) \quad \sigma^2 \sim GI(\nu_0^2, \nu_0 \sigma_0^2)$$

donde  $\mu_0, \gamma_0^2, \nu_0, \sigma_0^2$  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  $\alpha > 2$ .

## M2: Modelo Normal con medias específicas por departamento

### Distribución muestral

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

### Distribución previa

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

donde  $\mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu_0$  y  $\sigma_0^2$  son los hiperparametros del modelo.

## M3: Modelo Normal con medias y varianzas específicas por departamento

### Distribución muestral

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

### Distribución previa

$$\begin{aligned} \theta_j \mid \mu, \tau^2 &\stackrel{\text{iid}}{\sim} N(\mu, \tau^2) & \mu &\sim N(\mu_0, \gamma_0^2) & \tau^2 &\sim GI(\eta_0^2, \eta_0 \tau_0^2) \\ \sigma_j^2 \mid \nu, \sigma^2 &\sim GI(\nu^2, \nu \sigma^2) & \nu &= \text{Constante} & \sigma^2 &\sim G(\alpha_0^2, \beta_0^2) \end{aligned}$$

donde  $\mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu \alpha_0$  y  $\beta_0$  son los parámetros del modelo y  $G$  es la distribución Gamma con media  $\frac{\alpha}{\beta}$  y varianza  $\frac{\alpha}{\beta^2}$ .

## M4: Modelo Normal con medias específicas por municipio y departamento

### Distribución muestral

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

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

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

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

## M5: Modelo Normal con medias específicas por municipio y departamento

## Distribución muestral

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

## Distribución previa

$$\begin{aligned} \zeta_{j,k} \mid \theta_k, \sigma_k^2 &\stackrel{\text{iid}}{\sim} N(\theta_k, \sigma_k^2) & \kappa^2 &\sim \text{GI}(\xi_0^2, \xi_0 \kappa_0^2) & \theta_k \mid \mu, \tau^2 &\stackrel{\text{iid}}{\sim} N(\mu, \tau^2) \\ \mu &\sim N(\mu_0, \gamma_0^2) & \tau^2 &\sim \text{GI}(\eta_0^2, \eta_0 \tau_0^2) & \sigma_k^2 \mid \nu, \sigma^2 &\sim \text{GI}(\nu^2, \nu \sigma^2) \\ \nu &= \text{Constante} & \sigma^2 &\sim G(\alpha_0^2, \beta_0^2). \end{aligned}$$

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

## Desarrollo Metodológico

Los modelos presentados anteriormente se ajustan mediante muestreadores de Gibbs con un total de  $1000 + 10 \cdot 10000 = 101000$  iteraciones. Las primeras 1000 iteraciones del algoritmo constituyen el período de calentamiento del algoritmo y no se tienen en cuenta para realizar inferencia.

Además, con el fin de reducir la autocorrelación de la cadena después del período de calentamiento, se realiza un muestreo sistemático de amplitud 10. Por lo tanto, la cadena utilizada para realizar inferencias acerca de la distribución posterior de los parámetros de cada modelo consta de  $B = 10000$  iteraciones.

Para estos ajustes, se emplean distribuciones previas difusas definidas por los siguientes hiperparámetros a partir de la información de la prueba:

- M1:  $\mu_0 = 250, \gamma_0^2 = 50^2, \nu_0 = 1, \sigma_0^2 = 50^2$ .
- M2:  $\mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu_0 = 1, \sigma_0^2 = 50^2$ .
- M3:  $\mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu = 1, \alpha_0 = 1, \beta_0 = \frac{1}{50^2}$ .
- M4:  $\xi_0 = 1, \kappa_0^2 = 50^2, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu_0 = 1, \sigma_0^2 = 50^2$ .
- M5:  $\xi_0 = 1, \kappa_0^2 = 50^2, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu = 1, \alpha_0 = 1, \beta_0 = \frac{1}{50^2}$ .

## Preguntas

### Punto 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).

**Nota:** El archivo *pobreza monetaria.xls* disponible en la página web del curso, contiene los datos de la incidencia de la pobreza monetaria de 23 departamentos y Bogotá 7 D.C en el periodo 2002-2018. Estos datos son de carácter público y se pueden descargar gratuitamente de la página web del DANE.

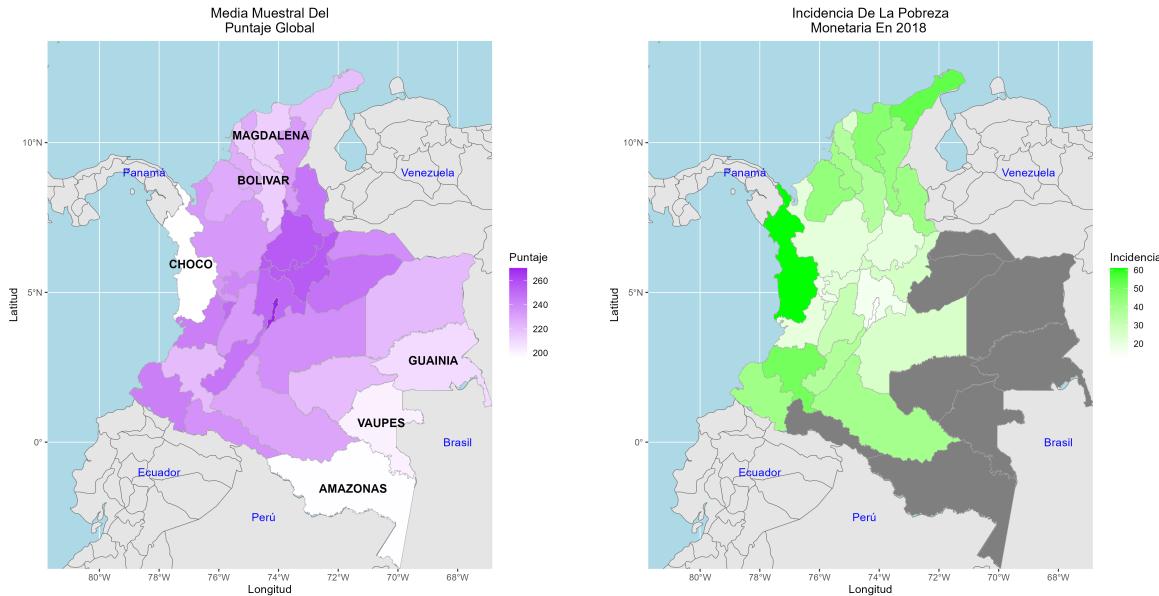


Figura 1: Puntaje global e incidencia de la pobreza monetaria

Podemos observar que la media muestral del puntaje global de los departamentos parece ser mayor en la región andina, mientras que en la región pacífica y la Amazonía tienen una media muestral del puntaje global menor. Hay que destacar que el departamento del Chocó tiene la media muestral del puntaje global más baja de todos los departamentos. Por otra parte, podemos observar que la región andina parece tener una incidencia de la pobreza monetaria en 2018. Además, destacan el departamento del Chocó y La Guajira con incidencias de la pobreza monetaria que sobresalen respecto a las demás.

## Punto 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).

**Nota:** El archivo estadísticas educación.xls disponible en la página web del curso, contiene estadísticas de los niveles preescolar, básica y media relacionada con indicadores sectoriales por municipio en el periodo 2011-2022. Estos datos son de carácter público y se pueden descargar gratuitamente de la página web del MEN en este enlace.

De acuerdo con el análisis anterior, podemos observar que los municipios con una media muestral mayor parecen estar en la región andina, mientras que la región amazónica, pacífica y caribe destacan por tener los puntos en el mapa con la menor media muestral. De igual manera, podemos observar que la cobertura neta secundaria destaca por su baja cobertura en estas mismas regiones.

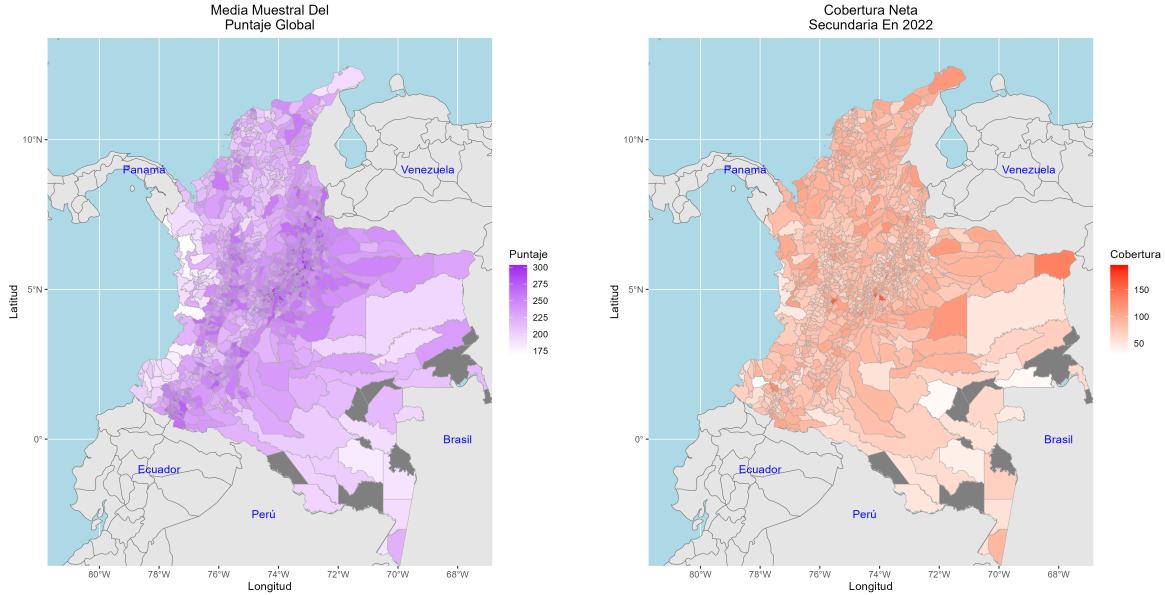


Figura 2: Puntaje global y cobertura neta secundaria en 2022

### Punto 3.

En un gráfico con cuatro paneles ( $2 \times 2$ ), hacer el DAG de  $\mathbf{M}_2$  (panel 1, esquina superior izquierda),  $\mathbf{M}_3$  (panel 2, esquina superior derecha),  $\mathbf{M}_4$  (panel 3, esquina inferior izquierda) y  $\mathbf{M}_5$  (panel 1, esquina inferior derecha)

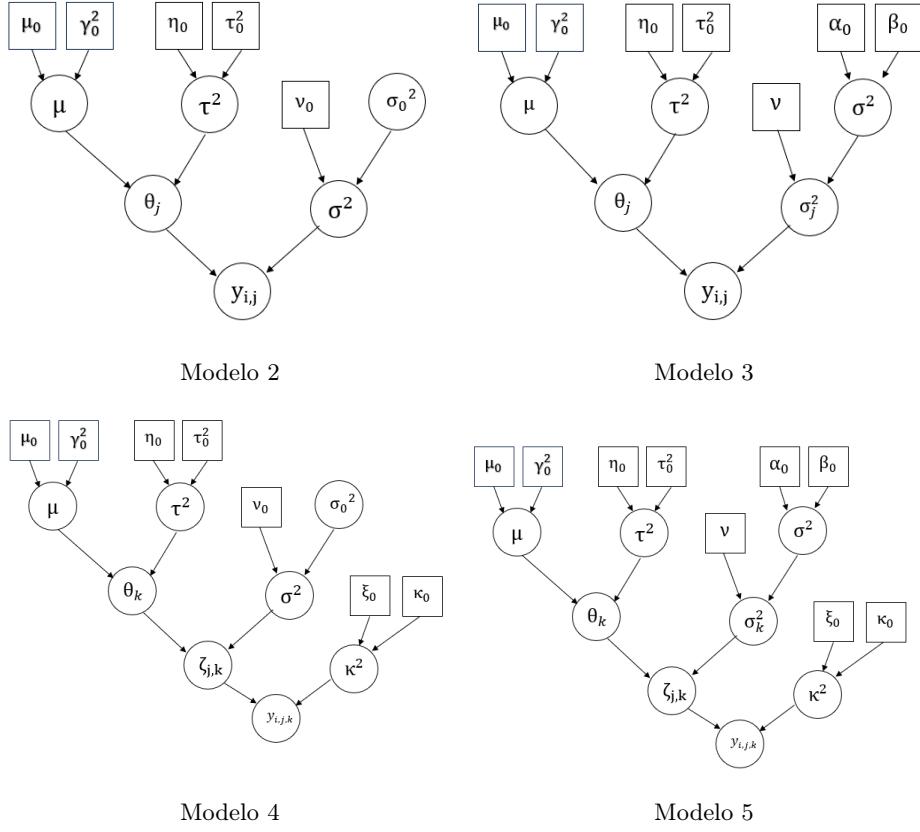


Figura 3: Gráficos DAG

## Punto 4.

En un gráfico con cuatro paneles ( $2 \times 2$ ), dibujar la cadena de la log-verosimilitud de  $M_2$  (panel 1, esquina superior izquierda),  $M_3$  (panel 2, esquina superior derecha),  $M_4$  (panel 3, esquina inferior izquierda) y  $M_5$  (panel 1, esquina inferior derecha). Los gráficos deben tener la misma escala para facilitar la comparación. Interpretar los resultados obtenidos (máximo 100 palabras).

**Nota:** Incluir un apéndice al final del informe con las distribuciones condicionales completas (no incluir la demostración, solo cada distribución con sus respectivos parámetros) y un resumen de los coeficientes de variación de Monte Carlo de cada parámetro de cada modelo. Tomar como ejemplo la Sección 4.3 del artículo Some Developments in Bayesian Hierarchical Linear Regression Modeling

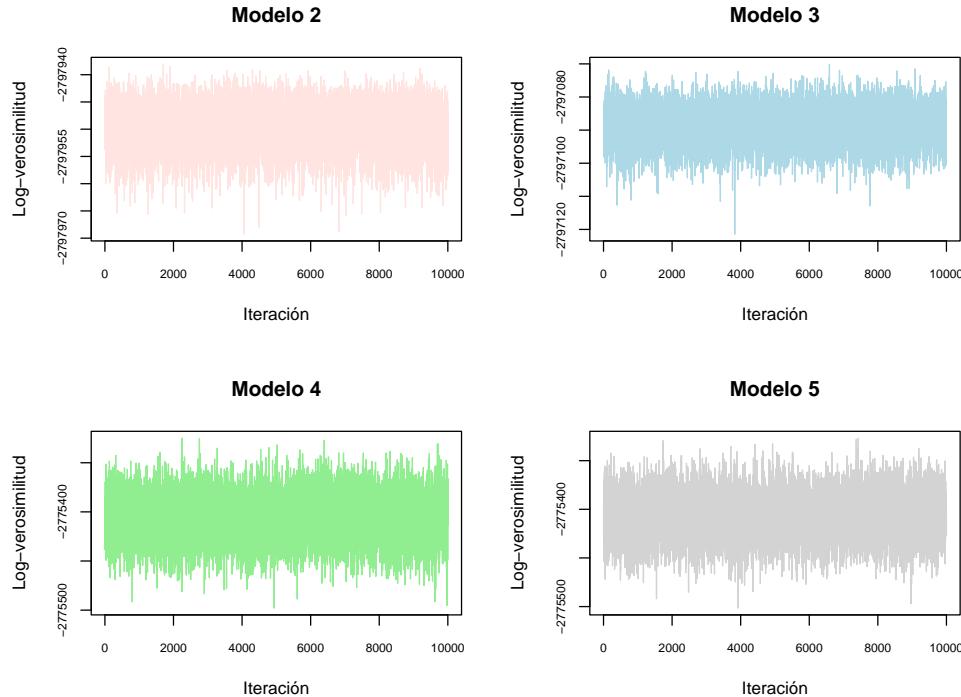


Figura 4: Log-verosimilitud

Al observar las cadenas de log-verosimilitud de los modelos 2, 3, 4 y 5, no hay evidencia para afirmar que no se ha alcanzado la convergencia global de los parámetros del modelo.

## Punto 5.

Calcular el DIC y el WAIC de cada  $M_k$ , para  $k = 1, \dots, 5$ . Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Modelo	DIC	WAIC
1	5,636,785	5,636,785
2	5,553,385	5,595,930
3	5,548,387	5,594,236
4	5,457,409.35	5,775,621
5	5,457,409.41	5,778,099

Tabla 1: Criterios de información.

Podemos observar que el modelo 4 tiene el menor Criterio de Información de la Devianza (DIC) y un menor Criterio de Información de Watanabe-Akaike (WAIC) de los 5 modelos. El modelo 5 no destaca sobre el modelo 4,

probablemente porque se está incluyendo a Bogotá en el análisis como un departamento con un único municipio. Esto tal vez podría mejorarse añadiendo localidades en lugar de municipios al momento de incluir a Bogotá en el modelo.

## Punto 6.

Calcular la media posterior y el intervalo de credibilidad al 95 % basado en percentiles de  $\mu$  de cada  $M_k$ , para  $k = 1, \dots, 5$ . Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

Modelo	Q2.5 %	Media	Q9.75 %
Modelo 1	250.135	250.273	250.410
Modelo 2	235.518	242.283	249.029
Modelo 3	235.542	242.241	249.076
Modelo 4	225.705	232.526	239.455
Modelo 5	225.767	232.545	239.239

Tabla 2: Resumen  $\mu$ .

De acuerdo con la forma en que se han planteado los modelos, podemos interpretar a  $\mu$  como la media global de los departamentos. Por lo tanto, según los modelos 2,3,4 y 5,  $\mu$  es significativamente menor a 250. Esto sugiere un bajo rendimiento en la prueba a nivel global..

## Punto 7.

Usando  $M_5$ , hacer el ranking de los departamentos basado las medias específicas de los departamentos. Comparar los resultados con un ranking frecuentista basado en la muestra dia muestral. En un gráfico con dos paneles ( $1 \times 2$ ), hacer la visualización del ranking Bayesiano (panel 1, izquierda) y el ranking frecuentista. Las visualizaciones deben incluir simultáneamente las estimaciones puntuales y los intervalos de credibilidad/confianza al 95 %. Interpretar los resultados obtenidos (máximo 100 palabras).

**Nota:** Usar la siguiente convención de colores: rojo oscuro para promedios significativamente inferiores a 250; negro para promedios que no difieren significativamente de 250; y verde oscuro para promedios significativamente superiores a 250.

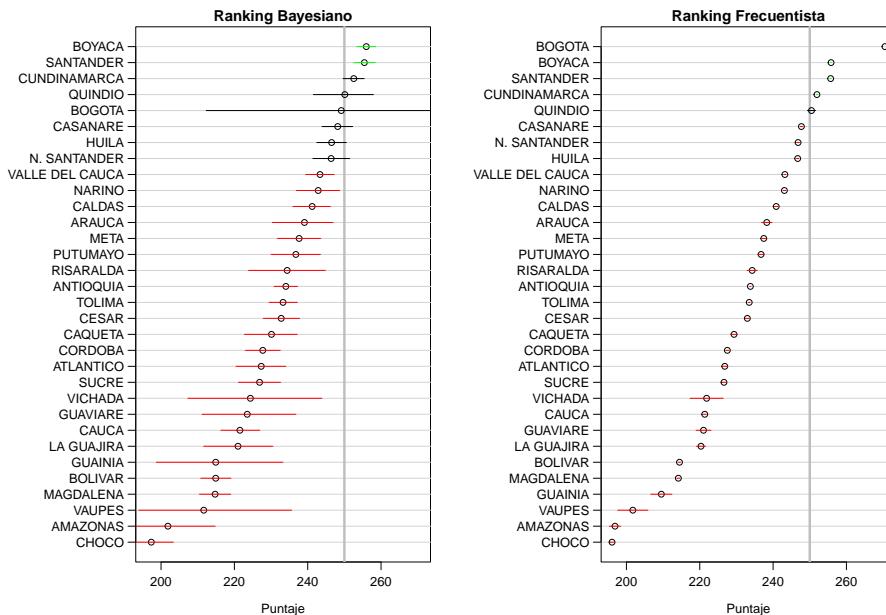


Figura 5: Ranking Departamentos

El ranking bayesiano indica que no existe evidencia suficiente para afirmar que al menos 8 departamentos están por debajo del promedio establecido en la prueba, mientras que el ranking frecuentista sugiere que son 5 departamentos. Esto demuestra que el enfoque bayesiano podría ser más eficaz para identificar departamentos con un desempeño inferior en comparación con otros, además de proporcionar información sobre la incertidumbre en relación al promedio de estos departamentos. Esta información es valiosa para tomar medidas y abordar este problema de manera más efectiva.

## Punto 8.

Usando  $M_5$ , hacer una segmentación de los departamentos usando las medias específicas de los departamentos, por medio del método de agrupamiento de K-medias con cinco 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:** Llevar a cabo la segmentación en cada iteración de la cadena de Markov asociada con  $M_5$ , y en cada iteración, establecer los departamentos pertenecen al mismo grupo. Tomar como ejemplo la Sección 7.5 de las notas de clase disponibles en este enlace. Para llevar a cabo la visualización del mapa se recomienda utilizar una segmentación de las medias posteriores de las medias específicas de los departamentos.

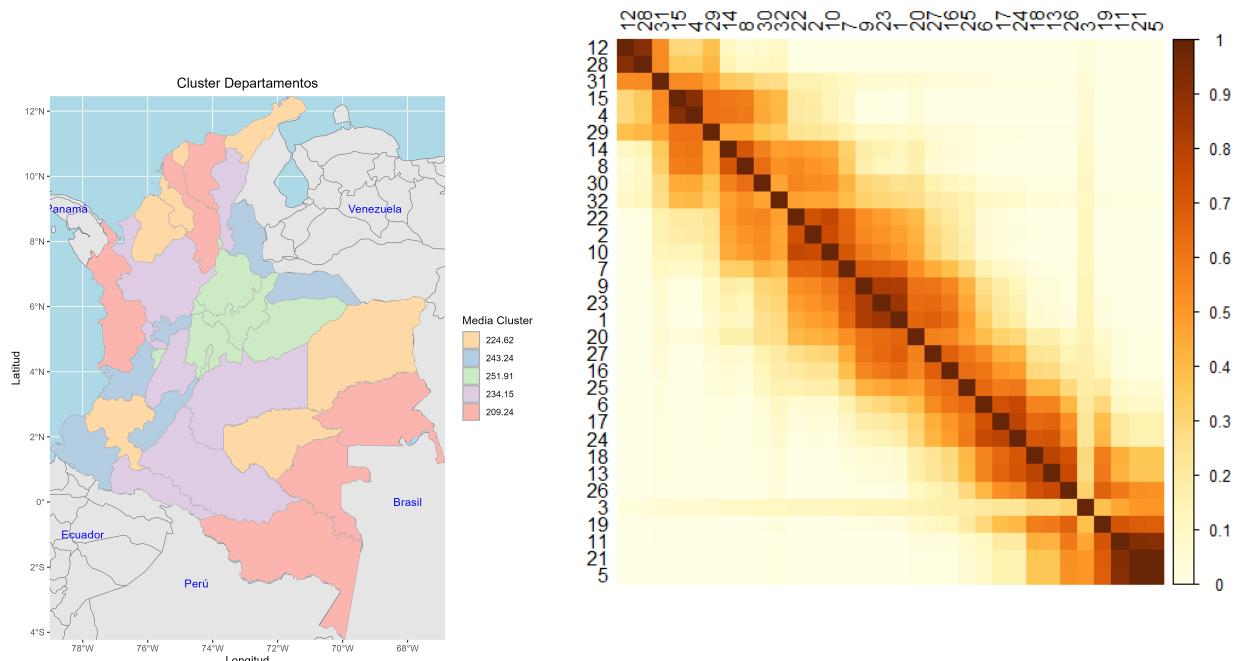


Figura 6: Segmentación departamentos

Al realizar la segmentación de los departamentos, podemos observar que aquellos con una media posterior menor se encuentran en los clusters rojo y amarillo, mientras que los departamentos ubicados en los clusters verde y azul tienen una media posterior mayor que los demás. Además, en el lado izquierdo, tenemos la matriz de incidencia que permite la formación de grupos de departamentos teniendo en cuenta cada iteración relacionada con  $M_5$ . Los grupos que se pueden formar están representados por los cuadrados que aparecen en esta matriz.

## Punto 9.

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  $M_5$ . Presentar los resultados tabularmente

(organizados descendente de acuerdo con la media posterior) y visualmente (por medio de un mapa usando la media posterior).

**Nota:** Llevar a cabo la regresión en cada iteración de la cadena de Markov asociada con  $M_5$ , y en cada iteración, hacer las predicciones de la IPM para los departamentos a los que halla lugar. Interpretar los resultados obtenidos (máximo 100 palabras).

Departamento	Media	Q2.5 %	Q97.5 %
Amazonas	52.74821	44.19476	60.61230
Vaupes	46.98228	32.83775	58.11086
Guainía	45.05898	34.16378	55.20538
Guaviare	39.98728	32.22978	47.72462
Vichada	39.47888	28.04725	49.81825
Putumayo	32.16099	28.09490	36.29749
Arauca	30.77843	26.01644	36.01081
Casanare	25.42365	22.45649	28.32395

Tabla 3: Estimaciones de la incidencia de la pobreza monetaria

Podemos observar que el departamento de Amazonas es el que tiene la media posterior más baja entre los departamentos que no fueron medidos, mientras que Casanare tiene la menor media posterior. Además, parece que en Vaupes es donde se encuentra la mayor variabilidad de todos, según el intervalo de credibilidad.

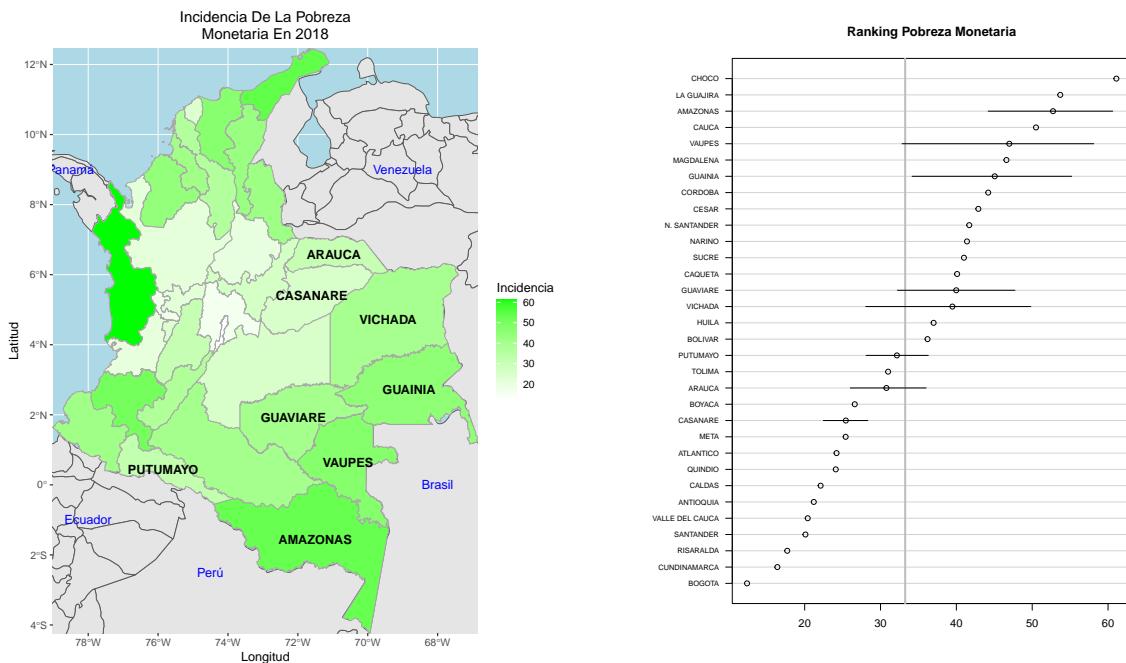


Figura 7: Estimaciones de la incidencia de la pobreza monetaria

## Punto 10.

Usando  $M_5$ , 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 con ocho 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).

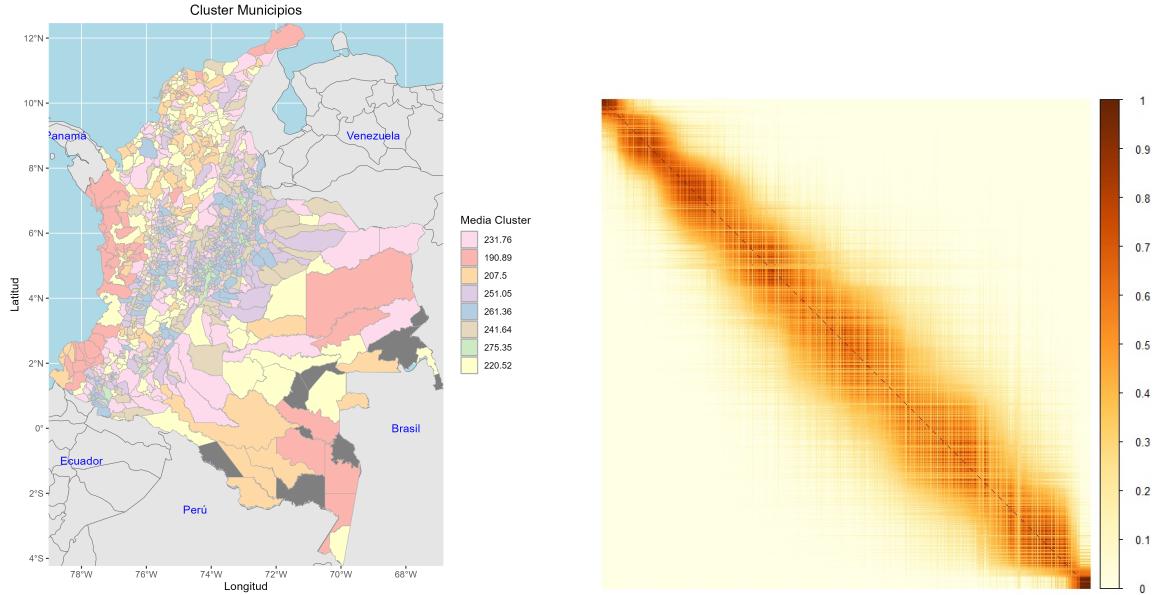


Figura 8: Segmentación municipios

Al realizar la segmentación, los departamentos con la media posterior más baja se encuentran en los clusters rojo, amarillo, naranja y rosa, mientras que aquellos con una mayor media posterior están en los clusters verde, azul, morado y café. Junto a esta información, encontramos la matriz de incidencia que nos permite conformar grupos teniendo en cuenta todas las iteraciones relacionadas con  $M_5$ . Estas segmentaciones nos ayudan a estudiar los departamentos de manera independiente y evaluar el desempeño de los municipios. Por ejemplo, en el departamento de Amazonas, observamos que hay 4municipios que pertenecen al cluster rojo, lo que sugiere que estos podrían requerir más atención al abordar este problema.

### Punto 11.

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).

Tabla 4: Datos municipales

Departamento	Municipio	Q0.025	Media	Q0.975
Guainía	Belén De Bajirá	79.27	80.64	82.35
Chocó	Mapiripana	78.82	83.41	85.69

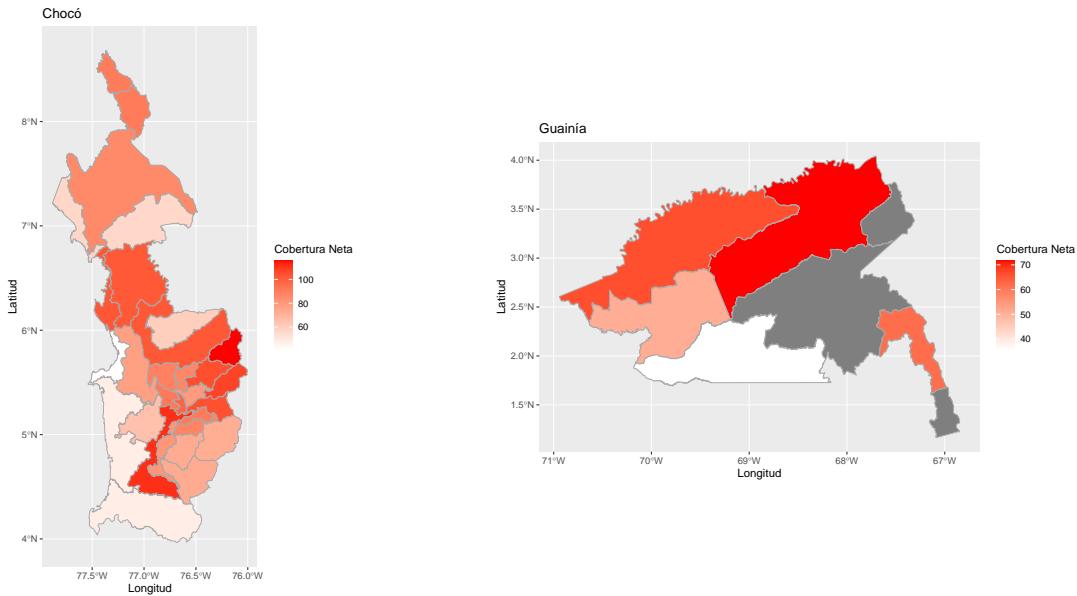


Figura 9: CNS municipios no medidos

Los municipios en los que se pudo estimar la cobertura neta secundaria (CNS) en 2022 fueron Belén de Bajirá en La Guajira y Mapiripana en el Chocó. Podemos observar que la media de ambos municipios es de 80.64 y 83.41, respectivamente, destacando que Belén de Bajirá parece tener una mayor CNS que los demás municipios.

## Punto 12.

Validar la bondad ajuste de  $M_5$  por medio de la distribución predictiva posterior en cada municipio, utilizando como estadísticos de prueba el mínimo, el máximo, el rango intercuartílico, la media, la mediana, y la desviación estándar. Presentar los resultados visualmente. Interpretar los resultados obtenidos (máximo 100 palabras).

**Nota:** Tomar como ejemplo la Sección 8.5 del artículo Some Developments in Bayesian Hierarchical Linear Regression Modeling .

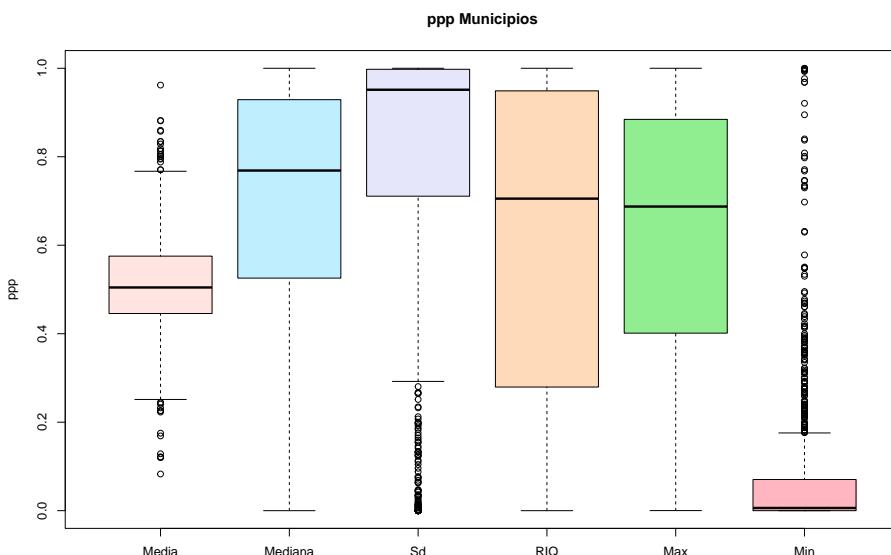


Figura 10: Validacion

Parece que nuestro modelo está capturando de manera efectiva las medias de los municipios y el valor máximo, mientras que se observa una gran variabilidad en lo que respecta al rango intercuartílico y el valor mínimo. Además, parece que se está sobreestimando la mediana y la desviación estándar, pero en general, parece que el modelo está captando adecuadamente la naturaleza de los datos.

## Apéndices:

### Distribuciones:

#### Modelo 1:

##### Distribución Posterior

$$\begin{aligned} p(\boldsymbol{\theta} \mid \text{resto}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= \prod_{j=1}^m \prod_{i=1}^{n_j} N(y_{i,j} \mid \theta, \sigma^2) \times \prod_{j=1}^m N(\theta \mid \mu_0, \gamma_0^2) \\ &\quad \times GI\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \end{aligned}$$

##### Distribuciones Condicionales Completas:

$$\begin{aligned} \theta \mid \text{resto} &\sim N\left(\frac{\mu_0/\gamma_0^2 + n\bar{y}/\sigma^2}{1/\gamma_0^2 + n/\sigma^2}, \frac{1}{1/\gamma_0^2 + n/\sigma^2}\right) \\ \sigma^2 \mid \text{resto} &\sim GI\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + S_\theta^2}{2}\right) \end{aligned}$$

#### Modelo 2:

##### Distribución Posterior:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= \prod_{j=1}^m \prod_{i=1}^{n_j} N(y_{i,j} \mid \theta_j, \sigma^2) \times \prod_{j=1}^m N(\theta_j \mid \mu, \tau^2) \\ &\quad \times GI\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \times N(\mu \mid \mu_0, \gamma_0^2) \times GI\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right) \end{aligned}$$

##### Distribuciones Condicionales Completas:

$$\begin{aligned} \theta_j \mid \text{resto} &\sim N\left(\frac{\mu/\tau^2 + n_j \bar{y}_j/\sigma^2}{1/\tau^2 + n_j/\sigma^2}, \frac{1}{1/\tau^2 + n_j/\sigma^2}\right) \\ \sigma^2 \mid \text{resto} &\sim GI\left(\frac{\nu_0 + \sum_{j=1}^m n_j}{2}, \frac{\nu_0 \sigma_0^2 + \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{i,j} - \theta_j)^2}{2}\right) \\ \mu \mid \text{resto} &\sim N\left(\frac{\mu_0/\gamma_0^2 + m\bar{\theta}/\tau^2}{1/\gamma_0^2 + m/\tau^2}, \frac{1}{1/\gamma_0^2 + m/\tau^2}\right) \\ \tau^2 \mid \text{resto} &\sim GI\left(\frac{\eta_0 + m}{2}, \frac{\eta_0 \tau_0^2 + \sum_{j=1}^m (\theta_j - \mu)^2}{2}\right) \end{aligned}$$

#### Modelo 3:

##### Distribución Posterior

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
&= \prod_{j=1}^m \prod_{i=1}^{n_j} N(y_{i,j} \mid \theta_j, \sigma^2) \\
&\quad \times \prod_{j=1}^m N(\theta_j \mid \mu, \tau^2) \times \prod_{j=1}^m GI\left(\sigma_j^2 \mid \frac{\nu}{2}, \frac{\nu \sigma^2}{2}\right) \\
&\quad \times N(\mu \mid \mu_0, \gamma_0^2) \times GI\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right) \\
&\quad \times e^{-\lambda_0 \nu} \times G(\sigma^2 \mid \alpha_0, \beta_0)
\end{aligned}$$

**Distribuciones Condicionales Completas:**

$$\begin{aligned}
\theta_j \mid \text{resto} &\sim N\left(\frac{\mu/\tau^2 + n_j \bar{y}_j / \sigma_j^2}{1/\tau^2 + n_j / \sigma_j^2}, \frac{1}{1/\tau^2 + n_j / \sigma_j^2}\right) \\
\sigma_j^2 \mid \text{resto} &\sim GI\left(\frac{\nu + n_j}{2}, \frac{\nu \sigma^2 + \sum_{i=1}^{n_j} (y_{i,j} - \theta_j)^2}{2}\right) \\
\mu \mid \text{resto} &\sim N\left(\frac{\mu_0 / \gamma_0^2 + m \bar{\theta} / \tau^2}{1 / \gamma_0^2 + m / \tau^2}, \frac{1}{1 / \gamma_0^2 + m / \tau^2}\right) \\
\tau^2 \mid \text{resto} &\sim GI\left(\frac{\eta_0 + m}{2}, \frac{\eta_0 \tau_0^2 + \sum_{j=1}^m (\theta_j - \mu)^2}{2}\right) \\
\sigma^2 \mid \text{resto} &\sim G\left(\alpha_0 + \frac{m\nu}{2}, \frac{\beta_0}{2} + \frac{\nu}{2} \sum_{j=1}^m \frac{1}{\sigma_j^2}\right)
\end{aligned}$$

**Modelo 4:**

**Distribución Posterior**

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \text{resto}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
&= \prod_{k=1}^m \prod_{j=1}^{n_k} \prod_{i=1}^{n_{j,k}} N(y_{i,j,k} \mid \zeta_{j,k}, \kappa^2) \\
&\quad \times \prod_{k=1}^m \prod_{j=1}^{n_k} N(\zeta_{j,k} \mid \theta_k, \sigma^2) \times GI\left(\kappa^2 \mid \frac{\xi_0}{2}, \frac{\xi_0 \kappa_0^2}{2}\right) \\
&\quad \times \prod_{k=1}^m N(\theta_k \mid \mu, \tau^2) \times GI\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \\
&\quad \times N(\mu \mid \mu_0, \gamma_0^2) \times GI\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2}\right)
\end{aligned}$$

### Distribuciones Condicionales Completas:

$$\begin{aligned}
\zeta_{j,k} \mid \text{resto} &\sim N\left(\frac{\theta_k/\sigma^2 + n_{j,k}\bar{y}_{j,k}/\kappa^2}{1/\sigma^2 + n_{j,k}/\kappa^2}, \frac{1}{1/\sigma^2 + n_{j,k}/\kappa^2}\right) \\
\kappa^2 \mid \text{resto} &\sim GI\left(\frac{\xi_0 + n}{2}, \frac{\xi_0\kappa_0^2 + \sum_{k=1}^m \sum_{j=1}^{n_k} \sum_{i=1}^{n_{j,k}} (y_{i,j,k} - \zeta_{j,k})^2}{2}\right) \\
\theta_k \mid \text{resto} &\sim N\left(\frac{\mu/\tau^2 + n_k\bar{\zeta}_k/\sigma^2}{1/\tau^2 + n_k/\sigma^2}, \frac{1}{1/\tau^2 + n_k/\sigma^2}\right) \\
\sigma^2 \mid \text{resto} &\sim GI\left(\frac{\nu_0 + \sum_{k=1}^m n_k}{2}, \frac{\nu_0\sigma_0^2 + \sum_{k=1}^m \sum_{j=1}^{n_k} (\zeta_{j,k} - \theta_k)^2}{2}\right) \\
\mu \mid \text{resto} &\sim N\left(\frac{\mu_0/\gamma_0^2 + m\bar{\theta}/\tau^2}{1/\gamma_0^2 + m/\tau^2}, \frac{1}{1/\gamma_0^2 + m/\tau^2}\right) \\
\tau^2 \mid \text{resto} &\sim GI\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{k=1}^m (\theta_k - \mu)^2}{2}\right)
\end{aligned}$$

**Modelo 5:**

### Distribución Posterior

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \text{resto}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \prod_{k=1}^m \prod_{j=1}^{n_k} \prod_{i=1}^{n_{j,k}} N(y_{i,j,k} \mid \zeta_{j,k}, \kappa^2) \\
&\quad \times \prod_{k=1}^m \prod_{j=1}^{n_k} N(\zeta_{j,k} \mid \theta_k, \sigma_k^2) \times GI\left(\kappa^2 \mid \frac{\xi_0}{2}, \frac{\xi_0\kappa_0^2}{2}\right) \\
&\quad \times \prod_{k=1}^m N(\theta_k \mid \mu, \tau^2) \times GI\left(\sigma_k^2 \mid \frac{\nu}{2}, \frac{\nu\sigma^2}{2}\right) \\
&\quad \times N(\mu \mid \mu_0, \gamma_0^2) \times GI\left(\tau^2 \mid \frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right) \times 1 \times G\left(\sigma^2 \mid \frac{\alpha_0}{2}, \frac{\beta_0}{2}\right)
\end{aligned}$$

### Distribuciones Condicionales Completas:

$$\begin{aligned}
\zeta_{j,k} \mid \text{resto} &\sim N\left(\frac{n_{j,k}\bar{y}_{j,k}/\kappa^2 + \theta_k/\sigma_k}{1/\sigma_k^2 + n_{j,k}/\kappa^2}, \frac{1}{1/\sigma_k^2 + n_{j,k}/\kappa^2}\right) \\
\kappa^2 \mid \text{resto} &\sim GI\left(\frac{\xi_0 + n}{2}, \frac{\xi_0\kappa_0^2 + \sum_{k=1}^m \sum_{j=1}^{n_k} \sum_{i=1}^{n_{j,k}} (y_{i,j,k} - \zeta_{j,k})^2}{2}\right) \\
\theta_k \mid \text{resto} &\sim N\left(\frac{\mu/\tau^2 + n_k\bar{\zeta}_k/\sigma_k^2}{1/\tau^2 + n_k/\sigma_k^2}, \frac{1}{1/\tau^2 + n_k/\sigma_k^2}\right) \\
\sigma_k^2 \mid \text{resto} &\sim GI\left(\frac{\nu + n_k}{2}, \frac{\nu\sigma^2 + \sum_{j=1}^{n_k} (\zeta_{j,k} - \theta_k)^2}{2}\right) \\
\mu \mid \text{resto} &\sim N\left(\frac{\mu_0/\gamma_0^2 + m\bar{\theta}/\tau^2}{1/\gamma_0^2 + m/\tau^2}, \frac{1}{1/\gamma_0^2 + m/\tau^2}\right) \\
\tau^2 \mid \text{resto} &\sim GI\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{k=1}^m (\theta_k - \mu)^2}{2}\right) \\
\sigma^2 \mid \text{resto} &\sim G\left(\frac{\alpha_0 + m\nu}{2}, \frac{\beta_0}{2} + \frac{\nu}{2} \sum_{k=1}^m \frac{1}{\sigma_k^2}\right)
\end{aligned}$$

## Coeficientes de variación de Monte Carlo

Modelo	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Modelo 1	7.300e-09	1.795e-04	3.590e-04	3.590e-04	5.386e-04	7.181e-04
Modelo 2	0.001799	0.003815	0.004766	0.037184	0.008452	0.991229
Modelo 3	0.001777	0.004639	0.121204	0.315856	0.330449	5.697986
Modelo 4	0.001743	0.027353	0.040060	0.044029	0.056188	1.025713
Modelo 5	0.0018	0.0275	0.0406	1.4451	0.0572	1603.9091

Tabla 5: Resumen coeficientes.

**Nota:** El valor máximo en los coeficientes de Monte Carlo del Modelo 5 corresponde al parámetro asociado con la varianza de Bogotá.

## 1. Bibliografía:

- [1] Universidad Nacional de Colombia (2023). *Estadística bayesiana*. <https://sites.google.com/view/juansosa/bayesian-statistics?authuser=0>
- [2] Universidad Nacional de Colombia (2023). *Computación Estadística*. <https://r2022.netlify.app>