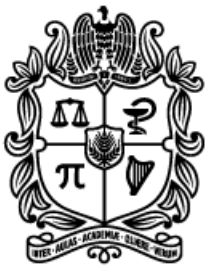

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Universidad Nacional de Colombia

Modelos Multinivel Bayesianos: Análisis Profundo de los Resultados de la Prueba Saber 11 para la Configuración Nacional, Departamental y Municipal en Colombia

Autor:

Juan Daniel Diaz Alvarez
juadiazal@unal.edu.co

Juan Sosa
Ph.D Estadística

Noviembre del 2023

Índice

1. Introducción

De acuerdo con la Guía de Usuario del Examen Saber 11, este examen, realizado semestralmente por el Icfes, tiene múltiples objetivos. En primer lugar, sirve como un criterio de admisión para estudiantes que desean ingresar a las Instituciones de Educación Superior. Además, su propósito incluye el monitoreo de la calidad de la educación ofrecida en instituciones de educación media y la generación de información relevante para la estimación del valor agregado de la educación superior ("Guia de Usario examen Saber 11", 2014).

Siguiendo los lineamientos de la Prueba Saber 11, este examen produce resultados a nivel individual de estudiantes que se encuentran en la fase final de su educación media. Estos resultados se expresan en puntajes obtenidos en cinco pruebas genéricas: Matemáticas, Lectura, Ciencias, Sociales e Inglés. Los puntajes están en una escala establecida en la segunda aplicación del año 2014, con un promedio de 50 y una desviación estándar de 10. Esta fijación de la media y la desviación estándar permite establecer una línea de base y proporciona un punto de referencia para las estimaciones. Además, se calcula un puntaje global, que se obtiene mediante un promedio ponderado de los puntajes en las cinco pruebas genéricas ("Documentación del examen Saber 11", 2014).

Así, el puntaje global (PG) de la prueba Saber 11 está dado por:

$$\text{PG} = 5 \cdot \left(\frac{5 \cdot \text{M} + 3 \cdot \text{L} + 3 \cdot \text{C} + 3 \cdot \text{S} + 1 \cdot \text{I}}{13} \right)$$

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 propósito de este estudio consiste en la configuración de modelos multinivel Bayesianos, empleando como conjunto de datos de entrenamiento el puntaje global de los estudiantes, para así modelar los resultados de las pruebas a nivel nacional, desglosados por Municipio y Departamento, con los siguientes objetivos:

- Establecer un ranking y una segmentación probabilística de los departamentos basados en su puntaje global promedio.
- Crear un ranking y una segmentación probabilística de los municipios en función de su puntaje global promedio.
- Desarrollar un modelo predictivo de la incidencia de la pobreza monetaria basado en el puntaje global promedio por departamento.

- Establecer un modelo predictivo de la cobertura neta de educación secundaria a partir del puntaje global promedio por municipio.

Para cumplir con los objetivos del estudio emplearemos los resultados de las pruebas Saber 11 desglosados municipio y departamento, acompañado de la incidencia de la pobreza monetaria y por último la cobertura neta secundaria. Por lo que excluiremos del estudio datos faltantes en la ubicación del colegio por Municipio, Departamento y Puntaje Global, igualmente se exluirán las estudiantes que no sean de nacionalidad colombiana y que no residen en Colombia, además se excluirá San Andrés y por último se usarán los registros que el Icfes haya catalogado como *Publicar*.

En la sección 2, se analizarán en detalle los modelos Bayesianos propuestos, así como su estructura jerárquica. En la sección 3, se investigará la relación entre la incidencia monetaria por departamento y los puntajes de las pruebas Saber 11, además de examinar la cobertura neta de la educación por municipio y los puntajes de las pruebas Saber 11 por municipio. En la sección 4, se procederá a ajustar los modelos Bayesianos y a establecer rankings Bayesianos basados en los puntajes obtenidos por departamento y municipio. En la sección 5, se llevará a cabo el ajuste de modelos predictivos de la incidencia de la pobreza monetaria y la cobertura neta en educación secundaria basados en los puntajes de las pruebas Saber 11. Y Finalmente, se presentarán las conclusiones más relevantes encontradas a lo largo de todo el documento.

2. Modelos Bayesianos y su estructura jerarquica

Consideraremos 5 modelos Bayesiano basados en la distribución normal, dadas las características de la prueba Saber 11, con esto veremos el comportamiento desde el modelo más simple hasta el más complejo, ganando flexibilidad y extrapolación de datos, pero con un costo computacional alto. A continuación se especifican los modelos considerados:

1. **Modelo Normal**.

   Distribución muestral:

   $$y_{ij} \mid \theta, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$$

   Para $i = 1, \ldots, n_j$ y $j = 1, \ldots, m$ donde y_{ij} es el untaje global del estudiante i en el departamento j.

Distribución previa:

$$\theta \sim \mathsf{N}(\mu_0, \tau_0^2) \qquad \sigma^2 \sim \mathsf{GI}\left(\frac{\nu_0}{2}, \frac{\nu_0\,\sigma_0^2}{2}\right)$$

Donde $\mu_0, \tau_0^2, \nu_0, \sigma_0^2$ son los hiperparámetros del modelo.

2. **Modelo Normal con medias especificas por departamento.**

Distribución muestral:

$$y_{ij} \mid \theta_j, \sigma^2 \overset{\text{ind}}{\sim} \mathsf{N}(\theta_j, \sigma^2)$$

Distribución previa:

$$\theta_j \mid \mu, \tau^2 \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu, \tau^2\right), \qquad \mu \sim \mathsf{N}(\mu_0, \gamma_0^2), \qquad \tau^2 \sim \mathsf{GI}\left(\frac{\eta_0}{2}, \frac{\eta_0\,\tau_0^2}{2}\right)$$
$$\sigma^2 \sim \mathsf{GI}\left(\frac{\nu_0}{2}, \frac{\nu_0\,\sigma_0^2}{2}\right)$$

Donde $\mu_0, \gamma_0^2, \tau_0^2, \nu_0, \sigma_0^2$ son los hiperparámetros del modelo.

3. **Modelo Normal con medias y varianzas específicas por departamento.**

Distribución muestral:

$$y_{ij} \mid \theta_j, \sigma_j^2 \overset{\text{ind}}{\sim} \mathsf{N}\left(\theta_j, \sigma_j^2\right)$$

Distribución previa:

$$\theta_j \mid \mu, \tau^2 \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu, \tau^2\right), \qquad \mu \sim \mathsf{N}(\mu_0, \gamma_0^2), \qquad \tau^2 \sim \mathsf{GI}\left(\frac{\eta_0}{2}, \frac{\eta_0\,\tau_0^2}{2}\right),$$
$$\sigma_j^2 \mid \nu, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{GI}\left(\frac{\nu}{2}, \frac{\nu\,\sigma^2}{2}\right), \qquad \nu = \text{Constante}, \qquad \sigma^2 \sim \mathsf{G}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right)$$

Donde $\mu_0, \gamma_0^2, \eta_0, \tau_0^2, \nu, \alpha_0, \beta_0$ son los hiperparámetros del modelo.

4. **Modelo Normal con medias específicas por municipio y departamento.**

Distribución muestral:

$$y_{ijk} \mid \zeta_{jk}, \kappa^2 \overset{\text{ind}}{\sim} \mathsf{N}\left(\zeta_{jk}, \kappa^2\right)$$

Para $i = 1,\ldots,n_{jk}$, $j = 1,\ldots,n_k$, $k = 1,\ldots,m$ donde y_{ijk} es el puntaje global del estudiante i en el municipio j del departameto k.

Distribución previa:

$$\zeta_{jk}|\theta_k,\sigma^2 \overset{\text{ind}}{\sim} \mathsf{N}(\theta_k,\sigma^2), \qquad \kappa^2 \sim \mathsf{GI}\left(\frac{\xi_0}{2},\frac{\xi_0\kappa_0^2}{2}\right)$$

$$\theta_k \mid \mu,\tau^2 \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu,\tau^2\right), \qquad \mu \sim \mathsf{N}(\mu_0,\gamma_0^2), \qquad \tau^2 \sim \mathsf{GI}\left(\frac{\eta_0}{2},\frac{\eta_0\,\tau_0^2}{2}\right)$$

$$\sigma^2 \sim \mathsf{GI}\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)$$

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

5. **Modelo Normal con medias específicas por municipio y departamento.**

Distribución muestral:

$$y_{ijk} \mid \zeta_{jk},\kappa^2 \overset{\text{ind}}{\sim} \mathsf{N}\left(\zeta_{jk},\kappa^2\right)$$

Distribución previa:

$$\zeta_{jk}|\theta_k,\sigma_k^2 \overset{\text{ind}}{\sim} \mathsf{N}(\theta_k,\sigma_k^2), \qquad \kappa^2 \sim \mathsf{GI}\left(\frac{\xi_0}{2},\frac{\xi_0\kappa_0^2}{2}\right)$$

$$\theta_k \mid \mu,\tau^2 \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu,\tau^2\right), \qquad \mu \sim \mathsf{N}(\mu_0,\gamma_0^2), \qquad \tau^2 \sim \mathsf{GI}\left(\frac{\eta_0}{2},\frac{\eta_0\,\tau_0^2}{2}\right)$$

$$\sigma_k^2 \mid \nu,\sigma^2 \sim \mathsf{GI}\left(\frac{\nu}{2},\frac{\nu\sigma^2}{2}\right), \qquad \nu = \text{Constante}, \qquad \sigma^2 \sim \mathsf{G}\left(\frac{\alpha_0}{2},\frac{\beta_0}{2}\right)$$

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

En la Figura 1, se presentan visualmente los Grafos Acíclicos Dirigidos (DAG, por sus siglas en inglés) correspondientes a los modelos previamente discutidos, con la excepción del Modelo 1, conocido como el 'Modelo Normal', el cual, debido a su simplicidad y características claramente establecidas, no se incluye en esta representación gráfica. Este conjunto de representaciones gráficas de los modelos forma la jerarquía de modelos, en la que en la útlima jerarquía se encuentran los hiperparámetros, en los que se logra una representación visual de cómo influyen en la estructura de los modelos los cuales nos permitirán darle vida a los modelos, y estos serán ajustados de acuerdo a la información que se obtiene de las pruebas Saber 11.

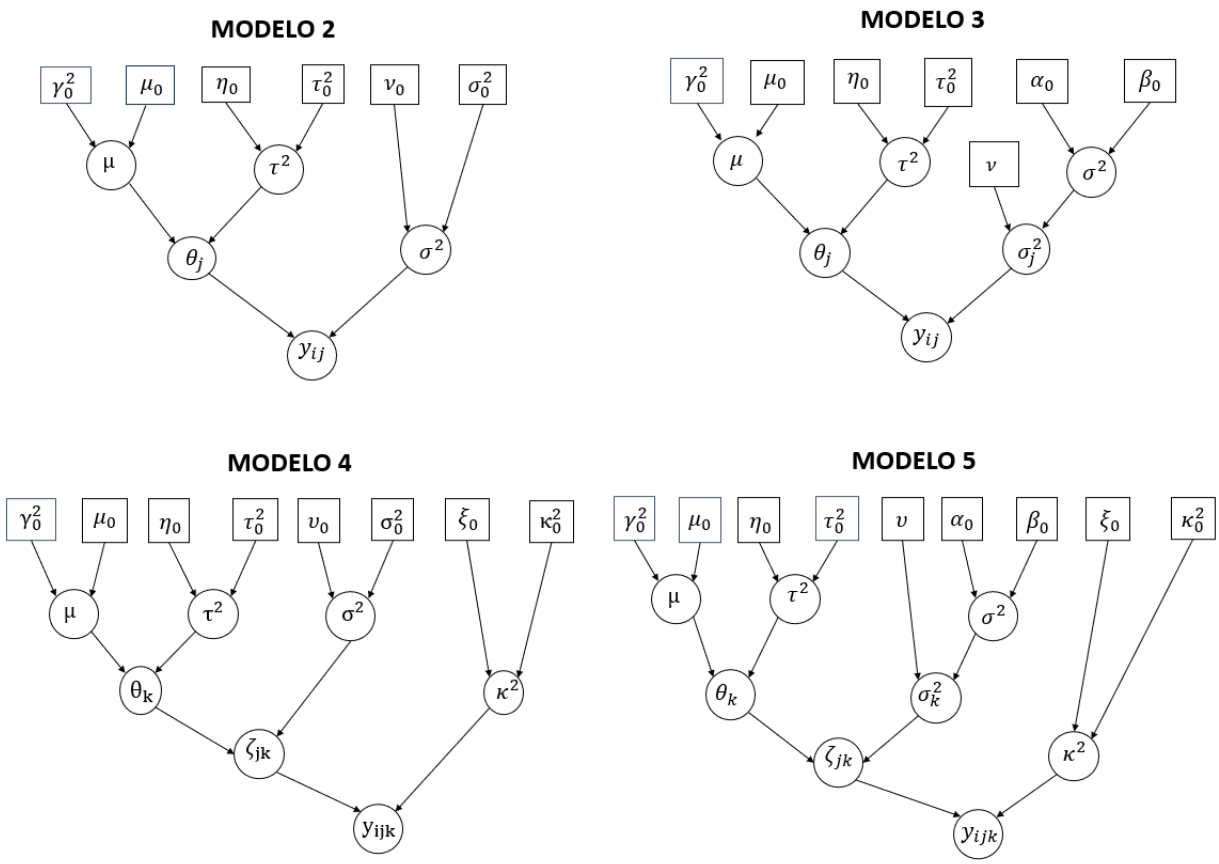

Figura 1: DAGs

Cada modelo será ajustado usando el muestreador de Gibbs, con un total de 101000 iteraciones, donde las primeras 1000 iteraciones del muestreador serán el período de calentamiento de la cadena. Una vez culminado este período, se realizará un muestreo sistemático con una amplitud de 10 y así completar un total de 10000 iteraciones para las inferencias posteriores.

3. Relaciones entre la incidencia de la pobreza monetaria, la cobertura educativa y los puntajes de las pruebas Saber 11

Las estadísticas de la incidencia de la pobreza monetaria se define como el porcentaje de las personas que son catalogadas como pobres. Teniendo como umbral el límite de la pobreza (LP) para determinar la caracterización. (Consejo Nacional de Política Económica y Social, 2012). Por otro lado el umbral, o bien llamado linea de pobreza, la CEPAl (como se citó en "Pobreza monetaria en Colombia: Resultados 2020") establece que: "La línea de pobreza representa un valor monetario en el cual se consideran dos componentes: el costo de adquirir una canasta básica de alimentos y

el costo de los demás bienes y servicios, expresado sobre la base de la relación entre el gasto total y el gasto en alimentos". Sobre esta base se construyen las estadísticas de la pobreza en Colombia. Por lo tanto, en la Figura 2 se muestra los puntajes medios de las pruebas Saber 11 y la incidencia de la pobreza monetaria para el año 2018 por departamento. Hay que aclarar que las estadísticas de la incidencia de la pobreza monetaria solo se obtuvieron para 23 departamentos.

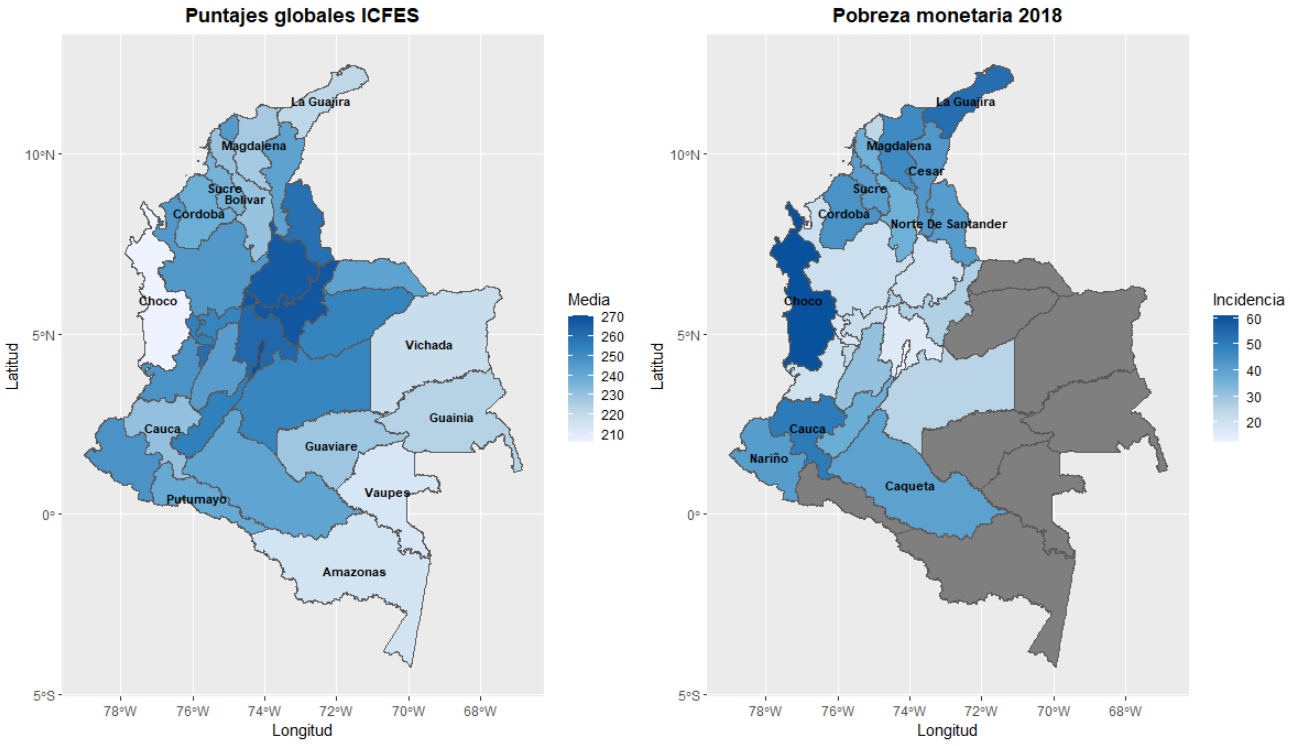

Figura 2: Pobreza monetaria 2018 y puntajes Saber 11 2022-2 por departamento

En la Figura 2 se destacan los departamentos donde el puntaje medio de la prueba Saber 11 fue menor a 240 puntos, mientras que para la pobreza monetaria se destaca los departamentos que tiene una incidencia mayor al 40 %. De manera particular se puede destacar el Departamento del Chocó, quien tuvo el puntaje medio en la prueba más bajo de todo del todo país y con la tasa de pobreza monetaria más alta. Lo mismo parece suceder con el departamento de La Guajira. Esto podría sugerir que las dificultades económicas pueden estar afectando la calidad de la educación o el rendimiento académico de los estudiantes en estos departamentos.

Para la Figura 3 se muestran los puntajes medios obtenidos en la prueba Saber 11 y la tasa de cobertura neta secundaria por municipio. Este último, es un indicador que mide el acceso y la permanencia de la población en edad escolar al sistema educativo y se calcula dividiendo el número

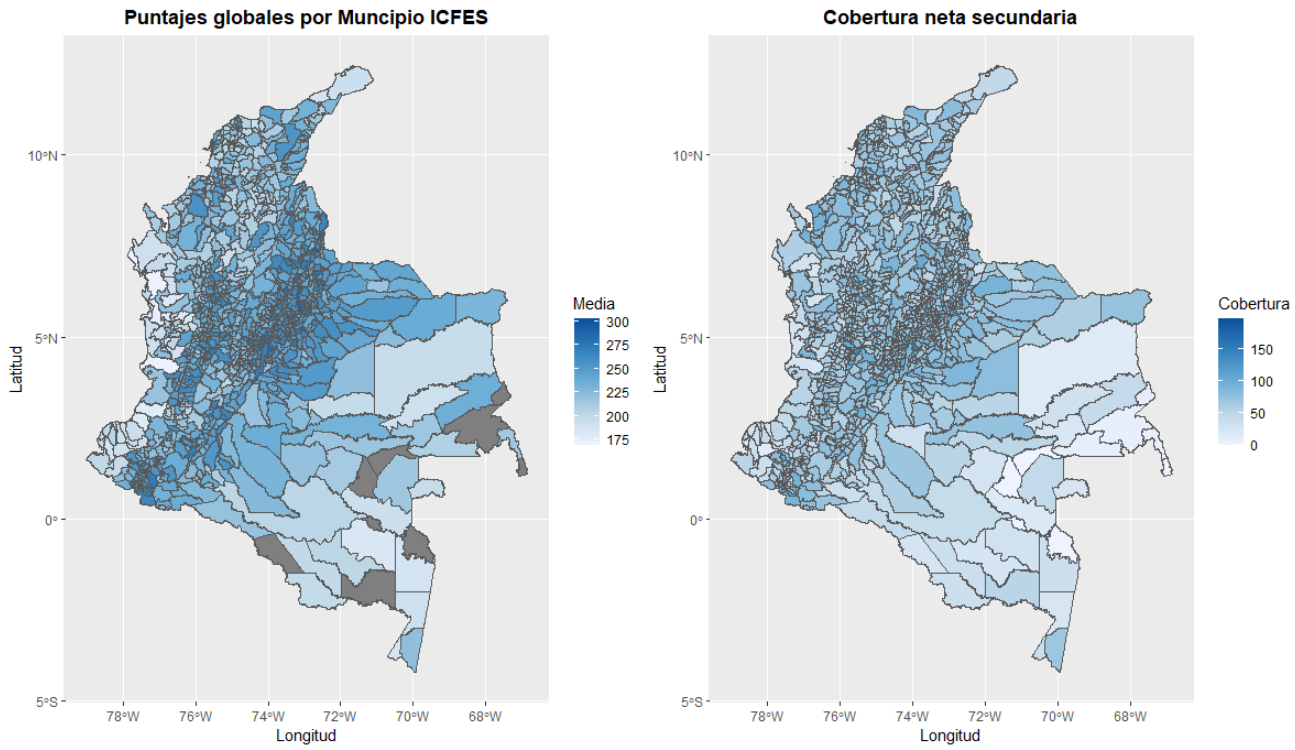

Figura 3: Cobertura neta secundaria y puntajes Saber 11 2022-2 por municipio

de estudiantes matriculados en secundaria que tienen la edad teórica para cursarlo por el total de la población correspondiente a esa misma edad (Ministerio de Educación Nacional, 2023).

El mapa sugiere una variación en las puntuaciones medias de los municipios como en la tasa de cobertura. Algunos municipios tienen puntuaciones altas y una alta tasa cobertura, lo que podría indicar un alto nivel de acceso y permanencia en el sistema educativo. Sin embargo, otros municipios tienen puntuaciones medias más bajas y una menor tasa de cobertura, lo que podría indicar desafíos en términos de acceso y permanencia en el sistema educativo. Lo que podría indicar que la tasa de cobertura neta secundaria esté afectando los resultados de los estudiantes en estos municipios.

4. Ajuste de los modelos

Como se mencionó en la Sección 2 se implementará el muestreador de Gibbs. Para tal fin, usaremos previas difusas, atendiendo además, a la información que proporciona la documentación

de la prueba Saber 11. Así los hiperparámetros de cada modelo, estarán dispuestos a continuación:

- $\mathbf{M}_1 : \mu_0 = 250, \gamma_0^2 = 50^2, \nu_0 = 1, \sigma_0^2 = 50^2$.
- $\mathbf{M}_2 : \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu_0^2 = 1, \sigma_0^2 = 50^2$.
- $\mathbf{M}_3 : \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu = 1, \alpha_0 = 1, \beta_0^2 = 1/50^2$.
- $\mathbf{M}_4 : \xi_0 = 1, \kappa_0^2 = 50^2, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu_0^2 = 1, \sigma_0^2 = 50^2$.
- $\mathbf{M}_5 : \xi_0 = 1, \kappa_0^2 = 50^2, \mu_0 = 250, \gamma_0^2 = 50^2, \eta_0 = 1, \tau_0^2 = 50^2, \nu = 1, \alpha_0 = 1, \beta_0^2 = 1/50^2$.

4.1. Validación de la convergencia

5. Apendices

5.1. Distribuciones Condicionales Completas - DDC

- **Modelo 3**

$$\theta_j|- \sim \mathsf{N}\left(\frac{n_j\bar{y}_j + \mu/\tau^2}{n_j/\sigma_j^2 + 1/\tau^2}, \frac{1}{n_j/\sigma_j^2 + 1/\tau^2}\right)$$

$$\sigma_j^2|- \sim \mathsf{GI}\left(\frac{\nu + n_j}{2}, \frac{\nu\sigma^2 + \sum_{i=1}^{n_j}(y_{i,j} - \theta_j)^2}{2}\right)$$

$$\mu|- \sim \mathsf{N}\left(\frac{m\bar{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \frac{1}{m/\tau^2 + 1/\gamma_0^2}\right)$$

$$\tau^2|- \sim \mathsf{GI}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{j=1}^{m}(\theta_j - \mu)^2}{2}\right)$$

$$\sigma^2|- \sim \mathsf{G}\left(\frac{\alpha_0 + \nu m}{2}, \frac{\beta_0}{2} + \frac{\nu}{2}\sum_{j=1}^{m}\frac{1}{\sigma_j^2}\right)$$

- **Modelo 4:** las DCC de μ y τ^2 ya se calcularon en el Modelo 3.

$$\theta_k|- \sim \mathsf{N}\left(\frac{n_k\bar{\zeta}_k/\sigma_k^2+\mu/\tau^2}{n_k/\sigma^2+1/\tau^2}, \frac{1}{n_k/\sigma_k^2+1/\tau^2}\right)$$

$$\zeta_{j,k}|- \sim \mathsf{N}\left(\frac{n_{jk}\bar{y}_{jk}/\kappa^2+\theta_k/\sigma^2}{n_{jk}/\kappa^2+1/\sigma_k^2}, \frac{1}{n_{jk}/\kappa^2+1/\sigma^2}\right)$$

$$\sigma^2|- \sim \mathsf{GI}\left(\frac{\nu_0+\sum_{k=1}^m n_k}{2}, \frac{\nu_0\sigma_0^2+\sum_{k=1}^m\sum_{j=1}^{n_{j,k}}(\zeta_{j,k}-\theta_k)^2}{2}\right)$$

$$\kappa^2|- \sim \mathsf{GI}\left(\frac{\xi_0+n}{2}, \frac{\xi_0\kappa_0^2+\sum_{k=1}^m\sum_{j=}^{n_k}\sum_{i=1}^{n_{jk}}(y_{ijk}-\zeta_{jk})^2}{2}\right)$$

- **Modelo 5**: las DCC de $\theta_k, \sigma^2, \mu, \tau^2, \zeta_{jk}$ y κ^2 se calcularon en modelos anteriores.

$$\sigma_k^2|- \sim \mathsf{GI}\left(\frac{\nu+n_k}{2}, \frac{\nu\sigma_0^2+\sum_{j=1}^{n_{jk}}(\zeta_{jk}-\theta_k)^2}{2}\right)$$

6. Referencias

Referencias

Consejo Nacional de Política Económica y Social. (2012). *Metodologías oficiales y arreglos institucionales para la medición de la pobreza en Colombia* (inf. téc.). Departamento Nacional de Planeación.

Departamento Administrativo Nacional de Estadística (DANE). (2021). Pobreza monetaria en Colombia: Resultados 2020. https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2020/Presentacion-pobreza-monetaria_2020.pdf

Documentación del examen Saber 11. (2014).

Guia de Usario examen Saber 11. (2014).

Ministerio de Educación Nacional. (2023). Tasa de Cobertura Neta Nacional [Accedido el 16 de octubre de 2023]. http://bi.mineducacion.gov.co:8380/eportal/web/planeacion-basica/tasa-de-cobertura-neta1