Estadística Bayesiana
Examen Parcial # 2

Instrucciones generales

- Este caso de estudio constituye el 60% de la calificación del Examen Parcial 2.
- Debe asociarse con otra persona, entendiendo que la calificación del examen será la misma para ambas personas.
- El reporte final se debe enviar a más tardar el **domingo 29 de octubre de 2023** a las 11:59 pm a la cuenta de correo:

  `jcsosam@unal.edu.co`
- Reportar las cifras utilizando la cantidad adecuada de decimales, dependiendo de lo que se quiera mostrar y las necesidades del problema.
- Numerar figuras y tablas y proporcionarles un tamaño adecuado que no distorsione la información que estas contienen.
- El archivo del reporte final debe ser un archivo `pdf`.
- Usar `LateX` o `Markdown` (en `R` o `Python`) para escribir el informe.
- El código fuente de `R` o `Python` debe reproducir exactamente todos los resultados (incluir semillas donde sea necesario).
- La presentación, la organización, la redacción, y la ortografía serán parte integral de la calificación.

- Si los estudiantes Juan Sosa y Ernesto Perez trabajan juntos, tanto el archivo `pdf` del informe, así como el código fuente, y el asunto del e-mail donde se adjuntan estos archivos, se deben llamar de la siguiente manera:

  `bayes - parcial 1 - juan sosa - ernesto perez`

  Esta condición es indispensable para que su examen sea calificado.

- Usar reglas APA para hacer las referencias correspondientes. No copiar texto de libros o internet sin hacer la cita correspondiente.

- El informe no tiene que ser extenso. Recuerde ser minimalista escribiendo el reporte. Se deben incluir solo aquellos gráficos, tablas, y ecuaciones que sean relevantes para la discusión.

- Cualquier evidencia de plagio o copia se castigará severamente tal y como el reglamento de la Universidad Nacional de Colombia lo estipula. Dejo a mi discreción el uso de software especializado para evaluar si hay copia o plagio de otros informes o internet.

  Si está claro que (por ejemplo) dos grupos han trabajado juntos en una parte de un problema que vale 20 puntos, y cada respuesta habría ganado 16 puntos (si no hubiera surgido de una colaboración ilegal), entonces cada grupo recibirá 8 de los 16 puntos obtenidos colectivamente (para una puntuación total de 8 de 20), y me reservo el derecho de imponer penalidades adicionales a mi discreción.

  Si un grupo resuelve un problema por su cuenta y luego comparte su solución con cualquier otro grupo (porque rutinariamente Usted hace esto, o por lástima, o bondad, o por cualquier motivo que pueda creer tener; ¡no importa!), Usted es tan culpable de colaboración ilegal como la persona que tomó su solución, y ambos recibirán la misma penalidad. Este tipo de cosas es necesario hacerlas ya que muchas personas no hacen trampa, y debo asegurarme de que sus puntajes son obtenidos de manera genuina. En otros semestres, unos estudiantes perdieron la materia debido a una colaboración ilegal; ¡no deje que le suceda a Usted!

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

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

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 de del puntaje global promedio por municipio.

Tratamiento de datos

Para ajustar los modelo propuestos, se consideran únicamente los estudiantes con:

- 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 **525061 registros**. Usar el diccionario de variables para realizar este proceso.

Modelos

M_1: Modelo Normal

Distribución muestral:

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

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

Distribución previa:

$$\theta \sim \mathsf{N}(\mu_0, \gamma_0^2)\,, \quad \sigma^2 \sim \mathsf{GI}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)\,,$$

donde $\mu_0, \gamma_0^2, \nu_0, \sigma_0^2$ son los hiperparámetros del modelo y $\mathsf{GI}(\alpha, \beta)$ 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 $\alpha > 2$.

Nota: Este modelo se encuentra desarrollado en este enlace.

M_2: Modelo Normal con medias específicas por departamento

Distribución muestral:

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

Distribución previa:

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

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

Nota: Este modelo se encuentra desarrollado en este enlace.

M_3: Modelo Normal con medias y varianzas específicas por departamento

Distribución muestral:

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

Distribución previa:

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

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

Nota: Este modelo se encuentra desarrollado en este enlace. ¡Cuidado! La parametrización de la previa de σ^2 es $\mathsf{G}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right)$ en lugar de $\mathsf{G}\left(\alpha_0, \beta_0\right)$.

M_4: Modelo Normal con medias específicas por municipio y departamento)

Distribución muestral:

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

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

Distribución previa:

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

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

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

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.

M_5: Modelo Normal con medias específicas por municipio y departamento

Distribución muestral:

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

Distribución previa:

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

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

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

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

Desarrollo metodológico

Los modelos presentados anteriormente se ajustan por medio de **muestreadores de Gibbs** con $1000 + 10 \cdot 10000 = 101000$ iteraciones. Las primeras 1000 iteraciones del algoritmo constituyen el periodo de calentamiento del algoritmo (no se tienen en cuenta para realizar inferencia). Además, con el fin de reducir la autocorrelación de la cadena después del periodo de calentamiento, se hace un muestreo sistemático de amplitud 10, de forma la cadena para realizar inferencias acerca de la distribución posterior de los parámetros de cada modelo consta de $B = 10000$ iteraciones.

Para tal fin se emplean distribuciones previas difusas definidas por los siguientes hiperparámetros a partir de la información de la prueba:

- M_1: $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\nu_0 = 1$, $\sigma_0^2 = 50^2$.
- M_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_3: $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\eta_0 = 1$, $\tau_0^2 = 50^2$, $\nu = 1$, $\alpha_0 = 1$, $\beta_0 = 1/50^2$.
- M_4: $\xi_0 = 1$, $\kappa_0^2 = 50^2$, $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\eta_0 = 1$, $\tau_0^2 = 50^2$, $\nu_0 = 1$, $\sigma_0^2 = 50^2$.
- M_5: $\xi_0 = 1$, $\kappa_0^2 = 50^2$, $\mu_0 = 250$, $\gamma_0^2 = 50^2$, $\eta_0 = 1$, $\tau_0^2 = 50^2$, $\nu = 1$, $\alpha_0 = 1$, $\beta_0 = 1/50^2$.

Preguntas

1. En un gráfico con dos paneles (1×2), 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

D.C en el periodo 2002-2018. Estos datos son de carácter público y se pueden descargar gratuitamente de la página web del DANE en este enlace.

2. En un gráfico con dos paneles (1×2), 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.

3. En un gráfico con cuatro paneles (2×2), hacer el DAG 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).

   Nota: Tomar como ejemplo la Figura 1 del artículo *Some Developments in Bayesian Hierarchical Linear Regression Modeling*.

4. En un gráfico con cuatro paneles (2×2), 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*.

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

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

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 me-

dia muestral. En un gráfico con dos paneles (1×2), 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.

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.

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).

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 perteneces al mismo grupo. Interpretar los resultados obtenidos (máximo 100 palabras).

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).

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*.