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 **miércoles 17 de mayo de 2023** a las 11:59 am 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!

Distribución de los ingresos en Colombia

La base de datos `personas.csv` disponible en la página web del curso, corresponde a una muestra aleatoria del módulo de **Personas** (para las que el ingreso total está disponible y es mayor que cero) de la encuesta **Medición de Pobreza Monetaria y Desigualdad 2021** llevada a cabo en Colombia por el DANE, la cual se encuentra disponible en el enlace https://microdatos.dane.gov.co/index.php/catalog/733.

El Universo de la encuesta está conformado por la población civil no institucional, residente en todo el territorio nacional; va dirigida a todos los hogares encontrados en la vivienda. La encuesta utiliza informante directo para las personas de 18 años y más, y para aquellas de 10 a 17 años que trabajen o estén buscando trabajo. Para los demás se acepta informante idóneo (persona del hogar mayor de 18 años, que a falta del informante directo pueda responder correctamente las preguntas). No se acepta información de empleados del servicio doméstico, pensionistas, vecinos o menores, excepto cuando el menor de edad es el jefe del hogar o cónyuge.

El objetivo de este trabajo es construir un modelo multinivel completamente Bayesiano, tomando como datos de entrenamiento el **ingreso total** (`ingtot`; ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas), con el fin de modelar los ingresos por **dominio** (`dominio`; cada una de las 24 a.M., otras cabeceras y resto), y establecer un ranking junto con una segmentación de los mismos. Por lo tanto, se toma como variable de agrupamiento el dominio, y como variable respuesta el ingreso total.

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 la variable respuesta del individuo i en el grupo 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 https://rpubs.com/jstats1702/944440.

M_2: Modelo Normal con medias específicas

Distribución muestral:

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

Distribución previa:

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

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

Nota: este modelo se encuentra desarrollado en https://rpubs.com/jstats1702/950834.

M_3: Modelo Normal con medias y varianzas específicas

Distribución muestral:

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

Distribución previa:

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

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 https://rpubs.com/jstats1702/954522. ¡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 t

Distribución muestral:

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

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

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

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

Si $X \mid \kappa, \theta, \sigma^2 \sim \mathsf{t}_\kappa(\theta, \sigma^2)$, entonces $\mathsf{E}(X) = \theta$, para $\kappa > 1$, y $\mathsf{Var}(X) = \frac{\kappa}{\kappa-2}\,\sigma^2$, para $\kappa > 2$.

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 $y_{i,j} \mid \theta, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{t}_\kappa(\theta, \sigma^2)$ es equivalente a la distribución jerárquica dada por

$$y_{i,j} \mid \theta, \varsigma^2_{i,j} \overset{\text{ind}}{\sim} \mathsf{N}(\theta, \varsigma^2_{i,j})\,, \qquad \varsigma^2_{i,j} \mid \sigma^2 \overset{\text{iid}}{\sim} \mathsf{GI}\left(\frac{\kappa}{2}, \frac{\kappa\sigma^2}{2}\right)\,,$$

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

La inclusión de las variables $\varsigma^2_{i,j}$ en el modelo permite que todas las distribuciones condicionales completas de las cantidades desconocidas (incluyendo las mismas variables auxiliares) tengan forma probabilística conocida. En Gelman et al. (2013, pp. 293-294) hay una discusión detallada al respecto. Si no se consideran las variables $\varsigma^2_{i,j}$ 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 θ como σ^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 5 y 6.

Distribución previa:

$$\theta \sim \mathsf{N}(\mu_0, \gamma_0^2)\,, \quad \sigma^2 \sim \mathsf{G}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right)\,, \quad \kappa \sim \text{Constante}\,,$$

donde $\mu_0, \gamma_0^2, \alpha_0, \beta_0, \kappa$ son los hiperparámetros del modelo.

Nota: el muestreador de Gibbs para ajustar este modelo tiene tres pasos fundamentales, a saber, muestrear θ, muestrear σ^2 y muestrear cada $\varsigma_{i,j}^2$, de las distribuciones condicionales completas correspondientes, dados los valores más recientes de los demás parámetros. Aunque sí es imperativo muestrear las variables $\varsigma_{i,j}^2$, no es necesario almacenarlas porque no son objeto de inferencia. Además, la verosimilitud se puede calcular utilizando la función `dt.scaled` del paquete `metRology`.

M_5: Modelo t con medias específicas

Distribución muestral:

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

Distribución previa:

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

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

M_6: Modelo t con medias y varianzas específicas

Distribución muestral:

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

Distribución previa:

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

$$\sigma_j^2 \overset{\text{iid}}{\sim} \mathsf{G}\left(\frac{\alpha}{2}, \frac{\beta}{2}\right), \qquad \alpha \sim \text{Constante}, \qquad \beta \sim \mathsf{G}\left(\frac{a_\beta}{2}, \frac{b_\beta}{2}\right),$$
$$\kappa \sim \text{Constante},$$

donde $\mu_0, \gamma_0^2, \eta_0, \tau_0^2, \alpha, a_\beta, b_\beta, \kappa$ son los hiperparámetros del modelo.

Desarrollo metodológico

En todos los casos, se utiliza la siguiente convención:

- n : número total de personas en la muestra.
- m : número de dominios.
- n_j : número de personas en el dominio j.
- $y_{i,j}$: ingreso total en **escala logarítmica** de la persona i en el dominio j.

Los modelos presentados anteriormente se ajustan por medio de muestreadores de Gibbs con $B = 11000$ iteraciones. Las primeras 1000 iteraciones del algoritmo constituyen el periodo de calentamiento del algoritmo, de manera que no se tienen en cuenta para realizar las inferencias. Para tal fin se emplean distribuciones previas empíricas difusas definidas por los siguientes hiperparámetros:

- M_1: $\mu_0 = 13.495$, $\gamma_0^2 = 11.382$, $\nu_0 = 1$, $\sigma_0^2 = 1.182$.
- M_2: $\mu_0 = 13.495$, $\gamma_0^2 = 11.382$, $\eta_0 = 1$, $\tau_0^2 = 1.182$, $\nu_0 = 1$, $\sigma_0^2 = 1.182$.
- M_3: $\mu_0 = 13.495$, $\gamma_0^2 = 11.382$, $\eta_0 = 1$, $\tau_0^2 = 1.182$, $\nu = 1$, $\alpha_0 = 1$, $\beta_0 = 0.846$.
- M_4: $\mu_0 = 13.495$, $\gamma_0^2 = 11.382$, $\alpha_0 = 1$, $\beta_0 = 0.846$, $\kappa = 3$.
- M_5: $\mu_0 = 13.495$, $\gamma_0^2 = 11.382$, $\eta_0 = 1$, $\tau_0^2 = 1.182$, $\alpha_0 = 1$, $\beta_0 = 0.846$, $\kappa = 3$.
- M_6: $\mu_0 = 13.495$, $\gamma_0^2 = 11.382$, $\eta_0 = 1$, $\tau_0^2 = 1.182$, $\alpha = 1$, $a_\beta = 1$, $b_\beta = 1.182$, $\kappa = 3$.

Los hiperparámetros se establecieron teniendo en cuenta que la media muestral es $\bar{y} = 13.495$, la varianza muestral es $s_y^2 = 1.182$ y el inverso de la varianza muestral es $1/s_y^2 = 0.846$.

Preguntas

1. Hacer el DAG de M_6 incluyendo las variables auxiliares.

2. Graficar la cadena de la log-verosimilitud de cada M_k, para $k = 1, \ldots, 6$. Incluir un anexo con las distribuciones condicionales completas de cada modelo (no incluir la demostración, solo cada distribución con sus respectivos parámetros).

   Nota: usar un solo panel para facilitar la comparación.

3. Calcular el DIC y el p_{DIC} de cada M_k, para $k = 1, \ldots, 6$. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

4. Calcular los ppp's para Bogotá asociados con la media, la mediana, la desviación estándar, el coeficiente de variación, el rango y el rango intercuartílico usando M_6. Presentar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

5. Hacer un ranking Bayesiano de los dominios basado en los efectos promedio de M_6. Hacer una visualización que incluya simultáneamente las estimaciones puntuales y los intervalos de credibilidad al 95%. Interpretar los resultados obtenidos (máximo 100 palabras).

   Nota: hacer la visualización en escala logarítmica. Además, usar la siguiente convención de colores: rojo oscuro para efectos promedio significativamente inferiores a 13.830, negro para efectos promedio que no difieren significativamente de 13.830 y verde oscuro para efectos promedio significativamente superiores a 13.830 (observe que 13.830 corresponde a un SMLMV de 2022 en escala logarítmica).

6. Estimar puntualmente la media, la desviación estándar y el coeficiente de variación de los ingresos para el Top 5 del ranking usando M_6. Reportar los resultados tabularmente. Interpretar los resultados obtenidos (máximo 100 palabras).

   Nota: reportar las estimaciones de la media y la desviación estándar usando la escala real en pesos y del coeficiente de variación en puntos porcentuales.

7. Conformar un arreglo que contenga las estimaciones puntuales de los efectos promedio y las desviaciones estándar de todos los dominios usando M_6. Usando este arreglo como insumo, segmentar los dominios por medio de agrupación jerárquica con cuatro grupos. Presentar los resultados visualmente. Interpretar los resultados obtenidos (máximo 100 palabras).

   Nota: estandarizar las columnas del arreglo antes de hacer el proceso de segmentación.

8. (Bono) 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 $y_i \mid \theta, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{t}_\kappa(\theta, \sigma^2)$, para $i = 1, \ldots, n$, es equivalente a la distribución jerárquica dada por

$$y_i \mid \theta, V_i \overset{\text{ind}}{\sim} \mathsf{N}(\theta, V_i)\,, \qquad V_i \mid \sigma^2 \overset{\text{iid}}{\sim} \mathsf{GI}\left(\frac{\kappa}{2}, \frac{\kappa\sigma^2}{2}\right)\,.$$

   Sugerencia: obtener la función de densidad de la distribución $\mathsf{t}_\kappa(\theta, \sigma^2)$, para $i = 1, \ldots, n$, a partir de

$$p(y_i \mid \theta, \sigma^2) = \int_0^\infty p(y_i, V_i \mid \theta, \sigma^2)\,\mathrm{d}V_i\,.$$

Referencias

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.