Estadística Bayesiana
Examen Parcial # 3

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 **martes 28 de noviembre 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!

1 Alcaldía de Bogotá 2023

En esta publicación de La Silla Vacía se "revisó el desempeño de cada encuestadora que hizo mediciones en las cinco ciudades principales frente a los resultados" de las elecciones regionales de Colombia de 2023. De acuerdo con este medio, Invamer fue una de las encuestadoras con *menor error total* en Bogotá. En esta hoja de cálculo se pueden consultar algunos detalles técnicos incluyendo los margenes de error de las encuestas.

Así, Invamer S.A.S realizó una encuesta del 17 al 23 de octubre de 2023 para Noticias Caracol, Blu Radio y El Espectador, con el fin de medir la intención de voto en Bogotá, Medellín, Cali, Barranquilla, y Bucaramanga para las elecciones de alcaldes de 2023. En Bogotá se encuestaron 1200 hombres y mujeres de 18 años en adelante, de todos los niveles socio-económicos a nivel nacional, aptos para votar en las elecciones y que sean residentes de la ciudad. Se realizaron encuestas personales en el hogar de los encuestados a través de tablets y para las preguntas de intención de voto se utilizó tarjetón. La ficha técnica se puede descargar aquí. Se obtuvieron los resultados que se presentan en la Tabla 1.

Intención de voto: Bogotá		
Candidato	Cantidad	Proporción
C. F. Galán	493	0.411
G. Bolivar	257	0.214
J. D. Oviedo	227	0.189
D. Molano	48	0.040
R. Lara	41	0.034
J. L. Vargas	38	0.032
J. E. Robledo	28	0.023
N. Ramos	11	0.009
R. A. Quintero	3	0.003
Voto en Blango	54	0.045
Total	1200	1.000

Tabla 1: Si las elecciones para la Alcaldía de Bogotá fueran mañana, y los candidatos fueran los que aparecen en este tarjetón, ¿por cuál de ellos votaría usted?

Aunque Invamer utilizó una clase particular de muestreo aleatorio sin reemplazo (ver Ficha Técnica), es posible considerar la muestra como una muestra aleatoria simple con reemplazo (IID), dado que el tamaño total de la muestra es muy pequeño en comparación con el tamaño del Universo. El principal interés científico y político en este problema se centra en estimar la

proporción poblacional de votos que recibirá cada candidato de acuerdo con los datos proporcionados por Invamer.

Bajo las condiciones dadas anteriormente, dado que nuestra incertidumbre acerca de las respuestas de las 1200 personas en la encuesta es intercambiable, una versión particular del Teorema de De Finetti (Bernardo and Smith, 2000, pág. 176) garantiza que la única distribución muestral apropiada para datos de esta índole es la distribución multinomial.

Suponga que una población de interés tiene artículos de $k \geq 2$ tipos, y además, que la proporción de artículos de tipo j es $0 < \theta_j < 1$, para $j = 1, \ldots, k$. Siendo $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, note que las componentes de $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ son tales que $\sum_{j=1}^{k} \theta_j = 1$. Ahora, suponga que se toma una muestra IID $\boldsymbol{y} = (y_1, \ldots, y_n)$ de tamaño n de la población. Sea $\boldsymbol{n} = (n_1, \ldots, n_k)$ el vector aleatorio que almacena los conteos asociados con cada tipo de artículo, así que n_j es el número de elementos en la muestra aleatoria de tipo j, para $j = 1, \ldots, k$. En esta situación, se dice que $\boldsymbol{n}$ sigue una distribución multinomial con parámetros n y $\boldsymbol{\theta}$, la cual se define como sigue: $\boldsymbol{n} \mid n, \boldsymbol{\theta} \sim \mathsf{Multinomial}(n, \boldsymbol{\theta})$ si y solo si

$$p(\boldsymbol{n} \mid n, \boldsymbol{\theta}) = \frac{n!}{\prod_{j=1}^{k} n_j!} \prod_{j=1}^{k} \theta_j^{n_j} \tag{1}$$

siempre que $\sum_{j=1}^{k} n_j = n$ y $0 \leq n_j \leq n$ para todo $j = 1, \ldots, k$.

Considere el modelo con distribución muestral $\boldsymbol{n} \mid n, \boldsymbol{\theta} \sim \mathsf{Multinomial}(n, \boldsymbol{\theta})$ y distribución previa jerárquica dada por

$$\boldsymbol{\theta} \mid \alpha \sim \mathsf{Dirichlet}(\alpha \mathbf{1}_k) \qquad \text{y} \qquad \alpha \sim \mathsf{Gamma}(a, b)\,,$$

donde $\mathbf{1}_k$ es el vector de unos de $k \times 1$ y a y b son los hiperparámetros del modelo (en el Cap. 8 de Gelman et al. (2013) se discuten modelos más sofisticados que consideran la estrategia de muestreo).

Preguntas

Ajustar el modelo propuesto usando un muestreador de Gibbs con $a = b = 1$ (incluir un anexo con todos los detalles). Reportar visual y tabularmente las estimaciones puntuales, los intervalos de credibilidad al 95% y los resultados oficiales de la Registraduría Nacional del Estado Civil para Galán, Bolivar y Oviedo, expresando todas las cifras en puntos porcentuales. Interpretar los resultados obtenidos (máximo 500 palabras).

2 Selección de modelos

Puede ocurrir que en un análisis de regresión haya un gran número de variables independientes x, aunque puede que la mayoría de estas variables no tengan una relación sustancial con la variable dependiente y. En estas situaciones, incluir todas las variables regresoras en el modelo de regresión conduce a modelos saturados poco parsimoniosos difíciles de interpretar con un rendimiento deficiente. Por lo tanto, se recomienda considerar en el modelo final solo aquellas variables x para las que exista evidencia sustancial de una asociación con y. Esto no solo produce análisis de datos más simples, sino que también proporciona modelos con mejores propiedades estadísticas en términos de predicción y estimación.

Datos de diabetes

Considere la base de datos de diabetes dada en la Sección 9.3 de Hoff (2009, p. 161), que contiene datos asociados con 10 medidas basales $x_1, \ldots, x_{10}$ en un grupo de 442 pacientes diabéticos, así como una medida de progresión de la enfermedad y tomada un año después de las medidas basales. Los datos se pueden descargar de este enlace en `yX.diabetes.train` y `yX.diabetes.test`. A partir de estos datos, el objetivo es hacer un modelo predictivo para y basado en $x_1, \ldots, x_{10}$ (tanto y como las x_j se encuentran estandarizadas). Si bien un modelo de regresión con diez variables no sería abrumadoramente complejo, se sospecha que la relación entre y y las x_j puede no ser lineal, así que se recomienda considerar términos de segundo orden de la forma x_j^2 y $x_j x_k$ para potenciar la capacidad predictiva del modelo. Así, las variables regresoras incluyen diez efectos principales x_j, $\binom{10}{2} = 45$ interacciones $x_j x_k$ y nueve términos cuadráticos x_j^2 (no es necesario considerar x_2^2 en el modelo porque $x_2 =$ sexo es binaria, y por lo tanto $x_2 = x_2^2$). Esto da un total de $p = 64$ variables regresoras (no es necesario considerar el intercepto porque todas las variables se encuentran estandarizadas).

Modelamiento

Se considera un modelo de regresión de la forma $\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$, donde $\boldsymbol{y}$ es un vector de $n \times 1$ que contiene los valores de la variable respuesta, $\mathbf{X}$ es una matriz de $n \times p$ que contiene los valores de las variables regresoras, $\boldsymbol{\beta}$ es un vector de $p \times 1$ que contiene los

parámetros desconocidos, y finalmente, $\mathbf{I}_n$ es la matriz identidad de $n \times n$.

Para evaluar los modelos de regresión, se dividieron aleatoriamente a los 442 individuos con diabetes en 342 individuos de entrenamiento y 100 individuos de prueba, lo que provee un conjunto de datos de entrenamiento $(\boldsymbol{y}_{\text{train}}, \mathbf{X}_{\text{train}})$ y un conjunto de datos de prueba $(\boldsymbol{y}_{\text{test}}, \mathbf{X}_{\text{test}})$. Así, se ajustan los modelos usando $(\boldsymbol{y}_{\text{train}}, \mathbf{X}_{\text{train}})$, y luego, usando los coeficientes de regresión estimados $\hat{\boldsymbol{\beta}} = \mathsf{E}(\boldsymbol{\beta} \mid \boldsymbol{y}_{\text{train}})$, se genera $\hat{\boldsymbol{y}}_{\text{test}} = \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}$. Luego, se evalúa el rendimiento predictivo del modelo comparando $\hat{\boldsymbol{y}}_{\text{test}}$ con $\boldsymbol{y}_{\text{test}}$ por medio de una métrica apropiada.

Modelo 1: Regresión clásica previa unitaria

Distribución previa: Previa unitaria (*unit information prior*; Kass y Wasserman, 1995).

Modelo 2: Regresión clásica previa g

Distribución previa: Previa g (*g-prior*; Zellner, 1986).

Modelo 3: Regresión rígida

Distribución previa:

$$p(\boldsymbol{\beta}, \sigma^2, \lambda) = \mathsf{N}(\boldsymbol{\beta} \mid \mathbf{0}_p, \tfrac{\sigma^2}{\lambda}\mathbf{I}_p) \cdot \mathsf{GI}(\sigma^2 \mid \nu_0/2, \nu_0\,\sigma_0^2/2) \cdot \mathsf{G}(\lambda \mid a_\lambda, b_\lambda)\,,$$

con $\nu_0 = 1$, $\sigma_0^2 = \hat{\sigma}^2_{\text{OLS}}$, $a_\lambda = 1$ y $b_\lambda = 2$.

Modelo 4: Regresión con errores correlacionados

Distribución muestral:

$$\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2, \rho \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{C}_\rho)\,,$$

donde $\mathbf{C}_\rho$ es una matriz con estructura autoregresiva de primer orden de la forma

$$\mathbf{C}_\rho = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}.$$

Distribución previa:

$$p(\boldsymbol{\beta}, \sigma^2, \rho) = \prod_{j=1}^{p} \mathsf{N}(\beta_j \mid 0, \tau_0^2) \cdot \mathsf{GI}(\sigma^2 \mid \nu_0/2, \nu_0\, \sigma_0^2/2) \cdot \mathsf{U}(\rho \mid a_\rho, b_\rho)$$

con $\tau_0^2 = 50$, $\nu_0 = 1$, $\sigma_0^2 = \hat{\sigma}^2_{\text{OLS}}$, $a_\rho = 0$ y $b_\rho = 1$.

Preguntas

Ajustar cada modelo utilizando los datos de entrenamiento $(\boldsymbol{y}_{\text{train}}, \mathbf{X}_{\text{train}})$ (incluir un anexo con todos los detalles).

1. Para cada modelo, generar $\hat{\boldsymbol{y}}_{\text{test}} = \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}$ usando los coeficientes de regresión estimados $\hat{\boldsymbol{\beta}} = \mathsf{E}(\boldsymbol{\beta} \mid \boldsymbol{y}_{\text{train}})$. Graficar $\hat{y}_{\text{test}}$ frente y_{test} y calcular el error absoluto medio $\frac{1}{n}\sum_i |y_{\text{test},i} - \hat{y}_{\text{test},i}|$ correspondiente.

   Nota:

   - Incluir todos los gráficos en una sola figura con 2×2 paneles (todos los paneles deben tener la misma escala en los ejes).
   - En cada gráfico, superponer la recta $\hat{y}_{\text{test}} = y_{\text{test}}$ como punto de referencia (ver el segundo panel de la Figura 9.7 de Hoff 2009, p. 170).
   - En el encabezado de cada gráfica, incluir el error absoluto medio usando tres (3) cifras decimales.

2. Para cada modelo, chequear la bondad de ajuste usando la media como estadístico de prueba. Graficar la distribución predictiva posterior por medio de un histograma.

Nota:

- Incluir todos los gráficos en una sola figura con 2×2 paneles (todos los paneles deben tener la misma escala en los ejes).
- En cada gráfico, superponer el valor observado de la media como punto de referencia.
- En el encabezado de cada gráfica, incluir el valor p predictivo posterior correspondiente usando tres (3) cifras decimales.

3. Para cada modelo, calcular el DIC. Presentar los resultados tabularmente usando tres (3) cifras decimales.

4. Interpretar los resultados obtenidos en los numerales anteriores (máximo 500 palabras).

Referencias

Bernardo, J. M. and Smith, A. F. (2000). *Bayesian theory*. John Wiley & Sons.

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

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.