Regresión Avanzada ITAM
11 de marzo de 2024

Examen Parcial I

A contnuación se dan las reglas de este examen.

- El examen es individual y no puede haber comunicación entre los estudiantes.
- La fecha de entrega es a más tardar el domingo antes de las 12:00am
- Si se indica el uso de datos adicionales, se pueden encontrar en el folder de datos de la página de Canvas
- El examen deberá subirse a Canvas en el lugar creado para este examen.
- Se abrirá un foro de discusión para aclarar preguntas del examen. No se atenderán las preguntas que no se hagan por este medio. La intención es que todos cuenten con la misma información. En este foro, el único que puede responder a las dudas es el instructor.

Preguntas

1. Sea θ la tasa de créditos hipotecarios otorgados por un banco en Argentina. Durante el 2023 la tasa promedio fue de 60 % y la desviación estándar de la tasa fue de 0.04. En lo que va del año 2024 se han solicitado 100 créditos, de los cuales se han otorgado únicamente 50.

   a. Usando la información del año pasado, encuentra la distribución beta que mejor describe el conocimiento inicial.
   b. Usando la información del año pasado, encuentra la distribución normal transformada que mejor describa el conocimiento inicial.
   c. Determina la distribución inicial de referencia.
   d. Usando los datos del año 2024 encuentra la distribución final para cada una de las distribuciones iniciales de los incisos (a) – (c).
   e. Estima la tasa de créditos otorgados, usando las 3 distribuciones finales del inciso (d).

f. Estima el momio de otorgar un crédito, i.e., $\phi = \frac{\theta}{1-\theta}$, usando las 3 distribuciones finales del inciso (d).

2. Las utilidades mensuales de una compañía tienen una distribución $\mathcal{N}(\mu, \sigma^2)$ (aquí se da la varianza, no la precisión). Suponer que una muestra de 10 meses de esta compañía dio como resultado las siguientes utilidades: (212, 207, 210, 196, 223, 193, 196, 210, 202, 221).

   a. La incertidumbre sobre la utilidad promedio anual μ se puede representar por una distribución $\mathcal{N}(200, 40)$, y la incertidumbre de la desviación estándar de las utilidades mensuales se puede representar mediante una distribución $\mathcal{G}(10, 1)$. Mediante la distribución posterior estima μ y σ^2.

   b. Utilizando una distribución inicial no informativa, estima mediante la correspondiente distribución inicial μ y σ^2.

3. A continuación se presenta una base de datos de calificaciones de 20 empresas financieras hechas por las dos compañías calificadores más importantes S&P y Moody's (ver el archivo `calificaciones.txt`. Realiza un análisis Bayesiano completo de los datos, ajustando un modelo de regresión lineal, tomando como variable respuesta las calificaciones de S&P y como variable explicativa las calificaciones de Moody's.

4. Un investigador desea evaluar la relación entre el salario anual de trabajadores de una compañía de nivel medio y alto (Y, en miles de dólares) y el índice de calidad de trabajo (X_1), número de años de experiencia (X_2) y el índice de éxito en publicaciones (X_3). La muestra consiste de 24 trabajadores. Realiza un análisis Bayesiano completo de los datos y obtén las predicciones de salarios para 3 nuevos empleados con variables explicativas:

$$x'_{1F} = (5{,}4, 17, 6{,}0), x'_{2F} = (6{,}2, 12, 5{,}8), x'_{3F} = (6{,}4, 21, 6{,}1)$$

   Los datos se encuentran en el archivo `salarios.txt`.

5. Una compañía de seguros quiere lanzar un nuevo seguro médico para mineros. Para ello desea estimar la probabilidad de muerte (π_i), con base en el tiempo de exposición al mineral (x_i en horas). Se cuenta con información de las muertes registradas entre 1950 y 1959, junto con el tiempo de exposición al mineral y el número de mineros expuestos. Realiza un análisis Bayesiano de los datos y obtén la distribución predictiva del número de muertes suponiendo que hay 100 mineros con un tiempo de exposición de 200 horas. Los datos se encuentran en el archivo `mortality.txt`. El modelo es el siguiente: Para $i = 1, \ldots, N$

$$\begin{aligned} Y_i|\pi_i &\sim Bin(n_i, \pi_i) \\ logit(\pi_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

   con $\beta_0 \sim N(0, 0{,}001)$ y $\beta_1 \sim N(0, 0{,}001)$.

a. En el mismo contexto del problema enunciado (que hicimos en la última clase), supongamos ahora que la compañía de seguros está interesada en modelar el número total de desastres (Y_t) que ocurren en la mina. Se cuenta con $N = 112$ observaciones durante los años 1851 a 1962. Se proponen tres modelos:

   i. Modelo con tasa variable en función del tiempo:

   $$\begin{array}{rcl} Y_t|\mu_t & \sim & Poi(\mu_t) \\ log(\mu_t) & = & \beta_0 + \beta_1 x_t \end{array}$$

   con $\beta_0 \sim N(0, 0{,}001)$ y $\beta_1 \sim N(0, 0{,}001)$.

   ii. Modelo con tasa constante en dos períodos: Se cree que la tasa promedio de desastres es constante, pero que en el siglo XX la tasa ha disminuido. Esto se traduce en el siguiente modelo:

   $$\begin{array}{rcl} Y_t|\mu_t & \sim & Poi(\mu_t) \\ log(\mu_t) & = & \beta_0 + \beta_1 I(t \geq \tau) \end{array}$$

   con $\beta_0 \sim N(0, 0{,}001)$ y $\beta_1 \sim N(0, 0{,}001)$ y $\tau \sim U\{1, \ldots, N\}$.